Pharos Journal of Theology ISSN 2414-3324 online Volume 107 (1) Regular Issue (2026)

Hate Speech, Takfīr, and the Theology of Mercy: A Corpus Study of Arabic Comments on Online Sermons

Nursukma Suri*
Faculty of Cultural Studies, Universitas Sumatera Utara, Medan, Indonesia
https://orcid.org/0000-0003-2367-7439
*Corresponding Author: nursukma.suri@usu.ac.id

Aminuddin
Faculty of Ushuluddin and Islamic Studies, Universitas Islam Negeri Sumatera Utara
Sumatera Utara, Indonesia
https://orcid.org/0009-0004-6375-2960

Naqil Sayyaf Al Mujahid
Faculty of Social Sciences, Universitas Islam Negeri Sumatera Utara
https://orcid.org/0009-0008-0076-7603

https://doi.org/10.46222/pharosjot.107.26

Abstract

This study looks at hate speech in the digital preaching environment which includes social media platforms like YouTube and TikTok in the Arab world; we looked at the practice of takfīr and its role in the theological thought of mercy. We collected data from 9 representative comments posted between July and August 2025 and did a qualitative descriptive analysis and content analysis. We found that hate speech uses the same language which includes assertions of unbelief, innovation (bid'ah), misguidance/polytheism (shirk), threats of the hereafter, and the Muslim vs non-Muslim divide. These are not only random outbursts but also what we may term as exclusive speech acts which put down other groups, promote in group vs out group dynamics in the Muslim community, and which in turn shut out interfaith dialogue. Also we saw that platform algorithms play a role in the spread of this type of content which in turn reaches wider audiences faster. We see in this the role of the theology of mercy which puts forth a message of compassion, gentleness and tolerance as the base of Islamic communication. We did a study which included linguistic and theological analysis which we present as a new framework for digital preaching that is more inclusive. Also we assert that the fight against hate speech is not just a technical issue but a theological mission to re claim digital spaces as forums for the promotion of brotherhood and peace.

Keywords: Hate Speech, Takfir, Theology of Mercy, Linguistics, Online Sermons

Introduction

Recent studies report on the large scale issue of Arabic language hate speech which we see in the Jordanian Hate Speech Corpus that has over 300,000 tweets which they have labeled as neutral, hate, or irrelevant, also the ADHAR Corpus which has thousands of multi-dialect texts that they put into categories of religious, ethnic, or racial based hate speech (Charfi et al., 2024). We also have in our research which looks at hate speech detection with NLP models like AraBERT that we see very high accuracy rates of over 90% which in turn indicates that there are large scale consistent linguistic patterns in Arabic hate speech across many digital platforms (Zaghouani Biswas, 2025). This we see to not be a random thing but a large scale reality with social impact which is amenable to academic study.

In today's world we see an increase of hate speech in what should be used for enlightenment in digital preaching spaces. This rise is largely driven by algorithm-driven amplification, sectarian contestation, and the growing normalization of takfīr rhetoric within online religious subcultures. Sermons and khutbahs which are accessed by millions see to have comments which put forth terms like "kafir", "bid'ah", and "musyrik" directed at different groups. These labels which may pass for words are in fact very serious in a theological sense as they do away with some one's religious standing. What by design are meant to be public discussion forums instead become stages for rhetoric which in turn divide communities.

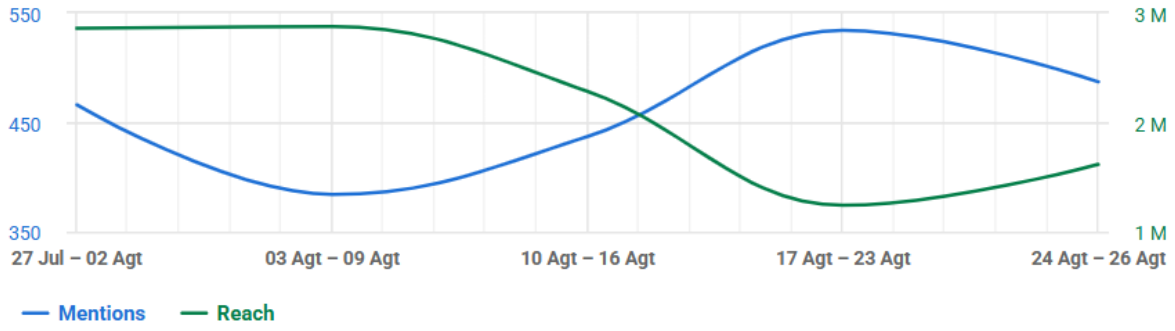

Figure 1. Dynamics of Hate Speech in the Digital Space (Source: Brand 24, 2025)

Digital reporting shows that there is a large gap between what we are seeing in terms of the amount of hate speech put out there and the number of people it is reaching over the past month. We saw an increase in the number of mentions which went from approximately 420 at the start of August to over 530 in the week of 17 23 August, at the same time the audience reached out was reduced which went down to a low of around 1.6 million at the same time. Though the audience may have gone down, what we are seeing is still very large numbers of people which are in the millions exposed to this hate speech. This also brings to light that hate speech is still very much a large scale issue even as the groups it is reaching out to has reduced.

Takfīr which scholars for years have reported as a dangerous practice does indeed issue out accusations of unbelief against fellow Muslims without proper proof may in fact hurt the accuser and damage social ties within the Muslim community (Alhazmi et al., 2024). In the digital sphere the meaning of takfīr has changed from a fatwa context to a every day language used by the public which makes it a form of everyday hate. Terms like kafir, bid'ah, and musyrik which serve as labels also function as pragmatic speech acts that put down other groups, break social ties, and create moral hierarchies (Shannag et al., 2022). Also we see through linguistic study that these words form systematic derogatory patterns and play the role of ideological communication strategies (Prati et al., 2024).

