ENGLISH REVIEW: Journal of English Education
Volume 11, Issue 2, June 2023
p-ISSN 2301-7554, e-ISSN 2541-3643
https://journal.uniku.ac.id/index.php/ERJEE

HATE SPEECH ON SOCIAL MEDIA: A CASE STUDY OF BLASPHEMY IN INDONESIAN CONTEXT

Rotua Elfrida
English Department, Faculty of Language and Arts
Universitas HKBP Nommensen
Email: rotuapangaribuan@ac.id

Arsen Nahum Pasaribu (Corresponding author)
English Department, Faculty of Language and Arts
Universitas HKBP Nommensen
Email: arsen.pasaribu@uhn.ac.id

APA Citation: Melfrida, R., & Pasaribu, A. N. (2023). Hate speech on social media: A case study of Blasphemy in Indonesian context. *English Review: Journal of English Education, 11*(2), 433-440. https://doi.org/10.25134/erjee.v11i2.7909

Received: 28-02-2023 Accepted: 27-04-2023 Published: 30-06-2023

Abstract: Many scholars have shown research on hate speech, spanning from hate speech detection methods to the bias interpretation of dataset they create, as well as the role of technology in the dissemination of hate speech on social media. However, research on hate speech categories and degrees on the issue of religious blasphemy is still relatively unexplored. Therefore, the purpose of this research is to uncover the strategies and levels of hate speech on social media, primarily YouTube channels, in response to the Minister of Religion's comments about the sound of mosque loudspeakers that need to be adjusted in volume. This comment has generated both positive and negative reactions in Indonesian society. This research looks into netizen comments in the comments column on the YouTube channel that carries the statement. Purposive sampling was used to select 300 comments from among the 840 comments in the comments section of the YouTube. For the purposes of this study, the sample was obtained in the form of comments containing hate speech. The data was then analyzed using content analysis, in which the data was coded and categorized according to hate speech level of classification. The study revealed that there are three types of hate speech in netizen comments: early warning, dehumanization and demonization, and violence and incitement. Early warning is the most common type of hate speech, followed by violence and hostility, as well as dehumanization and demonization. Due to cultural influences and contrasts in rank and power between the commentator and the person who is the subject of the hate speech, hate speech delivered by Indonesian netizens tends to be dominated by disagreement, negative character, and action.
Keyword: *content analysis; hate speech; Indonesian context; religion blasphemy; social media; YouTube comments*

INTRODUCTION

Hate speech is a global phenomenon in social interaction. It can happen anywhere and anytime. In public spaces, such as public transportation, parks, schools, markets, shopping centers, offices, and even in the family environment like at home. Hate speech can also occur when someone is with friends, family, teachers, co-workers, and even when with parents.

Technological developments, such as the invention of the internet, have had a huge impact on social interactions around the world. The emergence of social media platforms such as Facebook, Twitter, Instagram, and YouTube, to name a few, has enabled people around the world to interact widely and instantly. Real communication has changed to online with all the advantages and disadvantages. Likewise, the practice of hate speech has moved into cyberspace with various available internet-based communication facilities (Matamoros-Fernández & Farkas, 2021; Poletto *et al.*, 2021; Chetty & Alathur, 2018).

The existence of hate speech in social media has been revealed by some scholars. Most of the hate speech research found on flat form Twitter were carried out by several scholars (Sanguinetti *et al.*, 2019; Albadi *et al.*, 2018; Davidson *et al.*, 2019; Fortuna *et al.*, 2019; Ibrohim & Budi, 2019; Mulki *et al.*, 2019; Oriola & Kotze, 2020; Pereira-Kohatsu *et al.*, 2019; Alshalan & Al-Khalifa, 2020; Mozafari *et al.*, 2020; Roy *et al.*, 2020). A small portion of hate speech research is focused on other flat forms such as Facebook (Sigurbergsson &

Derczynski, 2020; Pasaribu, 2021). Even on other social media flat forms were not found.

Research on hate speech that focuses on the tools or methods used to detect hate speech on social media was conducted by Albadi *et al.*, (2018), Sanguinetti *et al.*, (2019), Davidson *et al.* (2019), Fortuna *et al.*, (2019), Ibrohim & Budi (2019), Mulki *et al.* (2019), Oriola & Kotze (2020), Pereira-Kohatsu *et al.* (2019), Alshalan & Al-Khalifa (2020), Mozafari *et al.* (2020).

