Enhancing Automated Hate Speech Detection: Addressing Islamophobia and Freedom of Speech in Online Discussions

Esraa Aldreabi
Department of Computer Science
Binghamton University
Binghamton, NY
ealdrea1@binghamton.edu

Jeremy Blackburn
Department of Computer Science
Binghamton University
Binghamton, NY
jblackbu@binghamton.edu

***Abstract*—This paper emphasizes the necessity of a precise definition of Islamophobia within the realm of social media platforms. The current broad understanding often leads to misclassification and poses challenges to the principles of freedom of speech. Differentiating between Islamophobia and legitimate criticism presents a complex task for automated hate speech detection models, particularly in the presence of offensive language and emotionally charged tones. Furthermore, the paper highlights the inadvertent discriminatory consequences that can arise from misusing Islamophobia detection models against atheists, feminists, ex-Muslims, and others, underscoring the importance of safeguarding their rights. Our study introduces a refined definition and employs advanced deep learning models. It demonstrates a reduction in the number of Islamophobic comments in the dataset while maintaining the accurate identification of genuine instances of Islamophobia. This distinction is made without compromising discussions related to religion and criticism. The results show promise in improving the precision of Islamophobia identification, all while upholding principles of free expression and open dialogue.**

***Index Terms*—Islamophobia, Freedom Of Speech, Topics, Deep Learning, Social Media**

I. Introduction

Islamophobia, which encompasses fear, prejudice, or discrimination against Islam and Muslims, is escalating as a concern in our interconnected world. Given the widespread influence of social media platforms as hubs for public discourse, it is imperative to closely examine the manifestation of hate speech within this digital landscape [1]. Joining online communities characterized by hate speech results in an expansion of hate speech beyond those communities, which persists for an extended period. This underscores the adverse effects of echo chambers and underscores the significance of moderation in addressing the adoption of hateful speech [2].

This paper aims to address the pressing need for a precise and nuanced definition of Islamophobia, particularly in the context of social media platforms. Within the digital realm, diverse communities including Ex-Muslims, atheists, feminists, and members of the LGBTQ+ community express their dissent toward certain religious aspects. They engage in critical examination and discussion surrounding Islamic texts, leaders, and historical events. However, the existing definition of Islamophobia lacks specificity [3], which can lead to potential mislabeling of such content. This misclassification poses a significant challenge to the principles of freedom of speech.

Differentiating Islamophobia from freedom of speech poses challenges for hate speech detection models, particularly when offensive language is involved [4] [5]. The presence of emotionally charged tones, often observed in content from ex-Muslims or individuals in the LGBTQ+ community, further complicates the process. Comprehensive hate speech detection models should understand context and sentiment to accurately distinguish valid criticism from expressions of hatred or discrimination. Furthermore, it is important to acknowledge that Islamophobia can be misused as a tool to discriminate against atheists, feminists, or ex-Muslims and stifle criticism of Islam [6] [7]. Safeguarding the rights and freedoms of marginalized groups becomes paramount in addressing this issue.To explore these implications, we use Reddit as a platform to observe how different communities, including atheists, ex-Muslims, and the LGBTQ+ community, are affected by varying definitions of Islamophobia.

This study focuses on defining Islamophobia accurately and its implications for labeling comments on Reddit using deep learning models. In comparison to prior research on Islamophobia detection, this paper brings forth several noteworthy contributions. Our study emphasizes the need for advanced algorithms that consider context, sentiment, and individual experiences in hate speech detection. We address the complexities introduced by offensive language and emo-

ASONAM '23, November 6–9, 2023, Kusadasi, Turkiye

ACM ISBN 979-8-4007-0409-3/23/11.
http://dx.doi.org/10.1145/3625007.3627487

tionally charged tones in online discussions, contributing to the development of more accurate automated systems. We draw attention to the potential misuse of Islamophobia as a tool to discriminate against marginalized groups, such as atheists, feminists, and ex-Muslims, emphasizing the significance of safeguarding their rights and freedoms. By employing deep learning models and refining the definition, we aim to enhance the accuracy of identifying genuine instances of Islamophobia while preserving the principles of free expression.