The social impact of hate speech in online preaching forums is a topic which we see play out in the division within the Muslim community itself (Karjo Ng, 2020; Naeem, 2022; Obermaier et al., 2023). We see that comments which put forth hate towards other groups contribute to internal fracture within a community which in actuality has a pluralistic past (Toscano, 2025). This polarization does not only divide ideologically but also shrinks the space for religious discussion. Instead of broadening the scope of faith hate speech fosters an atmosphere of mutual doubt which in turn weakens social cohesion.

Corpus analysis is key to which we study how hate speech comes to be a recurring theme, which in turn forms language and digital religious based discourses that are put under academic scrutiny (Lin Adolphs, 2023; Paquot Gries, 2021). Also it is noted that within the tenets of Islam which puts forth rahmah (grace) as the base element in social interaction, hate speech plays the role of the opposite of what religious communication is supposed to do which is to improve social bonds (Dalimunthe et al., 2024; Farid et al., 2024). From the perspective

of rahmah we put forth a counter narrative that which is to moderate between exclusive discourse and inclusive and thus we see the possibility of transforming digital platforms from which issue of takfīr is put forth into a setting for ethical and humanistic preaching which the Qur'an calls us to do through wise invitation.

This issue at hand is that which of looking into what we may term as Arabic language hate speech in online preaching platforms through the use of corpus linguistic analysis with a rahmah theological base. This study also fills in an empirical void by which it systematily reports out on the trends of hate speech but also puts forth what Islam brings to the table in the international discussion of hate speech and tolerance. By looking at hate speech through a dual lens of language and theology this research puts forth a positive framework for the development of a more ethical and peaceful digital environment.

Literature Review

This research aims to analyze hate speech in Arabic comments on online religious sermons as both a linguistic and theological phenomenon. Pragmatic linguistic analysis is employed to uncover derogatory patterns, labelling, and exclusive speech acts that emerge through terms such as كافر (kafir), بدعة (bid'ah), and مشرك (musyrik). This approach is complemented by corpus linguistic methods, which allow for the systematic identification of the frequency and context of these expressions (Paquot & Gries, 2021). The normative framework is supported by the theology of rahmah (mercy), which emphasises compassion and gentleness in religious communication. Thus, this research not only describes hate speech but also compares it to the fundamental principles of Islam in fostering an ethical digital public space for discourse.

Forensic linguistic studies, such as those conducted by Mubarok (2024) on abusive comments in Indonesian social media, show that hate speech often takes the form of harsh language, identity attacks, and exclusive labels that undermine social cohesion. This finding is relevant, as it highlights the linguistic dimension of hate speech, though this study focuses on Arabic-language comments and their theological implications. This perspective is further enriched by Omran's (2023) research, which compares various machine learning algorithms for detecting hate speech on social media. The results show that advanced algorithms can identify hate speech patterns with high accuracy, reinforcing the existence of linguistic regularities in online hate speech. The difference here lies in the focus of this research, which emphasises a qualitative reading of speech patterns and their theological meanings, rather than evaluating algorithmic performance.

The social dimension of hate speech is further explored by Perera (2023), who examined the behaviour of hate speech propagators on Twitter. This study found that propagators of hate speech exhibited characteristic interaction patterns, including frequent posting and aggressive rhetorical styles. This provides additional context that hate speech is not just a collection of words, but also a social behaviour that is consistently produced. While this perspective resonates with the current research, it differs by focusing on hate speech in online religious comments and its theological consequences. The political dimension is reinforced by Toscano's study (2025) on hate speech among far-right influencers in Spain, where hate speech functions as a tool for public polarization. This study emphasises the polarizing power of hate speech, although the context is different, as it focuses on the Arab religious space and seeks to offer a counter-discourse through the theology of rahmah.

Methods

This research employs a qualitative, descriptive approach to understand hate speech and takfīr in Arabic comments on online sermons (Edmonds & Kennedy, 2020; Gleeson, 2021). This approach is chosen because it allows for an in-depth reading of language patterns and

their theological meanings, beyond mere quantitative data. The focus of the analysis is on comments containing terms like كافر (kafir), بدعة (bid'ah), and مشرك (musyrik), which are viewed not only as linguistic keywords but also as theological expressions that shape social relations and religious identity in digital spaces.

The data sources in this study consist of primary and secondary data (Creswell & Creswell, 2022; Weyant, 2022). Primary data includes public comments on YouTube and TikTok platforms featuring Arabic-language sermons or khutbahs, as well as Qur'anic texts used as normative references to compare hate speech with the principles of rahmah. Secondary data consists of academic literature on Arabic-language hate speech, linguistic and pragmatic studies, and tafsir studies emphasising mercy in Islamic communication. Data collection techniques involve online documentation, literature review, and examination of available Arabic hate speech corpus datasets (Yin, 2013).

Data analysis is conducted using content analysis, combining linguistic categorisation and theological reading (Krippendorff, 2022). Each comment is categorised based on hate speech themes—accusations of kufr (disbelief), bid'ah (innovation), shirk (polytheism), or threats—and analysed for pragmatic patterns and ideological meaning. The nine comments were purposively selected from millions of interactions as representative exemplars of major hate-speech themes, chosen because they reflect distinct categories and show high engagement. The categorised results are presented in frequency tables and interpretative narratives that link empirical data to the theology of rahmah. Validity is ensured through source triangulation, comparing online comments, academic literature, and Qur'anic references (Flick, 2022). This process ensures that the research goes beyond descriptive data, providing critical interpretation that enriches religious linguistic studies and the discourse on Islamic theology in the digital space.