Meanwhile, other studies have focused on the bias of annotators (Sap *et al.*, 2020) and the bias caused by the word hate speech used (Kennedy *et al.*, 2020; Wich *et al.*, 2020; Xia *et al.*, 2020). Other topic research is the impact caused by exposure to hate speech (Bilewicz & Soral, 2020).

Research on hate speech with a religious background (religious hate speech) is also found in various flat forms of social media. Research conducted by Albadi *et al.* (2018) in Saudi Arabia regarding religious hate speech in the world of Twitter. They try to dismantle the practice of religious hate speech on Twitter by detecting it using various approaches. In India, Bohra *et al.* (2018) also tried to develop a hate speech detection method using the code-mixing method. This research reveals that this method is also effective in disclosing hate speech practices on Twitter.

From the research on hate speech that has been described above, several important points can be concluded that research on hate speech on social media tends to focus on the flat form of Twitter compared to other flat forms of social media such as Facebook or Instagram. Then the hate speech research that was carried out was dominated by the various methods and tools that can detect the presence of hate speech on social media. Research on the bias of interpreting hate speech datasets on social media is also a topic of interest to researchers. However, there are still relatively few research topics on religious hate speech on social media, especially research conducted by Indonesian scholars.

Based on the analysis above, research on hate speech on social media still leaves room for exploration. Detection of hate speech on other flat forms of social media such as Facebook, Instagram, and YouTube channels needs to be done to see the similarities and differences with the use of Twitter. Besides that, it is also interesting to use a qualitative approach to analyze, interpret, and categorize hate speech datasets. Research on hate speech with relation to religion matter is still relatively unaddressed well.

In Indonesian context, the issue of religious blasphemy has always attracted national attention. The issue of religious blasphemy which is the object of research is the case of religious blasphemy which was accused by the Indonesian Minister of Religion. The Religious Minister's comments were about the rule of mosque's loudspeaker volumes in Indonesia. The Religious Minister's statement has generated controversy among Indonesian Muslims. This issue has also been commented on immensely in social media, such as Twitter, Facebook, Instagram, and the YouTube channel. Thus, this study seeks to explore hate speech on the YouTube channel related to the issue of religious blasphemy in Indonesia.

Hate speech is the speech that assaults, dehumanizes, or incites violence or prejudice against individuals or groups on the basis of their race, ethnicity, national origin, religion, gender identity, sexual orientation, or other characteristics (Kovács *et al.*, 2021; Pasaribu, 2021). The varieties of hate speech can vary depending on the context and the intended audience (Wich *et al.*, 2020; Sanguinetti *et al.*, 2019; Mathew *et al.*, 2019; de Gibert *et al.*, 2018; Albadi *et al.*, 2018; Bohra *et al.*, 2018), but the following are common examples: hate speech directed at a specific race or ethnic group with the intention of inciting animosity or discrimination; Religious hate speech is hate speech directed at individuals or groups on the basis of their religion or religious beliefs. Homophobic hate speech is directed at individuals or organizations on the basis of their sexual orientation or gender identity. Ableist is hatred speech that targets people with disabilities and ridicules or demeans them. Misogynistic speech targets women or reinforces gender stereotypes and discrimination; xenophobic speech targets individuals or groups on the basis of their nationality or country of origin. Anti-immigrant speech targets immigrants or immigrant communities with the intent to incite hostility or discrimination, whereas anti-Semitic speech targets Jewish individuals or communities and frequently invokes stereotypes or conspiracy theories.

Hate speech is also classified into three types (Fortuna *et al.*, 2019; Mulki *et al.*, 2019). The first classification is early warning. The lowest level of the category is regarded to be hate speech. The level of hate speech in this category is still about disagreement, negative character, and negative conduct. The second type of dehumanization is demonization. This sort of hate speech includes

p-ISSN 2301-7554, e-ISSN 2541-3643
https://journal.uniku.ac.id/index.php/ERJEE

rhetoric that diminishes human dignity by referring to people as animals, devils, or demons. The final classification is violence and intent. Hate speech in this category takes the form of harsh statements intended to incite others to commit violence and murder. This category is more than just speech; it has also perpetrated acts of violence and murder.

