Toward Inclusivity: Rethinking Islamophobic Content Classification in the Digital Age

Esraa Aldreabi[1], Mukul Dev Chhangani[1], Khawlah M. Harahsheh[2], Justin M. Lee[3], 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
[3]Independent Scholar

Abstract

In this paper, we implement a comprehensive three-class system to categorize social media discussions about Islam and Muslims, enhancing the typical binary approach. These classes are: I) General Discourse About Islam and Muslims, II) Criticism of Islamic Teachings and Figures, and III) Comments Against Muslims. These categories are designed to balance the nuances of free speech while protecting diverse groups like Muslims, ex-Muslims, LGBTQ+ communities, and atheists. By utilizing machine learning and employing transformer-based models, we analyze the distribution and characteristics of these classes in social media content. Our findings reveal distinct patterns of user engagement with topics related to Islam, providing valuable insights into the complexities of digital discourse. This research contributes to the fields of quantitative social science by offering an improved method for understanding and moderating online discussions on sensitive religious and cultural subjects.

Introduction

In an era where social media platforms are pivotal in shaping public opinion, it is crucial to address the dynamic and sensitive discourse surrounding Islam and Muslims. We introduce a three-class system for analyzing Twitter and Reddit data: General Discourse About Islam and Muslims (Class I), Criticism of Islamic Teachings and Figures (Class II), and Comments Against Muslims (Class III). This approach aims to more accurately capture the complexities of online interactions, offering a deeper understanding of diverse perspectives. Balancing freedom of speech with the protection of groups from Islamophobic rhetoric is a key focus of our classification schema. We differentiate between academic or general discourse and direct attacks against Muslims, thus protecting Muslim communities while ensuring that the voices of ex-Muslims, LGBTQ+ individuals, atheists, and others are heard and appropriately contextualized, steering clear of being misconstrued as hate speech (Aldreabi and Blackburn 2024; Patel 2022; Imhoff and Recker 2012). Our initiative is underscored by a critical understanding of the dynamic online environment, notably the counterproductive effects of suspending users for toxic behavior. Such actions often lead to these users migrating to platforms with more lenient moderation, potentially amplifying the toxicity and increasing the risk of radicalization (Ali et al. 2021; Horta Ribeiro et al. 2021).

Our approach respects diverse viewpoints and aims to reduce harmful content. Class II facilitates critical engagement with Islamic teachings, crucial for groups like ex-Muslims or atheists. Class III, meanwhile, identifies content that could perpetuate stereotypes or incite hostility against Muslims, thereby serving as a defense against Islamophobia. Utilizing both traditional machine learning and advanced transformer-based models, we analyze content across Twitter and Reddit. This analysis demonstrates the efficacy of these models in text classification and illuminates user engagement patterns with Islam-related topics on these platforms.

The adoption of a three-class system offers numerous advantages. It increases accuracy beyond simple binary classification, encapsulates the complex dimensions of conversations regarding Islam and Muslims, and separates valid critiques from Islamophobic rhetoric. This approach helps preserve free speech and safeguarding vulnerable online groups. Our goal with this content classification tool is to enhance moderation practices and encourage respectful, well-informed discussions about Islam and Muslims, taking into account the varied experiences and perspectives of everyone involved.

Related Work

In the realm of social media, various scholars have offered important insights into Islamophobia and extremism, (Vidgen and Yasseri 2020) introduced a classification system for Islamophobia that moves beyond simple binary categories, categorizing content as non-Islamophobic, weakly Islamophobic, or strongly Islamophobic. This approach offers a more nuanced view of the subject. Regarding political groups, (Squire 2019) focused on Islamophobic sentiment within far-right groups on Facebook, while (Balcı, Sirivianos, and Blackburn 2023) examined left-wing extremism. Additionally, (Efstratiou et al. 2022) conducted a historical analysis of Reddit's political spaces, revealing diverse posting patterns and complex relationships between echo