Results and Findings

Hate Speech in the Qur'anic Literature

The Qur'an has long provided a normative foundation for preventing hate speech, belittling others, and negating their existence. The prohibition of mockery, bad naming, and prejudice serves as the basis for Islamic communication ethics. These verses not only affirm moral norms but also act as theological guidelines for interpreting the phenomenon of hate speech in digital spaces.

Table 1. Qur'anic Verses on the Prohibition of Hate Speech

No	Dalil (Surah & Ayat)	Translation
1	QS. Al-Hujurat [49]:11	"O you who have believed, let not a people ridicule [another] people..."
2	QS. Al-Hujurat [49]:12	"O you who have believed, avoid much [negative] assumption..."
3	QS. An-Nahl [16]:125	"Invite to the way of your Lord with wisdom and good instruction..."
4	QS. Fussilat [41]:34	"Good and evil are not equal. Repel [evil] by that which is better..."
5	QS. Al-Isra [17]:53	"And tell My servants to say that which is best..."

The Qur'an lays the ethical foundation for communication by rejecting all forms of mockery, slander, and labeling that degrade human dignity. This prohibition is explicitly discussed in the tafsir of Ibn Kathir, which emphasizes that mockery could lead to social division, while Zamakhsyari viewed it as linguistic insult against the principle of brotherhood (Al-Zamakhsyari, 1995; Katsir, 2018). Tafsir Hamka extends this to modern social spaces, including digital

media, where mockery and labelling such as “kafir” (disbeliever) or “bid'ah” (innovation) are prevalent.

The dimension of hate speech goes beyond mockery; it also includes prejudice and the practice of searching for faults in others. Tafsir Jalalain mentions that the prohibition of tajassus (spying) in Al-Hujurat [49:12] is an effort to maintain social privacy, while al-Maraghi stresses the preventive aspect of preventing the culture of accusations (As-Suyuthi & Al-Mahally, 2015; Maraghi, 1910). Quraish Shihab links this verse to the contemporary context, where practices like doxing and negative framing on social media represent new forms of ghibah (backbiting) and prejudice (Shihab, 2015). This verse demonstrates that hate speech is not a new phenomenon, but a transformed issue within the digital realm.

As an antithesis to hate speech, the Qur'an emphasizes the importance of communication with wisdom and gentleness. Jalalain interprets wisdom as rational argumentation, Ibn Kathir highlights the aspect of gentleness, and Thabathabai stresses love in preaching (Katsir, 2018; Thabathaba’i, 1991). Shihab (2015) connects this with cross-cultural communication, explaining that the call to the way of Allah must be relevant to the social context of the audience. This tafsir demonstrates that Islamic communication strategies rely on peaceful persuasion, not on harsh rhetoric like the prevalent digital takfīr.

The value of dialoguing with mercy is reinforced through the command to repel evil with good. Al-Razi interprets this principle as a foundation for universal conflict resolution, Ibn Kathir calls it a way to tame hostility, while Hamka stresses the effectiveness of gentle approaches in mitigating social conflict (Al-Razi, 2012; Hamka, 2012; Katsir, 2018). Modern tafsir connects this with the phenomenon of online comments: wise responses can break the chain of hate speech and create a healthier communication atmosphere. This verse emphasizes the importance of transforming the digital space from a realm of hate to one of friendship.

The directive for Muslims to choose the best words further serves as a reminder that language is the gateway to harmony. Jalalain emphasizes the preventive function of this verse, Ibn Kathir interprets it as a universal command to maintain speech, and Quraish Shihab stresses its urgency in the era of instant communication that is prone to provocation. According to this tafsir, any harsh comment opens the door for Satan to sow discord. In the digital context, such provocation quickly goes viral, making the command to choose good words even more relevant for maintaining social unity.

Taken together, the classical and modern tafsir demonstrate a consistent view that hate speech is prohibited because it damages social and spiritual relationships. Taken together, the classical and modern tafsir demonstrate a consistent view that hate speech is prohibited because it damages social and spiritual relationships, such as by provoking hostility among believers or eroding the bonds of mutual trust within the community. Classical tafsir emphasizes the normative prohibition, while modern tafsir focuses on the contemporary context. Both combine to provide guidance that hate speech in the form of labeling kafir, bidʿah, or musyrik is an antithesis to Islamic communication.

Thus, the Qur'anic literature not only presents prohibitions but also offers alternatives in the form of communication filled with wisdom, gentleness, and the best of words. Tafsir across time shows the continuity of epistemology: from the prohibition of mockery during the companions’ time to the ethics of digital communication in the 21st century. This framework affirms that the Qur'an provides a strong normative foundation to counter hate speech that is rampant in digital preaching platforms.

Linguistic Patterns of Hate Speech in Online Comments

The analysis of public comments on digital preaching platforms reveals that hate speech appears in explicit linguistic forms, with recurring keywords and pragmatic functions aimed at

excluding other groups. Terms such as *kafir* (disbeliever), *bid'ah* (innovation), and *musyrik* (polytheist) are used not just as labels but also as speech acts that reject the religious legitimacy of others. The following data presents nine representative comments found between July and August 2025.

