Beyond Binary: Revealing Variations in Islamophobic Content with Hierarchical Multi-Class Classification

Esraa Aldreabi[1], Khawlah M. Harahsheh[2], Mukul Dev Chhangani[1], Chung-Hao Chen[2], Jeremy Blackburn[1]

[1]Binghamton University, Binghamton, NY, {ealdrea1, mchhang1, jblackbu}@binghamton.edu
[2]Old Dominion University, Norfolk, VA, {Khara001, exchen}@odu.edu

Abstract

In the digital age, the rise of Islamophobia-marked by an irrational fear or discrimination against Islam and Muslims-has emerged as a pressing issue, especially on social media platforms. In this paper we employs a multi-class classification system, moving beyond traditional binary models. We categorize Islamophobic content into three main classes and various subclasses, covering a range from subtle biases to explicit incitement. Comparative analysis of data from Reddit and Twitter illuminates the distinct prevalence and types of Islamophobic content specific to each platform. This paper deepens our understanding of digital Islamophobia and provides insights for crafting targeted online counter-strategies. Additionally, it highlights the role of machine and deep learning in detecting and addressing Islamophobic content, emphasizing their significance in resolving complex social issues in the digital sphere.

Introduction

The digital era reshapes how we communicate, positioning social media sites like Reddit and Twitter as key influencers of public discourse. These platforms are adept at facilitating discussions and fostering relationships, yet they also become venues for the proliferation of hate speech (Burnap and Williams 2015). Islamophobia, defined by animosity or bias against Islam or Muslims, poses a profound challenge (Vidgen and Yasseri 2020). This context brings to light the vigorous debates around the obligations of social media in addressing prejudices and the success of their moderation policies in curbing toxic behavior without heightening the danger of radicalization (Ali et al. 2021; Horta Ribeiro et al. 2021). Our research investigates the manifestation and spread of Islamophobic content on Reddit and Twitter. The distinct nature of these platforms – Reddit with its in-depth, community-driven discussions, and Twitter with its rapid content turnover provides contrasting lenses through which to examine how Islamophobic discourse is presented and propagated.

Central to our study is a hierarchical multi-class classification framework that categorizes text data into a spectrum of Islamophobic expressions. This framework includes three primary classes: A: Non-Islamophobic, B: Implicit or Mild Islamophobia, and C: Explicit or Severe Islamophobia. Each of these is further divided into subclasses, such as A1: Islam-Related Content, A2: Religion Debate, B1: Stereotyping, B2: Misunderstanding/Negative Assumptions, C1: Hate Speech, and C2: Incitement/Discrimination. This detailed labeling system is pivotal for a nuanced analysis, enabling us to annotate and classify text data across various intensities and themes of Islamophobia. By acknowledging the inherent ambiguity in defining Islamophobia (Sealy 2021; Aldreabi and Blackburn 2024), our approach captures the broad spectrum of Islamophobic expressions, from subtle biases to outright hate speech. This comprehensive analysis is crucial for understanding the diverse ways in which Islamophobia manifests on these platforms. It also aids in differentiating between various levels of Islamophobic content, thus maintaining a balance between identifying hate speech and preserving free and civil discourse on religious matters.

Our findings are aimed at guiding technologists, policymakers, and researchers towards creating more inclusive and responsible digital spaces. Understanding the current landscape of online hate speech, specifically Islamophobia, is essential for developing targeted strategies to combat such biases on social media platforms. Through this research, we hope to contribute to the ongoing efforts to foster digital environments where dialogue is enriched by diversity and free from the blight of hate and prejudice.