chamber engagement, polarization, and hostility across different political leanings. (Puschmann et al. 2016) focused on a specific political movement in Germany, exploring Twitter data to understand its supporters' and opponents' media consumption. (Efstratiou et al. 2022) investigated the news ecosystem across various platforms, finding that polarized communities like Gab and the r/The_Donald subreddit often reference untrustworthy sources. Fringe communities were also found to disproportionately influence narratives on topics such as political elections, immigration, and foreign policy. Moreover, (Soral, Liu, and Bilewicz 2020) examined how users' choices of news sources on social media are linked to their exposure to anti-Muslim hate speech. Their findings point to a clear relationship, especially for regular social media users. (Ahmanideen and Iner 2024) demonstrated the tangible effects of online hate groups in real-world scenarios, particularly in anti-mosque campaigns post the war on terror. (Mehmmod, Kaleem, and Siddiqi 2022) utilized deep learning for detecting Islamophobic content, showing the importance of advanced technology in addressing online hate speech. (Ahmed Khan, Shah, and Ahmad 2020) combined various methods to study the Twitter hashtag #stopIslam. The increase in Islamophobia and hate speech during critical times, such as the COVID-19 pandemic, was documented by (Ghasiya and Sasahara 2022; Chandra et al. 2021; Tahmasbi et al. 2021), highlighting the link between global crises and the escalation of online hate speech. (Belal, Ullah, and Khan 2022) proposed a transfer learning approach using ULMFiT for detecting Islamophobia on Twitter. (Albadi, Kurdi, and Mishra 2018) focused on identifying religious hate speech in the Arabic-speaking online community, using a combination of methods to effectively distinguish complex expressions of hate. (Khan and Phillips 2021) suggested translating content into English for better detection of Islamophobia in different languages, acknowledging the varied linguistic landscape of online hate speech. On Reddit, (Aldreabi, Lee, and Blackburn 2023) used advanced analysis techniques to delve into Islamophobic comments, uncovering themes related to religion and politics. (Ali and Zannettou 2022) used a lexicon based approach to assess posts on 4chan's /pol/ board, finding a high percentage of posts with Antisemitic and Islamophobic terms to be hateful. While, (González-Pizarro and Zannettou 2023) utilized the CLIP model to analyze similar content on 4chan, emphasizing CLIP's effectiveness in detecting hateful content. Lastly, (Aldreabi and Blackburn 2024) discussed the fine line between Islamophobia and legitimate criticism, emphasizing the need for careful consideration in using detection tools to ensure they do not inadvertently impact various groups like atheists, feminists, ex-Muslims, and others. This highlights the importance of balanced approaches in tackling hate speech.

Dataset and Labeling

In the digital age, balancing protecting freedom of speech with ensuring the safety of Muslim communities is crucial. Our three-class system supports this balance, categorizing social media content into distinct groups with specific focus and criteria:

- **Class I: General Discourse About Islam and Muslims:** This category includes content that engages in a positive or neutral manner with Islam or Muslims. It covers a broad spectrum of discourse, from cultural exchanges to academic discussions, characterized by informative, respectful, and non-hostile dialogue. This class represents everyday conversations, educational content, and cultural exchanges that contribute to a broader narrative about Islam and Muslim communities. Examples:

  - "Wishing all my Muslim colleagues a peaceful and blessed Ramadan."

  - "It's heartwarming to see local mosques opening their doors to non-Muslims for cultural exchange events."

- **Class II: Criticism of Islamic Teachings and Figures:** This category is designated for content that offers critical perspectives on Islamic teachings or figures within Islam. It includes comments that may use strong or offensive language, provided they focus on religious doctrines or historical figures, not targeting the Muslim community as a whole. This class acknowledges the space for critical discourse in religious discussions, distinguishing it from derogatory comments aimed at Muslims. It includes critiques from various perspectives, including emotionally charged tones from ex-Muslims, LGBTQ+ individuals, and others who engage in debates or question aspects of Islamic teachings (Belal, Ullah, and Khan 2022; Aldreabi and Blackburn 2024; Albadi, Kurdi, and Mishra 2018). Examples:

  - "Most of the people who leave Islam don't show themselves because they would be murdered for it as Islam said kill exmuslims."