Table 2. Comments Containing Hate Speech on Digital Preaching Platforms (Translated from Arabic, Source: Brand24)

No	Da'i/Channel Name	Platform	Comment Content (Translation)	Date	Engagement (Views/Reach)	Category
1	Abu Al-Bara' Hilis	YouTube	"The weirdest dream I've ever interpreted: a tour around Hell, Jahannam!"	11-08-2025	Views: 410.3k / Reach: 632.3k	Threat/Curse
2	Walid Ismail Al-Dafi	YouTube	"For everyone deceived by the Shi'a... see what happens to an Iraqi woman."	02-08-2025	Views: 492k / Reach: 75k	Heresy Accusation
3	Durar Al-Utsaimin	YouTube	"Exposing the innovation (bid'ah) of celebrating the Prophet's birthday."	19-08-2025	Reach: 134.7k	Bid'ah Accusation
4	Bayan Inhiraf Al-Sufiyyah	YouTube	"Exposing the deviation (bid'ah) of celebrating the Prophet's birthday."	24-08-2025	Reach: 266.5k	Bid'ah Accusation
5	Mister Tuffah	YouTube	"Filter: Muslim or Kafir #explore."	14-08-2025	Reach: 264.6k	Kafir Accusation
6	Wave Zone Gaming	YouTube	"Don't run, oh Kafir."	14-08-2025	Reach: 264.6k	Kafir Accusation
7	Panda	YouTube	"Don't run, oh Kafir, this is not a song, spread the Qur'an."	24-08-2025	Reach: 266.5k	Kafir Accusation
8	alaaebrahim709	TikTok	"Don't eat the Maulid candy, it's bid'ah."	25-08-2025	Reach: 1.3M	Bid'ah Accusation
9	Muhanjamah	YouTube	"The Kafir seeing your eye repent."	24-08-2025	Reach: 266.5k	Kafir Accusation

The comments above show that hate speech in the digital preaching space does not appear in implicit forms but is open and direct, using religious terms with theological weight. The

accusations of *kafir*, *bid'ah*, and *sesat* (deviant) are used to negate other groups, while threats of hell are employed to instill fear.

Based on Brand24 Observation, YouTube is the platform with the dominant presence of hate speech, followed by TikTok, which although having only one case, generated the highest audience reach, exceeding 1.3 million. This pattern suggests that hate speech not only circulates but also gains massive amplification through the algorithms of digital platforms.

Table 3. Linguistic Patterns of Hate Speech in Digital Preaching Comments

Table 3. Linguistic patterns of hate speech in digital da'wah comments

Linguistic Pattern	Description	Pragmatic Function
Accusation of Infidelity (Kāfir)	The use of the term "kāfir" directly to negate the identity of the other party.	Creating religious exclusion and a moral hierarchy.
Labelling of Innovation (Bid'ah)	The designation of "bid'ah" upon certain religious practices or traditions.	Delegitimising practices with claims of orthodox authority.
Stigmatisation as Polytheist/Deviant	Accusations of doctrinal deviation against other groups.	Consolidating intra-umma polarisation.
Threats of Hell/Curse	Explicit references to hell or eschatological punishment.	Spreading fear and exerting pressure on others.
Muslim vs Non-Muslim Dichotomy	Separation of identity through questions or filters (e.g. "Muslim or kāfir").	Reinforcing identity boundaries and rejecting ambiguity.

The above linguistic patterns demonstrate regularities in the use of language for hate speech. The accusation of infidelity (kāfir) emerges as the most dominant form, functioning not merely as a term, but as a speech act that severs the legitimacy of another's faith. The labelling of bid'ah also plays a crucial role as a strategy for controlling the narrative of orthodoxy, designating certain practices as deviant. The stigmatisation of deviants and polytheists shows continuity with classical *takfīr* practices, while threats of hell are employed to reinforce psychological pressure. The Muslim–non-Muslim dichotomy closes the space for dialogue by simplifying identity into binary categories. Collectively, these patterns affirm that hate speech operates not only at the lexical level but also as a social mechanism that strengthens polarisation (Aldamen, 2023; Castellanos et al., 2024; Solovev, 2023).

The use of the term kāfir constitutes the most prominent form of hate speech within the sphere of digital da'wah. Historically, this term carries a heavy theological dimension as it relates to the status of one's faith. Yet in online comments, the word is employed loosely to negate interlocutors or other groups without authoritative grounds. This pattern illustrates the banality of *takfīr*: a sacred word reduced to a rhetorical weapon of attack. The consequence is not only the offence of the targeted individual but also the creation of an exclusive atmosphere that forecloses dialogue (Matamoros-Fernández, 2021).

The labelling of bid'ah appears as a linguistic strategy to control religious narratives (Sukamto & Pabbajah, 2022). Comments denouncing the celebration of the Prophet's birthday as bid'ah reveal that the term is employed to delegitimise practices perceived as inconsistent with orthodoxy. Pragmatically, this term is not merely evaluative but also an instrument of power to determine which traditions are valid and which must be rejected. In digital spaces, the usage of bid'ah is often reinforced with hashtags, indicating that narratives of orthodoxy are now produced and disseminated in an algorithmic format (Wei, 2024).

Accusations of deviance and polytheism expand the scope of hate speech by linking theological differences to moral stigma. Comments associating the Shīʿa with doctrinal

deviation illustrate how such accusations are deployed to reinforce intra-umma identity boundaries. Such labelling functions to draw a sharp symbolic boundary between Sunni and Shīʿa communities, signalling who is considered "inside" or "outside" orthodox identity and thereby intensifying group-based differentiation within the umma. Pragmatically, such utterances serve to construct a hierarchy, situating the majority group as the guardian of orthodoxy and others as deviants. The consequence is an intensification of internal polarisation, where theological differences are treated as existential threats (Alhazmi et al., 2024).

Threats of hell or curses represent a distinct form of hate speech, as they introduce an eschatological dimension into rhetoric. Comments referring to "a guided tour of Jahannam" or claims that certain individuals will enter hell serve as instruments of psychological pressure. By invoking eternal punishment, such utterances heighten emotional weight, casting recipients not only as excluded from the community of faith but also as eternally condemned. The effect reinforces fear and alienation, while normalising the use of hell as a rhetorical device (Bilal, 2022; Masud et al., 2022; Murphy, 2021).