METHOD

This study intends to uncover hate speech on the issue of blasphemy by Indonesians on social media, specifically the YouTube channel. This study uses a mixed method, the combination of descriptive qualitative and descriptive quantitative. The amount and percentage of different sorts of hate speech that arise in YouTube channel are conveyed using descriptive quantitative method. In the meantime, the qualitative research approach was being employed to describe and understand the meaning of the hate speech in the comments of the YouTube based on the context.

A total of 840 comments were found on the YouTube channel with the following link: https://www.youtube.com/watch?v=NwSGRfMFjns. The YouTube contains comments related to the religious blasphemy addressed to the Minister of Religion of the Republic of Indonesia. Around 300 comments were selected as the sample of the data. This research employed purposive sampling. The sample were taken in the form of comments containing hate speech from the three categories of hate speech above. The number of samples is considered sufficient until it reaches 300 comments that are detected hate speech. This study uses two ratters to re-check the research sample containing hate speech. The two ratters are lecturers who have in-depth knowledge of the theory and categorization of hate speech. The use of ratters in detecting the presence of hate speech in netizens' comments can increase the validity and reliability of the research data used.

Research data analysis uses qualitative content analysis (Schreier, 2014). The data analysis procedure for content analysis is as follows: (1) Data preparation. (2) Define the unit of analysis. (3) Develop categories and coding scheme. (4) Test your coding scheme on a sample of text. (5) Code all the text. (6) Assess your coding consistency. (7) Draw your conclusion from the coded data. (8) Report your method and finding.

To increase the validity and reliability of the data analysis procedure number 6 must be done properly until the coding consistency is gained. In addition, the method of the data analysis procedures was explained clearly to make the replication of research possible to do.

RESULTS AND DISCUSSION

This study attempts to uncover hate speech and its categories and level on social media, especially YouTube by Indonesians relating to the blasphemy problem that occurred in Indonesia. Comments made by Indonesian netizens on social media are examined and classified as hate speech in three categories: dehumanization and demonization, violence and provocation, and early warning. The table below divides the number of hate speech detected on social media into these three groups.

Table 1. *The category of hate speech in social media*

No	Category of hate speech	N	%
1	Dehumanization and demonization	58	19.33
2	Violence and incitement	104	34.67
3	Early warning	138	46.00
	Total	300	100.00

Table 1 depicts the many types of blasphemy-related hate speech on social media. According to the table, the early warning category dominated the 300 hate speech comments collected on social media, accounting for 138 comments or 46% of the entire data, followed by violence and incitement, accounting for 104 comments or 34.67% of the total research data. Finally, 58 comments or 19.33% of the total data indicated the sort of dehumanization and demonization.

Early warning category

As previously stated, the early warning category represents the lowest level of hate speech. Hate speech is classified as an early warning signal in this category because it has the potential to escalate to the level of violence or dehumanization. The focus of hate speech in this form is the struggle between one group and another, or "us" versus "them," who have opposing views, thoughts, and beliefs (Sanguinetti *et al.*, 2019). The early warning category is separated into three categories, with "disagreement" being the lowest. At this level, hate speech takes the form of inconsistencies about ideas, opinions, and beliefs that differ amongst social groupings. 'Wrong,' 'incorrect,' 'false,' 'persuade,' 'change opinion,' and 'challenge' are examples of words, statements, or acts employed at this level (Ibrohim & Budi, 2019). Then, in this group, the amount of hatred is negative action. The level of negative action is the nonviolent activity connected with a group. This type of hate speech

takes the shape of remarks or actions that are not violent in nature. The statement might be expressed using words or metaphorical language. Poor treatment, stealing, threatening, and outrageous deeds are examples of words, statements, or actions employed at the negative action level. Finally, there is a negative character on the third level. Hate speech is described at this level as rhetoric, which includes nonviolent characterization and insults. Stupid, fake, insane, and thief are some examples of this level (Mulki *et al.*, 2019; Pereira-Kohatsu *et al.*, 2019). The data as the representation of certain data in the early warning category can be seen in the extract 1 below.