II. RELATED WORK

In previous studies focused on detecting Islamophobia on social media platforms using machine learning and deep learning techniques, researchers have provided a specific definition of Islamophobia for social media. For instance, in [8], Islamophobia is defined as "any content that is produced or shared and expresses indiscriminate negativity against Islam or Muslims." Moreover, to address the complexities of Islamophobic content, the authors introduced a categorization system that goes beyond simple binary classification. This system includes three classifications: 1) non-Islamophobic, 2) weak Islamophobic, and 3) strong Islamophobic. The authors in [5] presented a study on detecting and analyzing religious hate speech in the Arabic Twittersphere. The authors develop various classification models, including lexicon-based and deep-learning-based approaches, to address the challenge of distinguishing hate speech from other forms of profane language. Their findings show that a simple Recurrent Neural Network (RNN) architecture with Gated Recurrent Units (GRU) achieves a satisfactory performance in detecting religious hate speech, with an Area Under the Receiver Operating Characteristic curve (AUROC) of 0.84. Additionally, their research reveals that a significant portion of discussions about religion in the Arabic Twittersphere involves hate towards various religious groups, particularly Jews, atheists, and Shia Muslims. The COVID-19 pandemic witnessed a surge in Islamophobia, as indicated by the research conducted by [9] and [10]. In [10], the authors focused on analyzing Islamophobia on the social media platform Twitter during the COVID-19 outbreak. The study examines the prevalence and patterns of Islamophobic content in relation to the pandemic, exploring how negative sentiments and discrimination towards Islam and Muslims were expressed on Twitter during this period. The authors followed the guidelines provided in [8] to define Islamophobia.

An investigation of the rapid dissemination of Islamophobic hate on the social media platform Facebook was performed by the authors in [9], specifically focusing on the Tablighi Jamaat controversy. The study examines how anti-Muslim hate groups and individuals exploited the controversy to spread misinformation and incite violence against Muslims in India. In this study, Islamophobia is defined as "hatred or fear of Muslims or their politics or culture". The authors in [11] utilized deep learning-based approach to detect Islamophobic hate speech in electronic media. The authors define Islamophobic hate speech as the indiscriminate negative attitude and behavior towards Muslims and Islam.

Addressing the detection of Islamophobia on Twitter, [4] proposed a transfer learning approach using Universal Language Model Fine Tuning (ULMFIT). they collected data based on hashtags that target Muslims but to protect freedom of speech some of the tweets were not considered. In their study [12], the authors tackled the challenge of detecting Islamophobia across multiple languages by focusing on the development of a model for detecting Islamophobic content across different languages on social media platforms. Their aim was to effectively address the challenge of language variations and cultural contexts in detecting Islamophobia. To achieve this, they proposed a language-agnostic approach that leveraged machine learning techniques to identify and classify Islamophobic content. The authors in [13] presents a study that utilizes deep learning techniques to identify Islamophobic content on the social media platform Reddit. The authors employ topic modeling to analyze the identified Islamophobic comments and uncover various topics such as the Islamic dress code, religious practices, marriage, and politics. The detection of Islamophobia is achieved through the use of deep learning models. The paper defines Islamophobic comments as those expressing negative sentiments towards Muslims, including derogatory remarks, advocating for restrictions on their entry or religious practices, and engaging in character assassination of religious figures.

III. DATASET COLLECTION AND LABELING

We retrieve a dataset of 200,000 comments by searching for keywords "Islam" and "Muslim" utilizing the Pushshift API dataset [14]. The choice of Reddit as a data source was deliberate due to its diverse range of subreddits and user communities. Subreddits such as "Atheism", "exmuslims", "feminine", and LGBTQ+ groups allow us to examine how different online communities are affected by variations in the definitions of Islamophobia. To ensure the homogeneity and quality of our dataset, we implement a language detection algorithm, which allow us to retain exclusively those comments composed in the English language, thereby maintaining data consistency.