The Muslim–non-Muslim dichotomy constitutes another pattern that illustrates the simplification of identity. Comments such as "Muslim or kāfir" reveal how digital spaces are frequently employed to impose binary boundaries. Pragmatically, this dichotomy erases grey areas, compelling individuals to adopt rigid identity positions. This phenomenon is perilous as it forecloses interfaith dialogue while simultaneously rejecting the reality of plurality that ought to be acknowledged.

These linguistic patterns illustrate that hate speech manifests not merely in isolated forms but as a coherent discursive system. Accusations of infidelity and labelling of bidʻah serve to uphold claims of orthodoxy, while stigmatisation as deviant and threats of hell reinforce exclusionary dimensions. The dichotomy of identity completes this pattern by enforcing rigid social boundaries. Together, they constitute an ecology of hate speech that functions simultaneously in digital spaces, reinforced by platform algorithms that amplify provocative content.

This phenomenon affirms that hate speech in digital daʻwah comments cannot be understood solely as spontaneous emotional expression (Ghenai, 2025). The consistent patterns and repeated use of terminology demonstrate that hate speech has become an ideological communication strategy (Atikuzzaman, 2023; Farrand, 2023; Okten, 2025). On the one hand, this strategy strengthens the authority of certain groups by delegitimising others. On the other hand, it undermines the possibility of digital daʻwah serving as an arena of learning and fellowship. Hence, the linguistic patterns of hate speech must be read as a structural phenomenon necessitating a counter-discourse grounded in a theology of compassion (*rahmah*) to balance exclusionary narratives.

Theological Implications and the Relevance of *Rahmah* in Hate Speech on Digital Daʻwah Platforms

The phenomenon of hate speech on digital daʻwah platforms demonstrates a sharp contradiction between *takfīr* discourse and the fundamental principles of Islamic communication, which emphasise compassion. Islam situates *rahmah* as both a theological and social foundation, whereas hate speech instead fosters exclusion and polarisation. To clarify this contrast, the following table summarises the categories of hate speech and their theological implications.

Table 4. Theological implications of hate speech categories

No	Hate Speech Category	Theological Implications (*Rahmah*)	Description

1	Accusation of Infidelity (*Kāfir*)	Contrary to *ukhuwah* and the respect for human dignity	*Rahmah* demands that faith be treated inclusively, not as an instrument of exclusion
2	Accusation of Innovation (*Bid'ah*)	Restricting the diversity of worship expressions	The theology of *rahmah* emphasises tolerance and respect for differences in tradition
3	Accusation of Deviance/Polytheism	Reinforcing intra-umma polarisation	*Rahmah* stresses the healing of social relations and the strengthening of brotherhood
4	Threat of Hell	Producing fear and alienation	*Rahmah* encourages guidance with gentleness, not cursing
5	Muslim–Non-Muslim Dichotomy	Closing the space for dialogue and erasing the grey areas of identity	*Rahmah* underscores recognition of plurality and the opening of dialogue space

The accusation of infidelity (*kāfir*) constitutes the most explicit form of hate speech, negating the identity of others. In Islamic theology, the label *kāfir* cannot be used carelessly, as it pertains to the status of faith. When the term is freely employed in digital comments, it not only damages individual dignity but also severs bonds of brotherhood. The perspective of *rahmah* teaches that human dignity must be respected, so the accusation of *kāfir* contradicts the ethics of Islamic communication, which prioritise respect and fraternity.

The category of *bid'ah* illustrates patterns of control over rigid orthodox narratives. This accusation is often used to reject certain acts of worship, even when those practices function as living religious expressions within society. Theologically, *rahmah* demands tolerance towards diversity of faith expression, so long as it does not contradict the principle of *tawḥīd* (Naldi et al., 2025). Therefore, the indiscriminate use of the *bid'ah* label aligns more closely with social exclusion than with the spirit of compassion taught in Islam.

Accusations of deviance or polytheism reveal attempts to reinforce intra-umma polarisation. Online comments accusing other groups of deviation serve to construct a moral hierarchy that affirms the dominance of particular groups. The theology of *rahmah* rejects this approach by emphasising the restoration of social relations. It not only prohibits stigmatisation but also encourages the creation of dialogical spaces where differences can be transformed into the strength of *ukhuwah*.

Threats of hell represent a discursive pattern that introduces an eschatological dimension into hate speech (Ibrohim & Budi, 2023). Comments employing hell as an instrument of cursing produce fear and alienation, while normalising the use of eschatological symbols to pressure others. The perspective of *rahmah* instead underscores the importance of guidance delivered with gentleness, so that people may return to the path of goodness. Speech that builds rather than curses is more consistent with the compassionate mission of da'wah.

The Muslim and non-Muslim dichotomy illustrates how hate speech simplifies identity into binary boundaries. Online comments that compel individuals to choose "Muslim or *kāfir*" foreclose the possibility of dialogue and erase the reality of plurality. The theology of *rahmah* requires recognition of diversity and openness to interfaith encounters. When identity is forced into a binary, *rahmah* emerges to defend diversity as part of divine will (Baider, 2023).

Overall, the patterns of hate speech appearing in digital da'wah spaces reveal that such utterances operate to exclude, intimidate, and divide. Conversely, the perspective of *rahmah* offers an inclusive theological alternative, emphasising respect, guidance, and recognition of plurality. Thus, *rahmah* is not merely a moral concept but a theological framework capable of

functioning as an effective counter-discourse against hate speech in digital spaces, while simultaneously broadening the function of da'wah as a means of social peace.