Related Work

In exploring the landscape of Islamophobia and hate speech, (Efstratiou et al. 2022) undertook a detailed historical exploration of Reddit's political forums, identifying a range of posting behaviors and the complex interplay between participation in echo chambers, polarization, and hostility among differing political ideologies. In a similar vein, (Squire 2019) investigated the Islamophobic sentiments prevalent within far-right factions on Facebook, and (Balcı, Sirivianos, and Blackburn 2023) shifted the focus towards examining the dynamics of extremism within left-wing groups. Furthermore, (Soral, Liu, and Bilewicz 2020) identified differences in exposure to Islamophobic content between users of social media versus traditional mass media, pointing to social media as a key arena for the dissemination of hate speech. (Vidgen and Yasseri 2020) in-

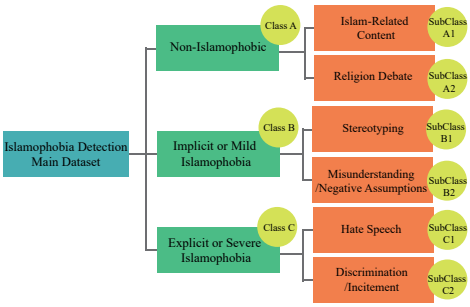


Figure 1: Hierarchical classification structure for Islamophobia detection.

troduced a classification system for Islamophobia that goes beyond simple binary distinctions, outlining categories of non-Islamophobic, weak Islamophobic, and strong Islamophobic content. (Mehmmod, Kaleem, and Siddiqi 2022) made a notable contribution to the evolving landscape of hate speech research by leveraging deep learning techniques specifically tailored for detecting Islamophobic hate speech. Advancements in technology for detecting hate speech, such as the deep learning techniques employed by (Mehmmod, Kaleem, and Siddiqi 2022) and the transfer learning approach of (Belal, Ullah, and Khan 2022). The work of (Albadi, Kurdi, and Mishra 2018) in the Arabic Twittersphere, utilizing classification models to distinguish religious hate speech, and (Khan and Phillips 2021)'s approach to multilingual data classification for detecting Islamophobia, reflect ongoing improvements in hate speech identification. (Aldreabi, Lee, and Blackburn 2023) provided analysis on Reddit using topic modeling to uncover specific Islamophobic topics, alongside (Ahmanideen and Iner 2024)'s study linking online hate speech with offline Islamophobic incidents, highlight the interconnectedness of digital platforms and real-world consequences. (Chandra et al. 2021; Tahmasbi et al. 2021) noted the surge in Islamophobia and hate speech during the COVID-19 pandemic. Lastly, (Aldreabi and Blackburn 2024) emphasizes the need for nuanced detection models to differentiate Islamophobia from valid criticism, aiming to identify hate speech without discrimination.

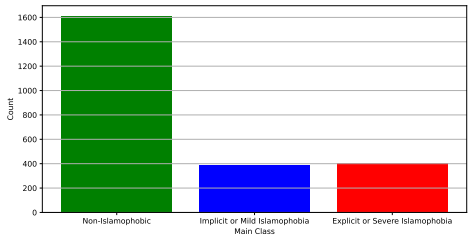
Dataset Labeling and Annotation

To deepen our grasp of Islamophobia, we are crafting a hierarchical classification system that goes beyond simple binary labeling. This advanced structure breaks down into three primary classes: A: Non-Islamophobic, B: Implicit or Mild Islamophobia, and C: Explicit or Severe Islamophobia. Each class then subdivides into specific categories, enabling a finer analysis of the nuances in Islamophobic expressions. Our scheme captures the spectrum of Islamophobia, ranging in severity and type, to classify content more accurately and encompass a broader array of Islamophobic expressions. We detail this classification approach below and provide a visual in Figure 1.