Extract 1

> "*Semoga bapak di beri akal yg cerdas lagi ya pak...dari jaman dulu hinggal saat ini aja baru toa mesjid di permasalah kan*". (Hoping (God) give you a wise thinking Sir…from old ages, only now the loud speaker of mosque to be blamed)

Comments from netizens on data extract 1: "*Semoga bapak diberikan akal yang cerdas*." (Hopefully, you will be granted a smart mind) imply that "*bapak*" alludes to Indonesia's Minister of Religion, Yaqut Cholil Qoumas. This comment implies that the minister's mind is not irrational to make such statement. According to the commentator, the minister of religion does not need to comment on the mosques' loudspeakers in Indonesia. According to the analyst, this has always been the case, and no one has yet restricted the noise produced by mosque loudspeakers in Indonesia.The phrase "*diberikan akal cerdas*" is a euphemism that can be used to replace the phrase "*Anda tidak cerdas*." or a more formal phrase "*Anda bodoh*". This strategy was devised by a commentator in order to improve the ability to convey information so that those who are affected by it do not become upset.

This statement may be classified as "early warning" hate speech with a 'negative character'. The expression "*diberikan akal cerdas*" (given a smart mind) is a type of euphemism that attempts to lighten the meaning of the original sentence "Anda tidak cerdas." (You are not smart) or a stronger statement "*Anda bodoh*" (you are stupid). This method is used by critics to retain an attitude of presenting thoughts such that the person referred to in the sentence is not insulted (Nozza, 2021). This statement falls within the category of "early warning" hate speech having a 'negative character'.

Extract 2

> "*Mengibaratkan suara azan dengan gonggongan anjing itu tidak tepat, sebaiknya lebih hati hati lagi bicaranya*." (Comparing the sound of the *azan* with the barking of a dog is not correct, you should be more careful what you say).

In extract 2, the commentators attempt to transmit that *azan* (the call to prayer at a mosque) is distinct from the sound of a dog barking. This assertion is considered to be false. Therefore, the phrase "*Sebaiknya lebih hati-hati lagi bicaranya*" (You should be more careful what you say) is a warning to the individual you are addressing not to communicate controversial statements that offend religious communities in Indonesia. The statement in extract 2 can be classified as a "early warning" with a level of "disagreement" due to the fact that this type of hate speech is at the most fundamental level. The commentator disagreed with the speaker's assertion. The commentators then attempt to provide a gentle warning to never restore it again.

According to the previous research, the early warning category is the most prevalent type of hate speech found in the comments section of YouTube channels. This category includes the most basic or initial form of hate speech, which may escalate to a harsher or more violent form. These forms and categories are more prevalent in netizen comments due to two factors. First, Indonesian culture influences the politeness of netizen communication. Second, the individual discussed by netizens is a minister with a higher status and greater authority than the commentators. This phenomenon is consistent with Pasaribu (2021) assertion that differences in status and power can influence communication in people's real-world and cyberspace social interactions.

Dehumanization and demonization

The next category of hate speech is dehumanization and demonization, which includes statements that refer to humans as animals, demons, or spirits, or statements that diminish the degree or status of humanity (Wich *et al.*, 2020). This statement will have a negative effect on the spirit and mind of the individual to whom it is directed. Here are excerpts illustrating hate discourse in this category:

Extract 3

> "*Pak menteri saya sarankan silaturahmi ke ulama atau ustad minta diruqiyah...biar bisa menikmati suara adzan*"(Mr. Minister, I suggest

p-ISSN 2301-7554, e-ISSN 2541-3643
https://journal.uniku.ac.id/index.php/ERJEE

> a visit to the ulema or ustad to ask for *diruqiyah*... so you can enjoy the sound).

The statement in extract 3 implies that the Minister of Religion is possessed by a genie or demon, as he does not appreciate hearing the mosque's call to prayer. Therefore, he was requested to perform "diruqiyaah," the exorcism of demons or spirits from the body of a person believed to be possessed. This remark is classified as "violence and incitement" because it implies that the target of the hate speech is in a stupor. He is considered a devil or a genie who does not like the sound *azan* (the call to prayer).