Discussion

The phenomenon of hate speech within digital da'wah spaces demonstrates the close interrelation between language, theology, and power. The patterns of accusation of infidelity (*kāfir*), bid'ah, deviance, and threats of hell are not merely coarse words but constitute communication strategies that regulate who is deemed legitimate within the community of faith. Within a pragmatic framework, such utterances function as exclusionary speech acts that generate social hierarchies. The perspective of the theology of *rahmah* offers a fundamental critique of this pattern, emphasising that religious communication ought to cultivate fellowship and compassion rather than hostility.

Corpus-based analysis demonstrates that hate speech in da'wah spaces is not incidental but systematic. The consistent use of the terms *kāfir* and bid'ah across various comments indicates a normalisation of hate language. Pragmatic linguistic theory assists in reading such utterances as performative acts that expel particular groups from the sphere of faith (Baumgarten, 2022). This illustrates that hate speech does not merely reflect emotion but also produces discriminatory social realities.

The presence of digital platforms further amplifies the effects of hate speech, as algorithms tend to promote provocative content in order to achieve higher engagement (Baykal et al, 2025). When comments containing harsh accusations go viral, exclusive language is multiplied, liked, and disseminated. From the perspective of the theology of *rahmah*, this phenomenon is paradoxical: a medium of da'wah, which ought to expand goodness, instead becomes a channel of exclusion. Accordingly, counter-discourse grounded in compassion becomes urgent to restore the function of digital spaces as healthy arenas of da'wah.

Hate speech containing threats of hell reveals an eschatological dimension that is misappropriated. Rhetoric concerning eternal punishment is employed to intimidate and exclude rather than to guide. Modern exegesis emphasises that *rahmah* is the essence of da'wah, in which threats function only contextually, not as rhetorical weapons. The transformation of hell rhetoric into arbitrary curses demonstrates the banality of eschatology, which is counterproductive to the spiritual aims of Islam.

The Muslim–non-Muslim dichotomy displayed in online comments shows how plurality is rejected through the simplification of identity. Within the framework of the theology of *rahmah*, the recognition of diversity is part of the core teachings of Islam. Digital spaces that impose binary identities obscure the inclusive message of Islam. Hence, communicative strategies of *rahmah* must emphasise openness to dialogue so that digital da'wah platforms do not become trapped within exclusive narratives (Ul Hassan & Ahmad, 2021; Zakirov, 2022).

The practical implications of these findings are the necessity of reorienting digital da'wah from exclusive rhetoric towards inclusive rhetoric (Murphy, 2021). Da'wah based on *rahmah* can construct a positive and productive communicative atmosphere. This approach is not only a theological solution but also a social strategy for reducing intra-umma polarisation. By prioritising compassion, digital da'wah can serve as a medium for strengthening interfaith solidarity amidst the complexities of modern society.

A rereading of hate speech through the lens of *rahmah* also demonstrates that digital space is not neutral but an arena of value contestation. If not filled with discourses of compassion, it will be dominated by exclusive rhetoric. Consequently, constructing counter-discourse grounded in rahmah is not only a theological responsibility but also a public communication strategy vital for sustaining social harmony, including in global digital environments where hate speech circulates transnationally.

This research is connected to the study by Mubarok (2024), which employed a forensic linguistic approach to examine hate speech in Indonesian social media. The difference is that Mubarok's study focuses more on the forensic aspects of language, while this research incorporates the theological dimension of *rahmah* as counter-discourse. This comparison shows that linguistic analysis can be enriched by theological perspectives to build a more comprehensive understanding.

The findings also relate to Toscano's (2025) study on hate speech by right-wing political influencers in Spain. Toscano emphasised the relationship between hate speech and political polarisation, whereas this research highlights its relation to theological polarisation. The similarity lies in the function of hate speech in reinforcing the dominance of certain groups, while the difference lies in the arena of contestation: secular politics in Europe versus digital da'wah in Islam.

The novelty of this research lies in the integration of corpus analysis of hate speech with the theology of *rahmah* as counter-discourse. Hitherto, studies on hate speech have tended to focus on algorithmic detection, linguistic patterns, or political impacts, while the theological dimension has often been marginalised. By positioning *rahmah* as a counter-framework, this study offers a unique analytical model in which digital space is read not only as a linguistic arena but also as a theological one. This approach underscores that combating hate speech is not merely a technical task but also a theological mission to bring compassion into public communication.

Conclusion

This research demonstrates that hate speech within digital da'wah spaces manifests in consistent linguistic patterns—accusations of *kāfir*, bid'ah, deviance, and threats of hell—that pragmatically function to marginalise and construct moral hierarchies. Corpus-based analysis confirms that hate speech is not spontaneous emotion but an ideological communication strategy reinforced by platform algorithms. The perspective of the theology of *rahmah* becomes a relevant counter-discourse as it affirms the values of compassion, wisdom, and respect for plurality, thereby redirecting digital da'wah towards its mission of brotherhood and peace. While rahmah provides a compelling theological alternative to exclusionary discourse, its application may encounter resistance from groups that prioritise more literalist or exclusivist doctrinal positions. Acknowledging this tension is essential, as the reception of counter-narratives in digital religious spaces reflects broader theological diversity within the Muslim community.

Digital da'wah must be directed towards inclusive narratives grounded in *rahmah*, by developing ethical communication guidelines, moderating comments, and producing alternative content that emphasises compassion and respect for differences. This research remains limited to corpus analysis of comments within a specific period, so future studies are recommended to expand across countries or sects and to integrate big data methods with Islamic theological frameworks in order to strengthen academic understanding as well as practical strategies against hate speech in digital spaces.