We employ a semi-automatic annotation method for labeling the dataset. Initially, we filter and select 2,000 comments from the original dataset using a list of keywords derived from a previous study [13]. This keyword list includes positive, negative, and neutral terms associated with Islam. To label the dataset accurately, we employ a hybrid method that combines automated keyword-based filtering with human validation. This integration of approaches ensures both balance and comprehensiveness in our labeling process.Two distinct definitions of Islamophobia, namely Definition-1 and Definition-2, guide our labeling efforts. Under Definition-1, Islamophobia was defined as in [8]:

> Definition-1: "Any content that is produced or shared and expresses indiscriminate negativity against Islam or Muslims."

Comments that fall under this definition are categorized as Islamophobic if they exhibit negative sentiments towards Mus-

lims, including the use of derogatory language, associating Muslims with terrorism or extremism, advocating for exclusionary measures against them, criticizing the history or teachings of Islam, or defaming religious figures. Comments that are not Islamophobic focus on discussing Islamic practices, miracles, teachings, and historical aspects without displaying Islamophobic content. Out of the 2,000 comments collected, 1,022 comments are classified as Islamophobic, while 977 comments are categorized as not Islamophobic. The Cohen's kappa coefficient for the annotation was 0.955, indicating almost perfect agreement between the annotators.

In Definition-2, the exclusion of the discussion or criticism of Islamic teachings as a criterion for labeling comments as Islamophobic is based on the recognition of the fundamental principle of freedom to express criticism towards religious ideas, which is deeply ingrained in Western democratic societies. This principle is supported by the First Amendment of the United States Constitution, which guarantees the right to freely criticize religions (U.S. Const. amend. I) [15]. It is important to note that the freedom of expression protects the act of criticizing specific beliefs, including religions, ideologies, or prophets, as long as it is conducted without threats or intimidation [16]. We define Islamophobia as:

Definition-2: "Any content that is produced or shared and expresses indiscriminate negativity against Muslims."

We categorize comments that employ derogatory language towards Muslims, associate Muslim communities with terrorism or extremism, dehumanize or demean Muslims, or advocate for exclusionary measures targeting Muslims as instances of Islamophobia. Applying this definition, we identify 520 out of the 2,000 comments as Islamophobic. To ensure a balanced dataset for model training, we decrease the number of comments categorized as non-Islamophobic to 709, resulting in a total of 1,229 comments. The annotation process involves independent labeling by the annotators. The Cohen's kappa coefficient for the annotation is 0.92, signifying almost perfect agreement between the annotators.

After labeling the datasets, we employ them to train our models. We conduct experiments to evaluate various deep learning models for text classification. To prepare the data for further processing, we perform preprocessing steps. Firstly, we convert all comments to lowercase and remove punctuation marks, unknown Unicode characters, and delimiters. For our neural network implementation, we utilize the PyTorch-pretrained-Bert library, which includes the pre-trained BERT model, text tokenizer, and pre-trained WordPiece. We compare the performance of three models: 1) pre-trained BERT with a 2-layer MLP (Multi-Layer Perceptron) inserted, 2) pre-trained BERT with a Convolutional Neural Network (CNN) inserted, and 3) pre-trained BERT with Long Short-Term Memory Networks (LSTM) inserted.

During the experiment, we train the classifiers for 10 epochs using a batch size of 32. We employ the Adam optimizer with a learning rate of 10^{-6} and set the dropout probability to 0.2. To prepare the input, we utilize a BERT tokenizer to tokenize each comment. We employ WordPiece tokenization, based on the original BERT, to split words into sub-word units. To ensure an unbiased evaluation, we split the dataset into training, validation, and testing sets with proportions of 80%, 10%, and 10%, respectively. We apply stratified sampling to maintain the class distribution ratios of 0.8, 0.1, and 0.1, respectively, to avoid overfitting.

TABLE I
EVALUATION RESULTS FOR DEFINITION 1 AND DEFINITION 2.