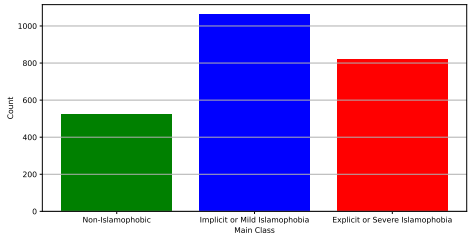
- ***Class A: Non-Islamophobic Content:*** This class is vital in our classification schema, as it represents content that discusses Islam or Muslims without conveying Islamophobic sentiment.
  - ***Subclass A1: Islam-Related Content:*** This subclass encompasses texts on Islamic practices, beliefs, or the Muslim community without Islamophobic tones. Examples include:
    - "Eid Mubarak to my Muslim friends. May your day be filled with joy!"
  - ***Subclass A2: Religion Debate:*** This subclass covers content engaging in respectful, analytical religious debates within freedom of expression bounds. It aims to separate constructive discussions from Islamophobic content, addressing the intricacies of religious discourse. During re-labeling, we found instances marked as "Islamophobic" in the original datasets that actually embodied this respectful debate, highlighting the need to differentiate critical engagement from Islamophobia. This approach emphasizes a nuanced understanding of Islamophobia, valuing diverse perspectives while guarding against hate speech (Aldreabi and Blackburn 2024). Examples in this subclass might include:
    - "People can't accept that the Quran has a homophobic verse, thus they make up nonsense to justify Islam being LGBTQ+ friendly Quran 7:81"
- ***Class B: Implicit or Mild Islamophobia:*** This class captures content that subtly or indirectly expresses Islamophobic sentiments. It is critical for identifying content that may not be overtly hateful but still perpetuates harmful stereotypes or assumptions about Islam and Muslims.
  - ***Subclass B1: Stereotyping:*** This subclass features content with generalizations or stereotypes about Muslims or Islam, endorsing unjustified stereotypes and derogatory language. Such content perpetuates harmful stereotypes, like falsely associating Muslims with terrorism, negatively affecting public perceptions and attitudes (Smith 2014; Cervi, Tejedor, and Gracia 2021). Examples:
    - "All Muslims are terrorists since childhood."
  - ***Subclass B2: Misunderstanding/Negative Assumptions:*** This subclass addresses content with misunderstandings or negative assumptions about Islam or its followers, including misconceptions about Islamic practices or generalized negative biases. Examples include incorrect generalizations about cultural practices among Muslims. Such content frequently uses mocking language to demean individuals of the Muslim faith (Cervi, Tejedor, and Gracia 2021; Nadal et al. 2012). Examples:
    - "Muslims don't believe in modern science. They only follow what's in their religious texts."
- ***Class C: Explicit or Severe Islamophobia:*** Class C addresses the most severe forms of Islamophobia, encompassing explicit hate speech, discrimination, and incitement against Muslims or Islam.
  - ***Subclass C1: Hate Speech:*** This subclass pertains to content classified as hate speech, characterized by

Table 1: Distribution of subclasses on Reddit and Twitter.

Platform	Class	Subclass	Count
Reddit	A:Non-Islamophobic	A1:Islam-Related Content	1,087
Reddit	A:Non-Islamophobic	A2:Debate Religion	527
Reddit	B:Implicit or Mild Islamophobia	B1:Stereotyping	234
Reddit	B:Implicit or Mild Islamophobia	B2:Misunderstanding/Negative Assumptions	154
Reddit	C:Explicit or Severe Islamophobia	C1:Hate Speech	352
Reddit	C:Explicit or Severe Islamophobia	C2:Discrimination/Incitement	54
Reddit		Total	2,408
Twitter	A:Non-Islamophobic	A1:Islam-Related Content	330
Twitter	A:Non-Islamophobic	A2:Debate Religion	194
Twitter	B:Implicit or Mild Islamophobia	B1:Stereotyping	841
Twitter	B:Implicit or Mild Islamophobia	B2:Misunderstanding/Negative Assumptions	221
Twitter	C:Explicit or Severe Islamophobia	C1:Hate Speech	512
Twitter	C:Explicit or Severe Islamophobia	C2:Discrimination/Incitement	310
Twitter		Total	2,408



(a) Reddit data.



(b) Twitter data.

Figure 2: Proportion of each main class in Reddit and Twitter data.

derogatory language or offensive comments aimed directly at Muslims or Islam. Examples of this subclass:

- "Muslims are nothing but fucking criminals, spreading hatred and violence everywhere they go."

– ***Subclass C2: Discrimination/Incitement:*** This subclass targets content with threatening language and overt discrimination, like advocating for banning Muslims (Patel 2022). It features direct threats, explicit harmful intentions toward Muslims, or portrays Islam as incompatible with Western culture (Evolvi 2018). Representing the extreme of Islamophobia, it highlights content that both discriminates and directly threatens based on faith and culture. Examples of this subclass include:

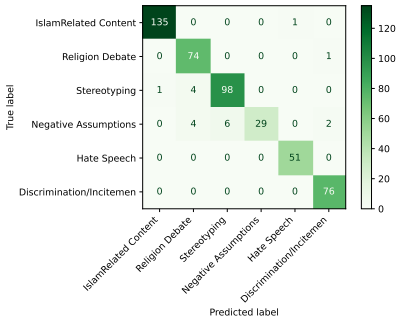
- "#BanIslam Kill Muslims save humanity."