  - "This combined with the disturbing homophobic verses in the Quran really does make me question my faith."

- **Class III: Comments Against Muslims:** This category identifies overtly Islamophobic comments, characterized by harmful stereotypes, derogatory language, or incitement of violence against Muslims. This class includes content that directly attacks individuals based on their religious identity, propagates false narratives, or calls for discriminatory actions against Muslims (Patel 2022; Evolvi 2018; Cervi, Tejedor, and Gracia 2021). By isolating this type of content, the classification system underscores the importance of distinguishing Islamophobia from general religious critique, highlighting the need for strategies to combat harmful rhetoric in online spaces. Examples:

  - "Muslim children should be slaughtered."

  - "Every Muslim is a potential terrorist."

In our study, we apply this three-class system to re-label publicly available datasets from Twitter (Khan and Phillips 2021) and Reddit (Aldreabi, Lee, and Blackburn 2023), originally categorized under a binary system as either Islamophobic or non-Islamophobic. By adopting this approach, we

Platform	Class	Count
Twitter	Class I: General Discourse About Islam and Muslims	2,401
Twitter	Class II: Criticism of Islamic Teachings and Figures	197
Twitter	Class III: Comments Against Muslims	1,879
Twitter Total		4,477
Reddit	Class I: General Discourse About Islam and Muslims	979
Reddit	Class II: Criticism of Islamic Teachings and Figures	659
Reddit	Class III: Comments Against Muslims	362
Reddit Total		2,000

Table 1: Class counts for Twitter and Reddit.

aim to capture a more comprehensive understanding of online discourse about Islam and Muslims. We include 2,000 comments from Reddit and 4,477 from Twitter in our study. To validate the precision of our relabeling, we compute the Cohen's Kappa scores, which assess the agreement among annotators. These scores, measuring 0.88 for Reddit and 0.91 for Twitter, imply a near-perfect level of agreement.

Our classification aims to create a balanced digital environment. This is illustrated in Table 1, which displays the distribution of discussions across Twitter and Reddit. The table organizes these conversations into three distinct classes, offering a clear insight into the prevalence of each class on the respective platforms.

Evaluation of Machine Learning and Transformer Models on The Dataset

In this section, we evaluate various models following the relabelling of Twitter and Reddit datasets into three-classes. We utilize a range of models, from traditional machine learning to advanced transformer-based ones, to categorize and analyze our dataset, taking into account the complexities introduced by the updated classification system. The evaluation is directly linked to the revised labeling strategy detailed in the previous section. Transitioning from a binary to a three-class system presents a challenge for the models: they must now not only differentiate between Islamophobic and non-Islamophobic content but also categorize content as general discourse, criticism, or direct comments against Muslims.

Preprocessing

For both types of models, the preprocessing of data is a critical step. It includes cleaning the text by removing URLs, emojis, and special characters, normalizing the text through lowercasing, and eliminating extra spaces. Additionally, we apply tokenization and stopword removal. After cleaning, the dataset is divided into three parts: 70% for training, 15% for validation, and 15% for testing, ensuring a thorough evaluation of the models.

Machine Learning Models Performance

Having prepared our dataset, we assess several traditional machine learning models: Multinomial Naive Bayes, Support Vector Machine, Logistic Regression, Random Forest, Gradient Boosting, Decision Tree, and K-Nearest Neighbors. These models are integrated into a TfidfVectorizer pipeline. We focus our evaluation on their performance in terms of precision, recall, and F1-score. The results of this evaluation are summarized in Table 2.

Model	Precision	Recall	F1-Score
Multinomial Naive Bayes	0.87	0.74	0.77
Support Vector Machine	0.88	0.84	**0.86**
Logistic Regression	0.88	0.83	0.85
Random Forest	0.87	0.79	0.82
Gradient Boosting	0.88	0.83	0.85
Decision Tree	0.84	0.82	0.83
K-Nearest Neighbors	0.79	0.76	0.77

Table 2: Model performance comparison.