Definition	Method	Precision	Recall	F1-Score
Definition1	$BERT_{base}$+MLP	0.735	0.71	0.70
	$BERT_{base}$+CNN	0.92	0.92	0.92
	$BERT_{base}$+LSTM	0.895	0.90	0.895
Definition2	$BERT_{base}$+MLP	0.765	0.655	0.64
	$BERT_{base}$+CNN	0.885	0.885	0.885
	$BERT_{base}$+LSTM	0.815	0.715	0.72

For evaluation purposes, we use the test datasets and consider three different metrics: F1-scores, precision, and recall. We summarize the results of these evaluations in Table I. Among the various models we evaluate for classification, we find that the $BERT_{base}$+CNN models consistently achieve higher F1-scores compared to the other models. Therefore, we utilize the two trained $BERT_{base}$+CNN models to classify the 200,000 comments for further analysis. Based on the classification results, Definition 1 identifies a total of 51,492 comments as Islamophobic, while Definition 2 classifies 49,383 comments as Islamophobic. This indicates that Definition 1 identified a slightly higher number of Islamophobic comments compared to Definition 2.

A. Ethics

We recognize the importance of ethical considerations when working with social media data. In our research, we strictly adhere to standard best practices [17] [18] to ensure the privacy and anonymity of individuals. We do not attempt to deanonymize any authors or disclose any personal information. It is important to note that our study solely utilizes publicly available data from social media platform.

IV. CONTENT ANALYSIS

In this section, we conduct a study to explore the effects of modifying the definition of Islamophobia on the classification process using deep learning models. Our objective is to analyze how these modifications affect the identification of Islamophobic comments in Reddit and the extraction of associated topics. The findings provide valuable insights into the changes observed in the frequencies of Islamophobic comments, the shifts in top subreddits associated with different definitions, and the prevalent topics within each defined context.

A. Subreddit Frequency Analysis

Figure 1 illustrates the top 20 subreddits associated with Definition 1, while Figure 2 displays the top 20 subreddits associated with Definition 2. Notably, the subreddit "exmuslim" consistently exhibits the highest frequency of Islamophobic comments across both definitions. However, Definition

2 demonstrates a substantial reduction in the prevalence of Islamophobic comments compared to Definition 1, as evidenced by the decreased frequency observed within "exmuslim" and the "atheism" subreddit. Moreover, there are variations in the top 20 subreddits when comparing the two definitions. Definition 1 highlights the prominence of subreddits like "Askgaybros" and "DebateReligion", where it is expected to encounter content that critically discusses and examines Islamic texts, leaders, or histories. Conversely, these particular subreddits are comparatively less prominent within Definition 2. According to [6] [7], Islamophobia can be utilized as a means to discriminate against various groups, including ex-Muslims, atheists, feminists, and homosexuals. With this understanding, our analysis focused on examining specific subreddits that are relevant to these groups, namely "exmuslim", "atheism", "askgaybros", "DebateReligion", as well as feminist subreddits and LGBTQ+ subreddits.

To assess the significance of these changes across selected subreddits, we conducted a comprehensive statistical analysis using Fisher's exact test. The frequencies for the selected subreddits, are presented in Table II. Under Definition 1, the subreddit "exmuslim" records a decrease in the number of Islamophobic comments from 4,284 to 2,997 under Definition 2, resulting in a percentage change of -30.04%. This change is highly significant, as indicated by a p-value of less than 0.001. Similarly, the subreddit "atheism" exhibits a decrease from 623 comments to 390 comments, reflecting a percentage change of -37.40%, with a corresponding p-value of less than 0.001. Furthermore, the subreddit "askgaybros" shows a decrease from 431 comments to 277 comments, signifying a percentage change of -35.73%, and a p-value of less than 0.001. Notably, the subreddit "DebateReligion" displays a substantial decrease, with the frequency of Islamophobic comments plummeting from 373 under Definition 1 to a mere 89 comments under Definition 2. The obtained p-value for this analysis is less than 0.001 indicating a remarkable change with a significant percentage decrease of -76.14%.