Following the establishment of our classification system, we sourced datasets from Reddit (Aldreabi, Lee, and Blackburn 2023) and Twitter (Khan and Phillips 2021), initially labeled as "Islamophobic" or "Non-Islamophobic". Three independent annotators undertook the task of re-labeling these datasets according to our scheme. To validate the reliability of the re-labeled data, we calculated the Fleiss' Kappa score, which yield a score of 0.86 for Twitter and 0.74 for Reddit. These figures indicate a near-perfect and substantial consensus among annotators, respectively. The comparatively lower agreement score for Reddit can be attributed to the longer text length and the complex nature of content within the "Debate Religion" subclass. The frequent occurrence of "Debate Religion" content in the Reddit dataset introduces unique challenges, as these discussions often navigate the delicate boundary between critique and offense, leading to diverse interpretations by annotators. Such variability underlines the subjective aspect of classifying content that entails nuanced and intricate debates. Our labeling system facilitates structured annotation of text data, enabling categorization by Islamophobia's degrees and types. The annotation results, showing the distribution across classes and subclasses, are detailed in Table 1.

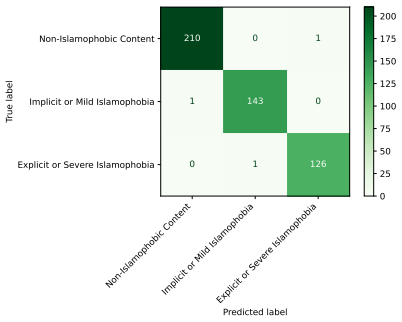
Classification Models and Performance Evaluation

We use the hateBERT model (Caselli et al. 2021), originally designed for hate speech detection, adapting it for the task of classifying Islamophobic content. This involves modifying hateBERT's classifier layer to fit our dataset's unique configuration (three classes and six distinct subclasses). We encode the labels and divide the dataset into training (80%), validation (10%), and test (10%) sets to thoroughly evaluate the model's ability to generalize.

In the training phase, we actively fine-tune the model across ten epochs. We use the AdamW optimizer at a learning rate of 2e-5 and use the CrossEntropyLoss function to minimize classification errors. By optimizing the model with a batch size of 32, we ensure efficient parameter updates and enhanced accuracy. We conduct the analysis in two main stages. Initially, we concentrate on subclassification, fine-tuning the model with a dataset divided into six specific categories to capture the various expressions of Islamophobia. We evaluate the model's performance on this subclassified dataset, and the confusion matrix(Figure 3a) visually demonstrates the model's precision in accurately classifying each subclass. Building on these results, we consolidate the subclasses into three main hierarchical classes: Non-Islamophobic, Implicit or Mild Islamophobia, and Explicit or Severe Islamophobia. The test confusion matrix in



(a) Confusion matrix for subclass classification.



(b) Severity-Based confusion matrix.

Figure 3: Confusion matrices for Islamophobic content classification.

Figure 3b for this broader classification demonstrates the model's high accuracy in categorizing content into these general severity levels.