References

Al-Razi, F. al-D. (2012). *Mafatih al-Ghaib al-Mushtahar Bi al-Tafsir al-Kabir* (7th ed.). Daar al-Hadith.

Al-Zamakhsyari. (1995). *al-Kasyaf an Haqoiqi al-Tanzil wa Uyuuni al-Aqowili fi al-wujuuh al-Takwil*. Daar al-Kutub al-'Ilmiah.

Aldamen, Y. (2023). Xenophobia and Hate Speech towards Refugees on Social Media:

Reinforcing Causes, Negative Effects, Defense and Response Mechanisms against That Speech. *Societies*, *13*(4). https://doi.org/10.3390/soc13040083

Alhazmi, A., Mahmud, R., Idris, N., Mohamed Abo, M. E., & Eke, C. I. (2024). Code-mixing unveiled: Enhancing the hate speech detection in Arabic dialect tweets using machine learning models. *PLOS ONE*, *19*(7), e0305657. https://doi.org/10.1371/journal.pone.0305657

As-Suyuthi, J., & Al-Mahally, J. M. I. A. (2015). Tafsir Al-Jalalain. *Tafsir Jalalain (Terjemah)*.

Atikuzzaman, M. (2023). Hate speech in social media: personal experiences and perceptions of university students in Bangladesh. *Global Knowledge Memory and Communication*, *72*(8), 835–851. https://doi.org/10.1108/GKMC-01-2022-0014

Baider, F. (2023). Accountability Issues, Online Covert Hate Speech, and the Efficacy of Counter-Speech. *Politics and Governance*, *11*(2). https://doi.org/10.17645/pag.v11i2.6465

Baumgarten, N. (2022). Contrastive pragmatics. In *The Routledge Handbook of Translation and Methodology*. https://doi.org/10.4324/9781315158945-14

Baykal, L., Nafus, D., & Alvarado, J. (2025). *YouTube recommendations reinforce negative emotions: Auditing algorithmic bias with emotionally-agentic sock puppets.* arXiv. https://arxiv.org/abs/2501.15048

Bilal, M. (2022). Context-Aware Deep Learning Model for Detection of Roman Urdu Hate Speech on Social Media Platform. *IEEE Access*, *10*, 121133–121151. https://doi.org/10.1109/ACCESS.2022.3216375

Castellanos, M., Wettstein, A., Wachs, S., & Bilz, L. (2024). Direct and indirect effects of social dominance orientation on hate speech perpetration via empathy and moral disengagement among adolescents: A multilevel mediation model. *Aggressive Behavior*, *50*(1). https://doi.org/10.1002/ab.22100

Charfi, A., Besghaier, M., Akasheh, R., Atalla, A., & Zaghouani, W. (2024). Hate speech detection with ADHAR: a multi-dialectal hate speech corpus in Arabic. *Frontiers in Artificial Intelligence*, *7*(May). https://doi.org/10.3389/frai.2024.1391472

Creswell, J. W., & Creswell, J. D. (2022). Research Design: Qualitative, Quantitative, and Mixed Methods Approaches 6th edition - John W. Creswell, J. David Creswell - SAGE. In *SAGE Publications, Inc.*

Dalimunthe, M. A., Djuniardi, R., & Siswanda, D. (2024). Pengaruh Media Sosial terhadap Perubahan Pola Komunikasi. *Ahkam*, *3*(1), 67–76. https://doi.org/10.58578/ahkam.v3i1.2486

Edmonds, W. A., & Kennedy, T. D. (2020). An Applied Guide to Research Designs: Quantitative, Qualitative, and Mixed Methods. In *An Applied Guide to Research Designs: Quantitative, Qualitative, and Mixed Methods*. https://doi.org/10.4135/9781071802779

Farid, A. S., Rahmah, R. A., Daulay, I., & Desiana. (2024). Theology and Ethics in Values-Based Journalism Communicating Islamic Perspectives. *Pharos Journal of Theology*, *105*(2). https://doi.org/10.46222/pharosjot.105.226

Farrand, B. (2023). 'Is This a Hate Speech?' The Difficulty in Combating Radicalisation in Coded Communications on Social media Platforms. *European Journal on Criminal Policy and Research*, *29*(3), 477–493. https://doi.org/10.1007/s10610-023-09543-z

Flick, U. (2022). Revitalising Triangulation for Designing Multi-perspective Qualitative Research. In *The SAGE Handbook of Qualitative Research Design*. https://doi.org/10.4135/9781529770278.n40

Ghenai, A. (2025). Exploring hate speech dynamics: The emotional, linguistic, and thematic impact on social media users. *Information Processing and Management*, *62*(3).

https://doi.org/10.1016/j.ipm.2025.104079

Gleeson, D. (2021). Research Questions and Research Design. In *Educational Research and Inquiry*. https://doi.org/10.5040/9781474243834.ch-005

Hamka. (2012). *Tafsir al-Azhar* (4th ed.). Pustaka Nasional.

Ibrohim, M. O., & Budi, I. (2023). Hate speech and abusive language detection in Indonesian social media: Progress and challenges. In *Heliyon* (Vol. 9, Issue 8). https://doi.org/10.1016/j.heliyon.2023.e18647

Karjo, C. H., & Ng, A. (2020). Hate Speech Propaganda from and against Muslims in Facebook Posts. *International Journal of Cyber Criminology*, *14*(2). https://doi.org/10.5281/zenodo.4766989

Katsir, I. (2018). *Tafsir al-Quran al-Adzim* (7th ed.). Insan Kamil.