Extract 4

> "*Kalau Bisa Menterinya Di Ganti, Ibarat Orang Pelihara 1 Bab1 Di Kebun Orang Lain, Maka Bab1 Itu Akan Merusak Kebun Tsb*." (If the minister can be replaced, it's like a person raising pigs in someone else's garden, then the pigs will destroy the garden)

The hate speech described in passage 4 falls into two categories: violence and incitement, and dehumanization and incitement. The first category is that of violence and intent. The preceding statement contains a provocative element, namely "...Menterinya diganti" (the minister is replaced) to refer to the Minister of Religion. The following category consists of dehumanization and incitement. The expression "Bab 1" in the statement is a form of "dehumanization" hate speech because the minister is referred to as a "Bab 1" (pig) animal. This is a very impolite statement, particularly in the context of Indonesian culture. A person's dignity is considered diminished when he or she is compared to animals, particularly swine, which Muslims consider impure. In other words, comments made by Internet users may contain more than one type of hate speech. Similar to this extract 4

Violence and incitement

This category includes hate speech in the form of statements about violent acts or incitements to perpetrate violent acts from one group to another. This category includes two distinct types of hate discourse. The first category consists of hate speech in the form of caustic and provocative remarks or calls to commit physical violence (Wich *et al.*, 2020; Matamoros-Fernández & Farkas, 2021). In contrast, the second category encompasses hate speech that incites murderous acts of violence. The data representation for this category is shown in Extracts 5 and 6 below.

Extract 5

> "*Harusnya di pecat nih menteri yg bikin gaduh*." (This minister should be fired for making noise)

The statement "*harusnya dipecat*" (must be fired) in extract 5 indicates a provocation for the Minister of Religion to be fired for his statement, which he considers to have offended Muslims in Indonesia. The group that claimed the statement on religious matters contained blasphemy attempted to discredit religion. Multiple comments comprising hate speech were posted on the Internet by Indonesian users. Even with a level that is more violent and cruel as demonstrated in this excerpt 5. Consequently, this statement falls under the category of violence and incitement.

Extract 6

> "*Adzan dimisalkan gonggongan anjing. Mungkin jika orang yang berkata seperti itu hidup di era sayyidina Umar al Khattab, kepala orang itu akan dipenggal*" (*Azan* is regarded as a dog barking. Maybe if the person who said that lived during the era of Sayyidina Umar al Khattab, that person's head would have been beheaded)

Extract 6 is a hate speech that demonstrates the extent of propaganda and death threats as a result of the claimed comment to the minister of religion that the call to prayer is analogous to a dog barking. This sentence implies that if the occurrence occurs under the reign of Syayaidina Umar al Khattab, the minister may also be decapitated. This discourse is characterized as hate speech, violence, and incitement since it contains remarks that promote violence and even murder (Chetty & Alathur, 2018).

According to the above research findings, the forms and categories of hate speech in social media communication can vary. It begins with an expression of disapproval and escalates to the level of incitement to perpetrate violence or murder (ElSherief *et al.*, 2018). Similar to the hate speech found in the responses of Indonesian netizens to the current statement by the Minister of Religion of the Republic of Indonesia, Yaqut Cholil Qoumas. The minister's statement that the intensity of the mosque's loudspeaker needs to be adjusted has elicited both pro and con arguments from Indonesian Muslims. Various parties subsequently distorted the minister's and the government's statement in an attempt to provoke public blame.

The previous research revealed the prevalence of hate discourse on social media. Using

euphemisms and cynicism, the commentators attempt to soften the hate speech. This is done to avoid conflict with the intended individual, who has a higher status and greater authority than the commentators (Ibrohim & Budi, 2019). Moreover, dehumanization and demonization-related hate speech, as well as violence and incitement-related hate speech, can be found in netizen comments. In the remarks of netizens, there is a prevalence of hate speech with harsh language, including death threats. The reason commentators dare to convey such hate speech is because communication between hate speech commentators and the intended recipient is indirect, such as through social media or YouTube channel (Sanguinetti *et al.*, 2019; Davidson *et al.*, 2019). In order to avoid the commentators' identity identified, they also created social media account with fake identities.

This research has confirmed that the hate speech found in the YouTube comments section related to the issue of blasphemy conducted by the Indonesian Minister of Religion consists of hate speech in the low (early warning), moderate (dehumanization and demonization) and dangerous (violence and incitement) categories. Besides that, the amount of hate speech found in one issue on YouTube is also considered massive. You can imagine hate speech products that are found in other flat forms of social media, maybe thousands or even millions of hate speeches delivered every day.