Regarding the feminist subreddits, we combine all relevant subreddits associated with feminism into one aggregated value. We make this decision due to the larger number of subreddits encompassing feminist perspectives. The list of subreddits is provided in [19]. The analysis shows a decrease in the frequency of Islamophobic comments from 114 comments under Definition 1 to 50 comments under Definition 2. This represents a remarkable percentage change of -56.14%, indicating a significant decrease in Islamophobic comments within the combined feminist subreddits. The analysis yields a p-value of less than 0.001, indicating statistical significance. Similarly, we conduct an analysis on the combined LGBTQ+ subreddits, which demonstrates a decrease in the frequency of comments related to Islamophobia. Under Definition 1, there are 526 comments, while under Definition 2, the number reduces to 356 comments. This change represents a percentage change of -32.32%. The p-value is less than 0.001, suggesting a notable decrease in Islamophobic discussions within the combined LGBTQ+ subreddits. Table III presents the results for the combination of the feminist and LGBTQ+ subreddits. Furthermore, Figure 3 provides a comprehensive visual representation of the variations in frequencies of Islamophobic comments across the top 50 subreddits. The plot enables a quick comparison between Definition 1 and Definition 2, visually highlighting the changes in the prevalence of Islamophobic comments.

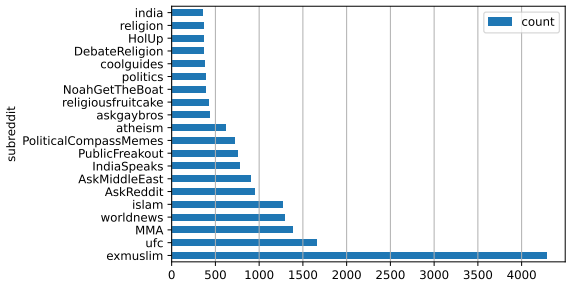

Fig. 1. Top 20 subreddits for Definition-1

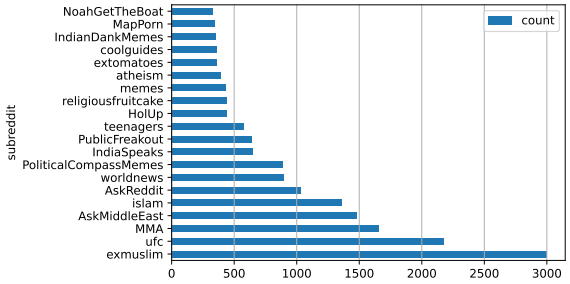

Fig. 2. Top 20 subreddits for Definition-2

TABLE II
FREQUENCIES AND PERCENTAGE CHANGES OF ISLAMOPHOBIC COMMENTS IN SELECTED SUBREDDITS UNDER DEFINITION 1 AND DEFINITION 2

Subreddit	Def.1 Freq	Def.2 Freq	change%	P-value
Exmuslim	4284	2997	-30.04	p<0.001
Atheism	623	390	-37.40	p<0.001
Askgaybros	431	277	-35.73	p<0.001
DebateReligion	373	89	-76.14	p<0.001

TABLE III
FREQUENCIES AND PERCENTAGE CHANGES FOR COMBINED RESULTS FOR FEMINIST AND LGBTQ+ SUBREDDITS

Group	Def.1 Freq	Def.2 Freq	change%	P-value
LGBTQ+	526	356	-32.32%	p<0.001
Feminist	114	50	-56.14%	p<0.001

B. Topic Analysis

To further investigate the impact of varying definitions on classification methods, we proceeded to extract topics from comments classified as Islamophobic and compared the outcomes between the two definitions. For this purpose, we employed Non-Negative Matrix Factorization (NMF) for topic modeling. NMF's ability to factorize a document-term matrix into non-negative matrices representing topics and their