Comparing Islamophobic Content on Reddit and Twitter

We undertake a review of how Islamophobic content is distributed across Reddit and Twitter. By examining Figures 2a and 2b, we pinpoint noticeable variations in the frequencies of "Non-Islamophobic," "Implicit or Mild Islamophobia," and "Explicit or Severe Islamophobia" categories across these platforms. The subclass distribution of Islamophobic content, detailed in Table 1, shows patterns reflecting each platform's discourse. On Reddit, Class A content, particularly "Islam-Related Content" (1087 instances) and "Debate Religion" (527 instances), dominates, suggesting a trend towards more in-depth and potentially constructive discussions. However, "Implicit or Mild Islamophobia," with subclasses like "Stereotyping" (234 instances) and "Misunderstanding/Negative Assumptions" (154 instances), and "Explicit or Severe Islamophobia," marked by "Hate Speech" (352 instances) and "Discrimination/Incitement" (54 instances), are also notably present. Twitter presents a different landscape. It houses a considerable amount of "Non-Islamophobic" content, with "Islam-Related Content" at 330 instances and "Debate Religion" at 194 instances, but more cases of Islamophobic content are evident "Implicit or Mild Islamophobia," represented by "Stereotyping" (841 instances) "Misunderstanding/Negative Assumptions" (221 instances), and "Explicit or Severe Islamophobia," indicated by "Hate Speech" (512 instances) and "Discrimination/Incitement" (310 instances), are more prevalent on Twitter, suggesting a broader and more intense manifestation of Islamophobia on this platform. The chi-squared test results ($\chi^2 = 1009.93$, $p < 0.001$) reinforce these observations, indicating a statistically significant association between the platform type and the distribution of subclasses, reflecting the distinct characteristics and user dynamics of each platform. Delving deeper, we notice the divergent nature of Islamophobic discourse. Reddit, with its forum-like structure, fosters longer posts (averaging 281.38 characters), which may contribute to a higher occurrence of content categories that necessitate elaborate expression. Twitter naturally encourages more succinct and direct communication, averaging 97.56 characters per post. This brevity might lead to oversimplification and, in some cases, more blatant expressions of Islamophobic sentiments, as seen in the higher counts of "Explicit or Severe Islamophobia." The platform's design, which emphasizes quick sharing and viral spread of content, might also contribute to the more widespread dissemination of severe forms of Islamophobia, including hate speech and incitement. These findings highlight distinct patterns in Islamophobic content across platforms: Reddit's discussions are more ideologically oriented, while Twitter's Islamophobic content often targets individuals directly, influenced by the platform's design favoring concise, direct communication.

Conclusion

Our study presents an advancement in the understanding of Islamophobic content on Reddit and Twitter, identifying unique patterns on each platform. Reddit displays a mix of in-depth discussions and Islamophobic content, ranging from stereotypes to hate speech, indicating a blend of constructive and negative discourse. Conversely, Twitter exhibits a higher prevalence of Islamophobia, both implicit and explicit, facilitated by its design that promotes brief communication. Our analysis reveals the problematic linkage of Muslims to terrorism and notable instances of explicit hate speech, with platform-specific focuses: ideological critiques on Reddit and personal attacks on Twitter. These findings highlight the intricate manifestation of Islamophobia online, emphasizing the importance of nuanced interventions to address it, providing insights for researchers, policymakers, and platforms to combat online hate speech. We face challenges such as category overlap, unclear subclass distinctions, and the complex cultural and contextual dimensions of Islamophobia. Future research should expand on our methods, explore more platforms, and use a broader range of data to deepen our understanding of Islamophobia and how to combat it.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. IIS-2046590.

References

Ahmanideen, G., and Iner, D. 2024. The interaction between online and offline islamophobia and anti-mosque campaigns: The literature review with a case study from an anti-mosque social media page. *Sociology Compass* 18(1):e13160.

Albadi, N.; Kurdi, M.; and Mishra, S. 2018. Are they our brothers? analysis and detection of religious hate speech in the arabic twittersphere. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 69–76.

Aldreabi, E., and Blackburn, J. 2024. Enhancing automated hate speech detection: Addressing islamophobia and freedom of speech in online discussions. In *Proceedings of the 2023 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ASONAM '23, 644–651. New York, NY, USA: Association for Computing Machinery.

Aldreabi, E.; Lee, J. M.; and Blackburn, J. 2023. Using deep learning to detect islamophobia on reddit. *The International FLAIRS Conference Proceedings* 36.

Ali, S.; Saeed, M. H.; Aldreabi, E.; Blackburn, J.; De Cristofaro, E.; Zannettou, S.; and Stringhini, G. 2021. Understanding the effect of deplatforming on social networks. In *Proceedings of the 13th ACM Web Science Conference 2021*, WebSci '21, 187–195. New York, NY, USA: Association for Computing Machinery.