Krippendorff, K. (2022). The Logic of Content Analysis Designs. In *Content Analysis: An Introduction to Its Methodology*. https://doi.org/10.4135/9781071878781.n5

Lin, P., & Adolphs, S. (2023). Corpus linguistics. In *The Routledge Handbook of Applied Linguistics*. https://doi.org/10.4324/9781003082644-25

Maraghi, M. M. (1910). *Tafsir al-Maraghi*. Dar al-Fikr.

Masud, S., Bedi, M., Khan, M. A., Akhtar, M. S., & Chakraborty, T. (2022). Proactively Reducing the Hate Intensity of Online Posts via Hate Speech Normalization. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. https://doi.org/10.1145/3534678.3539161

Matamoros-Fernández, A. (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

Mubarok, Y. (2024). Abusive Comments (Hate Speech) on Indonesian Social Media: A Forensic Linguistics Approach. *Theory and Practice in Language Studies*, *14*(5), 1440–1449. https://doi.org/10.17507/tpls.1405.16

Murphy, A. (2021). Political rhetoric and hate speech in the case of Shamima Begum. *Religions*, *12*(10). https://doi.org/10.3390/rel12100834

Naeem, T. (2022). Muslims as Terrorists: Hate Speech against Muslims. *International Journal of Islamic Thought*, *22*. https://doi.org/10.24035/ijit.22.2022.245

Naldi, A., Ginting, L. D. C. U., Rambe, R. F. A. K., & Damanik, F. H. S. (2025). Tawhid of The Sky and Tawhid of The Earth: Theological Reflections of the People of Medan Post Natural Disaster. *Pharos Journal of Theology*, *106*(2). https://doi.org/https://doi.org/10.46222/pharosjot.106.214

Obermaier, M., Schmuck, D., & Saleem, M. (2023). I'll be there for you? Effects of Islamophobic online hate speech and counter speech on Muslim in-group bystanders' intention to intervene. *New Media and Society*, *25*(9). https://doi.org/10.1177/14614448211017527

Okten, A. (2025). On Commemorating Hrant Dink: Affective Nationalism, Hate Speech, and Digital News Media Users. *Journalism and Mass Communication Quarterly*, *102*(2), 500–522. https://doi.org/10.1177/10776990221093920

Omran, E. (2023). A comparative analysis of machine learning algorithms for hate speech detection in social media. *Online Journal of Communication and Media Technologies*, *13*(4). https://doi.org/10.30935/ojcmt/13603

Paquot, M., & Gries, S. T. (2021). A Practical Handbook of Corpus Linguistics. In *A Practical Handbook of Corpus Linguistics*. https://doi.org/10.1007/978-3-030-46216-1

Perera, S. (2023). A comparative study of the characteristics of hate speech propagators and their behaviours over Twitter social media platform. *Heliyon*, *9*(8). https://doi.org/10.1016/j.heliyon.2023.e19097

Prati, F., Policardo, G. R., Hewstone, M., & Rubini, M. (2024). How Positive and Negative Intergroup Contact May Shape the Communication of Discrimination Toward Migrants. *Journal of Language and Social Psychology*, *43*(3). https://doi.org/10.1177/0261927X241237260

Shannag, F., Hammo, B. H., & Faris, H. (2022). The design, construction and evaluation of annotated Arabic cyberbullying corpus. *Education and Information Technologies*, *27*(8). https://doi.org/10.1007/s10639-022-11056-x

Shihab, M. Q. (2015). *Tafsir al-Mishbah* (2nd ed.). Lentera Hati.

Solovev, K. (2023). Moralized language predicts hate speech on social media. *Pnas Nexus*, *2*(1). https://doi.org/10.1093/pnasnexus/pgac281

Sukamto, F. K., & Pabbajah, M. (2022). *Sentiment prevalence on jihad, caliphate, and bid'ah among Indonesian students*. Cogent Social Sciences, 8(1), 2054532. https://doi.org/10.1080/23311886.2022.2054532

Thabathaba'i, S. M. H. (1991). *Al-Mizan fi Tafsir al-Qur'an*. Muassasah al-A'lami li al-Mathbu'at.

Toscano, G. B. (2025). Violence, Hate Speech, and Polarisation in Far-Right Political Influencers in Spain. *Javnost*, *32*(1), 92–114. https://doi.org/10.1080/13183222.2025.2469032

Ul Hassan, M. A., & Ahmad, H. I. (2021). A-1 Jurisprudencial Differences Between Sahabah (R.A), Causes and Effects: An Analytical Study. *Al-Aijaz Research Journal of Islamic Studies & Humanities*, *5*(1). https://doi.org/10.53575/a1.v5.01(21).1-11

Wei, M. (2024). Mediatization of Religion and Its Impact on Youth Identity Formation in Contemporary China. *Religions*, *15*(3). https://doi.org/10.3390/rel15030268

Weyant, E. (2022). Research Design: Qualitative, Quantitative, and Mixed Methods Approaches, 5th Edition. *Journal of Electronic Resources in Medical Libraries*, *19*(1–2). https://doi.org/10.1080/15424065.2022.2046231

Yin, R. K. (2013). Case study research: Design and methods. *Applied Social Research Methods Series*, *18*(2). https://doi.org/10.1097/00001610-199503000-00004

Zaghouani, W., & Biswas, M. R. (2025). *An Annotated Corpus of Arabic Tweets for Hate Speech Analysis*. Cornell University. https://doi.org/https://doi.org/10.48550/arXiv.2505.11969

Zakirov, A. A. (2022). F. Isac's Qur'anic Hermeneutics in Context of Islamic Liberation Theology. *Islam in the Modern World*, *17*(4). https://doi.org/10.22311/2074-1529-2021-17-4-63-78.

***Conflict of Interest Statement**: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.*