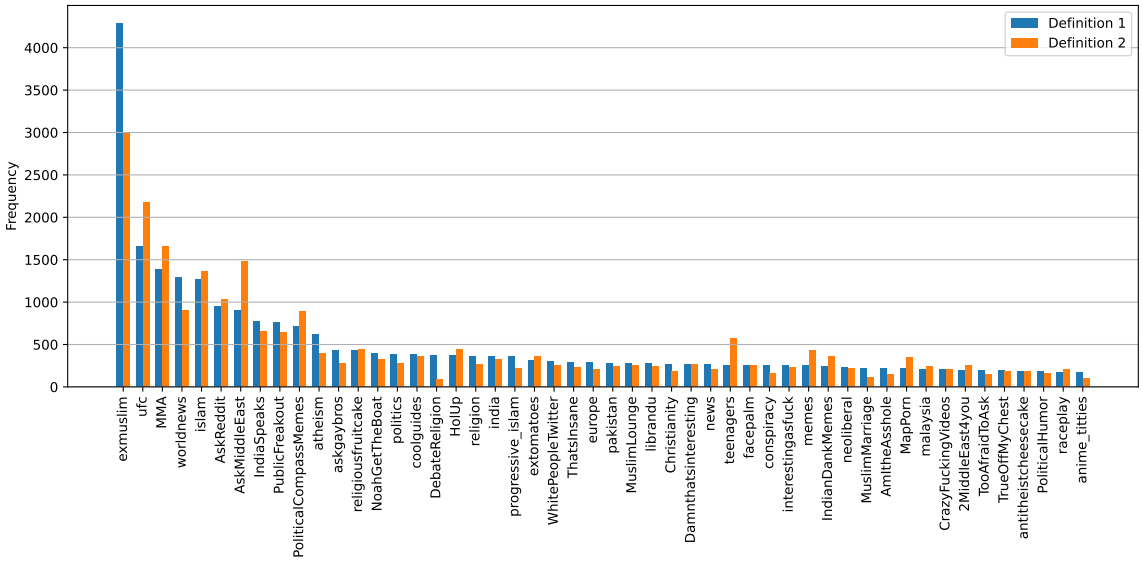

Fig. 3. Frequencies of Islamophobic Comments by Subreddit (Top 50)

TABLE IV
TOPICS EXTRACTED FROM THE COMMENTS IDENTIFIED AS ISLAMOPHOBIC USING DEFINITION 1.

Seq.No	Tokens	No. of Comments
1	women men muslim wear sex woman marry hijab rape man husband cover slave rap	2161
2	countries muslim gay country majority death western islamic middle saudi world law	1669
3	say allah quran islam prophet muslim muhammad one sex call man hadith comment word	3235
4	christianity christian believe jesus religions judaism allah hell heaven question world gods	1511
5	people white gay black racist hate muslim group trans call many like	2472
6	religion islam christianity religious religions peace world follow violent every cult muslim	2309
7	islam christianity convert cult leave bad nation worst radical religions shit	3977
8	kill convert islam leave muslim people hindus jews get name children apostates innocent	1174
9	muslims hindus india hindu ex hate christians non indian muslim majority even islam	2542
10	fuck islam muslim shit get ass country give go dumb bullshit yeah	2101

TABLE V
TOPICS EXTRACTED FROM THE COMMENTS IDENTIFIED AS ISLAMOPHOBIC USING DEFINITION 2.

Seq.No	Tokens	No. of Comments
1	islam leave nation women radical race nothing hate fuck	10424
2	muslims guy shit make call get mean christian thing hate kill bad yeah man leave	8887
3	christianity islam judaism religions hinduism abrahamic bad buddhism similar	1018
4	fuck islam muslim shit give ass dumb religion religions christianity mom	1970
5	muslims hate islam kill terrorists muslim race majority	2067
7	tap slut muslim favourite sex rat nudes go ppv cock	656
8	women muslim islam men wear hijab cover treat rape oppress	1115
9	christian muslim jewish atheist jew majority buddhist religious god white gay	1152

associated weights on each document aligns well with our objective of understanding the prevalent topics in Islamophobic comments. This decomposition not only allows us to identify dominant topics but also to interpret and analyze them in the context of the distinct definitions of Islamophobia.

To prepare the comments for topic modeling, we perform essential preprocessing steps, including the removal of stop words, white spaces, and punctuation. Additionally, we convert the text to lowercase. To determine the optimal number of topics that best represent the data, we employ the Coherence

Score as a metric. We conduct iterations across the range of 5 to 50 topics, with a step size of 5. For NMF, we utilize vectors created through the Term Frequency-Inverse Document Frequency (TF-IDF) based vectorization method. Table IV displays the topics we extract from Definition 1, while Table V showcases the topics we extract from Definition 2.