This research has revealed an interesting finding that Indonesian people tend to use low and medium categories of hate speech on social media, although the small number of high categories is still detected. This fact shows that Indonesian citizens still adhere to politeness traditions passed down from generation to generation.

These research findings will bring implications to the Indonesian government to make policies or rules to detect and reduce the spread of hate speech on social media. In addition, the government urgently needs to design an educational curriculum to reduce the emergence of hate speech in the future. For Indonesian people, they should be more aware of the future risks of posting hate speech on social media. Some cases of hate speech, especially related to the religious blasphemy have ended in law enforcement.

CONCLUSION

This study has disclosed the forms and categories and the level of hate speech found in the YouTube comments of Indonesian netizens. With the Minister of Religion as the target of hate speech, the issue of religious blasphemy is used as a pretext to spread hate speech through the YouTube comments section. This study affirms that netizen comments containing hate speech fall into three distinct categories: early warning, dehumanization and demonization, and violence and incitement. Due to the method of indirect communication and the disparity between the position or status of the commentators and those who are commented on, the number and types of hate speech discovered in the data vary.

This study has several limitations. The number of corpus data in this research must be increased so that appropriate conclusions can be drawn. In addition, this research employs manual methods for data collection and analysis. Therefore, data analysis takes longer time and has potential to be biases in data interpretation when compared to the use of applications or tools.

REFERENCES

Albadi, N., Kurdi, M., & Mishra, S. (2018). Are they our brothers? analysis and detection of religious hate speech in the Arabic Twittersphere. *Proceedings of the 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2018*, 69–76. https://doi.org/10.1109/ASONAM.2018.8508247

Alshalan, R., & Al-Khalifa, H. (2020). A deep learning approach for automatic hate speech detection in the saudi twittersphere. *Applied Sciences (Switzerland)*, *10*(23), 1–16. https://doi.org/10.3390/app10238614

Bilewicz, M., & Soral, W. (2020). Hate Speech Epidemic. The Dynamic Effects of Derogatory Language on Intergroup Relations and Political Radicalization. *Political Psychology*, *41*(S1), 3–33. https://doi.org/10.1111/pops.12670

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 2nd Workshop on Computational Modeling of PFople's Opinions, PersonaLity, and Emotions in Social Media, PEOPLES 2018 at the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language T*, 36–41. https://doi.org/10.18653/v1/w18-1105

Chetty, N., & Alathur, S. (2018). Hate speech review in the context of online social networks. *Aggression and Violent Behavior*, *40*, 108–118. https://doi.org/10.1016/j.avb.2018.05.003

Davidson, T., Bhattacharya, D., & Weber, I. (2019). *Racial Bias in Hate Speech and Abusive Language Detection Datasets*. 25–35. https://doi.org/10.18653/v1/w19-3504

de Gibert, O., Perez, N., García-Pablos, A., & Cuadros, M. (2018). Hate Speech Dataset from a White Supremacy Forum. *2nd Workshop on Abusive Language Online - Proceedings of the Workshop, Co-Located with EMNLP 2018*, 11–20. https://doi.org/10.18653/v1/w18-5102

ElSherief, M., Kulkarni, V., Nguyen, D., Wang, W. Y., & Belding, E. (2018). Hate lingo: A target-based linguistic analysis of hate speech in social media. *12th International AAAI Conference on Web and Social Media, ICWSM 2018*, *ICWSM*, 42–51. https://doi.org/10.1609/icwsm.v12i1.15041

Fortuna, P., Rocha da Silva, J., Soler-Company, J., Wanner, L., & Nunes, S. (2019). *A Hierarchically-Labeled Portuguese Hate Speech Dataset*. 94–104. https://doi.org/10.18653/v1/w19-3510

Ibrohim, M. O., & Budi, I. (2019). *Multi-label Hate Speech and Abusive Language Detection in Indonesian Twitter*. 46–57. https://doi.org/10.18653/v1/w19-3506

Kennedy, B., Jin, X., Davani, A. M., Dehghani, M., & Ren, X. (2020). Contextualizing hate speech classifiers with post-hoc explanation. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 5435–5442. https://doi.org/10.18653/v1/2020.acl-main.483