Model	Precision	Recall	F1-Score
RoBERTa	0.87	0.87	0.87
BERT	0.86	0.86	0.86
BART	0.87	0.86	0.86
HateBERT	0.87	0.89	**0.88**
BERTweet	0.85	0.84	0.84
ERNIE	0.86	0.86	0.86
DeBERTa	0.87	0.89	**0.88**

Table 3: Model performance comparison.

Transformer Models Performance

Transformer models are trained over 10 epochs with a batch size of 32 using the AdamW optimizer and a linear learning rate schedule with warmup. This training regimen is aimed at maximizing the models' effectiveness in handling the complexities of the classification system.

Our analysis indicates that DeBERTa and HateBERT stand out in terms of performance. DeBERTa achieves an F1-score of 0.88, with both precision and recall at 0.87. HateBERT follows closely, also achieving an F1-score of 0.88, but with precision at 0.87 and recall at 0.89. BART and BERTweet also show strong results: BART has an F1-score, precision of 0.86 and a recall of 0.87, while BERTweet maintains consistent scores of 0.84 in recall and F1-score, and 0.85 in precision. RoBERTa, BERT, and ERNIE, exhibit solid outcomes with precision, recall, and F1-scores of 0.87, 0.86, and 0.86, respectively. The similarity in performance metrics highlights the models' strengths and the nuanced differences that may influence their application in specific contexts. The comprehensive results for each model are detailed in Table 3.

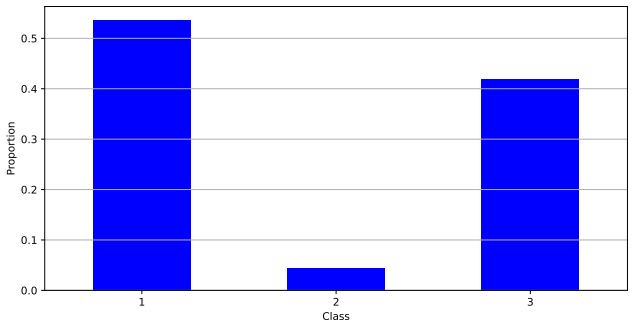

Figure 1: Class distribution in Twitter.

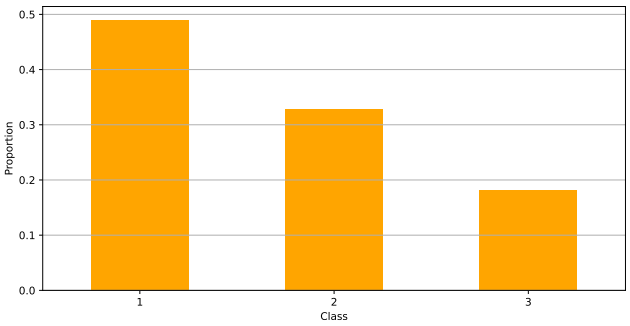

Figure 2: Class distribution in Reddit.

Class Distribution and Platform-Specific Trends

Analyzing the discourse on Twitter and Reddit reveals significant platform-specific trends in discussions about Islam and Muslims. On Twitter dataset, we observe a distribution of 53.63% for Class I (General Discourse About Islam and Muslims), 41.97% for Class III (Comments Against Muslims), and 4.40% for Class II (Criticism of Islamic Teachings and Figures), as seen in Figure 1. Similarly, on Reddit, we find 48.97% for Class I, 32.92% for Class II, and 18.11% for Class III, as shown in Figure 2. To assess whether the differences in class distribution between Twitter and Reddit are statistically significant, we perform a Chi-Square Test of Independence, resulting in a Chi-Square Statistic of 1,084.28 with 2 Degrees of Freedom, and a P-value of less than 0.001. This disparity suggests that Twitter users engage more in narratives classified as Islamophobic, evidenced by the higher prevalence of Class III content. Conversely, Reddit users demonstrate a greater tendency for engaging in critical discussions about Islamic teachings and figures, shown by the larger proportion of Class II content. These findings underscore distinct user engagement patterns on Twitter and Reddit, necessitating nuanced, platform-specific strategies for understanding and engaging with online discourse about Islam and Muslims. The stark contrast in class distribution between the two platforms highlights the complexity of digital communication and the importance of context-aware analysis in social media discussions about Islam and Muslims.