Balcı, U.; Sirivianos, M.; and Blackburn, J. 2023. A data-driven understanding of left-wing extremists on social media. https://doi.org/10.48550/arXiv.2307.06981.

Belal, M.; Ullah, G.; and Khan, A. A. 2022. Islamophobic tweet detection using transfer learning. In *2022 International Conference on Connected Systems & Intelligence (CSI)*, 1–9.

Burnap, P., and Williams, M. L. 2015. Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy & Internet* 7(2):223–242.

Caselli, T.; Basile, V.; Mitrović, J.; and Granitzer, M. 2021. HateBERT: Retraining BERT for abusive language detection in English. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, 17–25. Online: Association for Computational Linguistics.

Cervi, L.; Tejedor, S.; and Gracia, M. 2021. What kind of islamophobia? representation of muslims and islam in italian and spanish media. *Religions* 12(6).

Chandra, M.; Reddy, M.; Sehgal, S.; Gupta, S.; Buduru, A. B.; and Kumaraguru, P. 2021. "a virus has no religion": Analyzing islamophobia on twitter during the covid-19 outbreak. In *Proceedings of the 32nd ACM Conference on Hypertext and Social Media*, HT '21, 67–77. New York, NY, USA: Association for Computing Machinery.

Efstratiou, A.; Blackburn, J.; Caulfield, T.; Stringhini, G.; Zannettou, S.; and Cristofaro, E. D. 2022. Non-polar opposites: Analyzing the relationship between echo chambers and hostile intergroup interactions on reddit.

Evolvi, G. 2018. Hate in a tweet: Exploring internet-based islamophobic discourses. *Religions* 9(10).

Horta Ribeiro, M.; Jhaver, S.; Zannettou, S.; Blackburn, J.; Stringhini, G.; De Cristofaro, E.; and West, R. 2021. Do platform migrations compromise content moderation? evidence from r/the_donald and r/incels. *Proc. ACM Hum.-Comput. Interact.* 5(CSCW2).

Khan, H., and Phillips, J. L. 2021. Language agnostic model: Detecting islamophobic content on social media. In *Proceedings of the 2021 ACM Southeast Conference*, ACM SE '21, 229–233. New York, NY, USA: Association for Computing Machinery.

Mehmmod, Q.; Kaleem, A.; and Siddiqi, I. 2022. Islamophobic hate speech detection from electronic media using deep learning. *Mediterranean conference on pattern recognition and artificial intelligence* V:187–200.

Nadal, K. L. Y.; Griffin, K. E.; Hamit, S.; Leon, J.; Tobio, M.; and Rivera, D. P. 2012. Subtle and overt forms of islamophobia: Microaggressions toward muslim americans. *Journal of Muslim Mental Health* 6.

Patel, P. 2022. The appg, islamophobia and anti-muslim racism. *Feminist Dissent* (6):205–229.

Sealy, T. 2021. Islamophobia: With or without islam? *Religions* 12(6).

Smith, C. C. 2014. "ex-muslims," bible prophecy, and islamophobia: Rhetoric and reality in the narratives of walid shoebat, kamal saleem, ergun and emir caner. *Islamophobia Studies Journal* 2(2):76–93.

Soral, W.; Liu, J. H.; and Bilewicz, M. 2020. Media of contempt: Social media consumption predicts normative acceptance of anti-muslim hate speech and islamoprejudice. *International Journal of Conflict and Violence* 14:1–13.

Squire, M. 2019. Network, text, and image analysis of anti-muslim groups on facebook.

Tahmasbi, F.; Schild, L.; Ling, C.; Blackburn, J.; Stringhini, G.; Zhang, Y.; and Zannettou, S. 2021. "go eat a bat, chang!": On the emergence of sinophobic behavior on web communities in the face of covid-19. In *Proceedings of the Web Conference 2021*, WWW '21, 1122–1133. New York, NY, USA: Association for Computing Machinery.

Vidgen, B., and Yasseri, T. 2020. Detecting weak and strong islamophobic hate speech on social media. *Journal of Information Technology & Politics* 17(1):66–78.