Kovács, G., Alonso, P., & Saini, R. (2021). Challenges of Hate Speech Detection in Social Media: Data Scarcity, and Leveraging External Resources. *SN Computer Science*, *2*(2), 1–15. https://doi.org/10.1007/s42979-021-00457-3

Matamoros-Fernández, A., & Farkas, J. (2021). Racism, Hate Speech, and Social Media: A Systematic Review and Critique. *Television and New Media*, *22*(2), 205–224. https://doi.org/10.1177/1527476420982230

Mathew, B., Saha, P., Tharad, H., Rajgaria, S., Singhania, P., Maity, S. K., Goyal, P., & Mukherjee, A. (2019). Thou shalt not hate: Countering online hate speech. *Proceedings of the 13th International Conference on Web and Social Media, ICWSM 2019*, *Icwsm*, 369–380. https://doi.org/10.1609/icwsm.v13i01.3237

Mozafari, M., Farahbakhsh, R., & Crespi, N. (2020). Hate speech detection and racial bias mitigation in social media based on BERT model. *PLoS ONE*, *15*(8 August), 1–26. https://doi.org/10.1371/journal.pone.0237861

Mulki, H., Haddad, H., Bechikh Ali, C., & Alshabani, H. (2019). *L-HSAB: A Levantine Twitter Dataset for Hate Speech and Abusive Language*. 111–118. https://doi.org/10.18653/v1/w19-3512

Nozza, D. (2021). Exposing the limits of Zero-shot Cross-lingual Hate Speech Detection. *ACL-IJCNLP 2021 - 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, *2*, 907–914. https://doi.org/10.18653/v1/2021.acl-short.114

Oriola, O., & Kotze, E. (2020). Evaluating Machine Learning Techniques for Detecting Offensive and Hate Speech in South African Tweets. *IEEE Access*, *8*, 21496–21509. https://doi.org/10.1109/ACCESS.2020.2968173

Pasaribu, A. N. (2021). Hate Speech on Joko Widodo'S Offical Facebook: an Analysis of Impoliteness Strategies Used By Different Gender. *ELTIN JOURNAL, Journal of English Language Teaching in Indonesia*, *9*(1), 56–64.

Pereira-Kohatsu, J. C., Quijano-Sánchez, L., Liberatore, F., & Camacho-Collados, M. (2019). Detecting and monitoring hate speech in twitter. *Sensors (Switzerland)*, *19*(21), 1–37. https://doi.org/10.3390/s19214654

Poletto, F., Basile, V., Sanguinetti, M., Bosco, C., & Patti, V. (2021). Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation*, *55*(2), 477–523. https://doi.org/10.1007/s10579-020-09502-8

Roy, P. K., Tripathy, A. K., Das, T. K., & Gao, X. Z. (2020). A framework for hate speech detection using deep convolutional neural network. *IEEE Access*, *8*, 204951–204962. https://doi.org/10.1109/ACCESS.2020.3037073

Sanguinetti, M., Poletto, F., Bosco, C., Patti, V., & Stranisci, M. (2019). An Italian twitter corpus of hate speech against immigrants. *LREC 2018 - 11th International Conference on Language Resources and Evaluation*, 2798–2805.

Sap, M., Card, D., Gabriel, S., Choi, Y., & Smith, N. A. (2020). The risk of racial bias in hate speech detection. *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, 1668–1678. https://doi.org/10.18653/v1/p19-1163

Schreier, M. (2014). Qualitative Content Analysis. *The SAGE Handbook of Qualitative Data Analysis*, 170–183. https://doi.org/10.4135/9781446282243.n12

Sigurbergsson, G. I., & Derczynski, L. (2020). Offensive language and hate speech detection for danish. *LREC 2020 - 12th International Conference on Language Resources and Evaluation, Conference Proceedings*, 3498–3508.

Wich, M., Bauer, J., & Groh, G. (2020). *Impact of Politically Biased Data on Hate Speech Classification*. 54–64. https://doi.org/10.18653/v1/2020.alw-1.7

Xia, M., Field, A., & Tsvetkov, Y. (2020). *Demoting Racial Bias in Hate Speech Detection*. 7–14. https://doi.org/10.18653/v1/2020.socialnlp-1.2