Complexity of Nuanced Language: Complementing our class distribution analysis, we conduct sentiment analysis to gain deeper insights into the emotional tones within these classes. Utilizing the TweetEval model (Barbieri et al. 2020), our analysis not only offers a nuanced understanding of the emotions present in each class but also highlights the varied nature of discussions on Twitter and Reddit.

On Twitter, Class I (General Discourse) primarily exhibits positive sentiments (56.02%), with neutral sentiments at 33.99% and negative sentiments at 9.99%. However, Classes II and III lean heavily towards negative sentiments, at 73.60% and 73.18% respectively, hinting at a more confrontational tone. On Reddit, Class I is mostly neutral (63.13%), yet there is a considerable presence of negative sentiments (28.60%). This could be attributed to Muslims defending their beliefs with emotion-laden language. In Classes II and III, there is a marked dominance of negative sentiments, with 86.32% in Class II and 95.30% in Class III.

The high negative sentiments in Classes II and III on Twitter may point to more aggressive, possibly Islamophobic narratives. Conversely, on Reddit, the negative sentiments in Class I, possibly resulting from Muslims' emotional defense of their faith, underscore the platform's capacity for deeper, though at times contentious, discussions.

Conclusion and Future Work

Moving from a binary to a three-class classification system provides a more comprehensive understanding and categorization of discussions about Islam and Muslims on social media. This approach captures a wider variety of dialogues, from positive and critical discussions to Islamophobic comments, improving the accuracy of content classification. Our research applies machine learning and advanced transformer models, highlighting the capabilities and limitations of these technologies in processing complex online discussions. For future work, expanding the diversity of data sources to include more social media platforms could offer a broader perspective on online discourse, accommodating varied demographics and communication styles across platforms. It is crucial to address the biases inherent in machine learning and transformer models to enhance classification fairness and accuracy. Incorporating multilingual content and adapting to different cultural contexts will enrich future analysis, capturing nuances missed by focusing solely on English-language content. These steps not only aim to refine the accuracy and fairness of content classification but also enrich our grasp of the complex dynamics that characterize online discussions. Ultimately, our endeavor seeks to pave the way for future research that is both more inclusive and reflective of the global, multifaceted nature of social media conversations.

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.

Ahmed Khan, R.; Shah, M.; and Ahmad, N. 2020. Securitization of islam and muslims through social media: A content analysis of stopislam in twitter. *Global Mass Communication Review* V.

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, M., and Zannettou, S. 2022. Analyzing antisemitism and islamophobia using a lexicon-based approach. In *ICWSM Workshops*.

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.

Barbieri, F.; Camacho-Collados, J.; Espinosa Anke, L.; and Neves, L. 2020. TweetEval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, 1644–1650. Online: Association for Computational Linguistics.

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.

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).

Ghasiya, P., and Sasahara, K. 2022. Rapid sharing of islamophobic hate on facebook: The case of the tablighi jamaat controversy. *Social Media + Society* 8(4):20563051221129151.

González-Pizarro, F., and Zannettou, S. 2023. Understanding and detecting hateful content using contrastive learning. *Proceedings of the International AAAI Conference on Web and Social Media* 17(1):257–268.

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).

Imhoff, R., and Recker, J. 2012. Differentiating islamophobia: Introducing a new scale to measure islamoprejudice and secular islam critique. *Political Psychology* 33:811–824.

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.

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

Puschmann, C.; Ausserhofer, J.; Maan, N.; and Hametner, M. 2016. Information laundering and counter-publics: The news sources of islamophobic groups on twitter. *Proceedings of the International AAAI Conference on Web and Social Media*.

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.