The tokens we extract from Definition 1 encompass a wide range of topics related to women, men, Islam, hijab, sexuality, and negative expressions. These tokens reflect the presence of discussions and references to various aspects of women's treatment within Islam. We observe references to the wearing of hijab, discussions about issues of oppression faced by women, and even conversations about sexuality within the context of Islamic teachings. Furthermore, the tokens touch upon issues related to Muslim-majority countries, such as homosexuality, ex-Muslims, death penalty, and laws. These discussions indicate a reflection of the diverse perspectives and opinions surrounding these topics within the online discourse. Additionally, the tokens mention Allah, the Quran, and other elements of Islamic faith, hinting at discussions about religious texts, beliefs, and interpretations.

In Definition 2, the tokens revolve around themes of Islam, women, radicalism, race, and negative expressions. The presence of terms like "hate", "kill", and derogatory language signals the existence of negative sentiments and emotions directed towards Muslims. Furthermore, the tokens imply that Islam is associated with radicalism, with references to terms like "terrorists" and negative opinions regarding its followers.

It suggests the presence of discussions that express animosity and hostility towards the religion. Additionally, the tokens highlight the existence of controversial and provocative language surrounding Islam, such as derogatory slurs and sexually explicit remarks. These expressions not only target Muslims but also indicate the presence of hateful content and comparisons with other religions, such as Christianity, Judaism, Hinduism, and Buddhism. This suggests that derogatory remarks and negative sentiments are not exclusive to Islam but may extend to other faiths as well. These tokens highlight the existence of controversial and provocative language surrounding Islam and the perception of it as a problematic or negative religion. Additionally, the tokens contain references to race, specifically linking it to Islam and Muslims. This implies the existence of discussions that associate Islam with a particular race or ethnic group, which can contribute to stereotypes and prejudices. Analyzing these findings, it becomes evident that both Definition 1 and Definition 2 capture different aspects and perspectives related to Islam. Definition 1 primarily focuses on discussions concerning women's rights, cultural practices, and interpretations of Islamic teachings. On the other hand, Definition 2 sheds light on negative sentiments, controversial language, and the presence of discussions that involve other religions in relation to Islam.

These results emphasize the importance of understanding the underlying definitions when studying topics related to Islam. They highlight the need to recognize the diverse viewpoints and narratives that emerge based on how Islamophobia is defined and perceived. It is crucial to differentiate between discussions that pertain to Islamic texts, histories, or rituals and instances of anti-Muslim racism and discrimination. Prioritizing critical analysis and questioning of religious tenets as fundamental components of an open society should be accompanied by ensuring that such discussions do not infringe upon the rights and dignity of individuals or contribute to a climate of hatred and discrimination. By promoting nuanced and respectful discussions, we can foster an environment of understanding, tolerance, and inclusivity.

C. Topics in Discussions Not Associated with Islamophobia

This subsection focuses on the topics we extract from comments identified as Not-Islamophobic based on Definition 2. Table VI presents these topics, encompassing a diverse range of discussions related to various aspects of Islam and Muslim communities. It is noteworthy that the model accurately recognizes these discussions as not-Islamophobic, acknowledging their alignment with the principles of freedom of speech.

One prominent topic we identify is women's rights within Islam, where commenters engage in debates and conversations regarding the treatment of women, the choice to wear a hijab, gender equality, and the role of women in Muslim societies. These discussions reflect a legitimate exploration of cultural practices, interpretations of religious texts, and societal norms.

Islamic law, or Sharia, is another topic that emerges in the not-Islamophobic discussions. Commenters delve into debates about the implementation and interpretation of Islamic laws, the role of Sharia in governance, and the extent to which it should be enforced. These discussions address a critical aspect of Muslim societies and the diversity of opinions within them.

Additionally, the model recognizes discussions about conversion from Islam, including debates on the reasons individuals leave the religion, the consequences they may face, and the societal and familial implications of apostasy. These conversations touch upon sensitive topics, such as the persecution of individuals who choose to leave Islam.

Marriage practices, such as the allowed age for marriage in Islam, also emerge as a topic in the not-Islamophobic discussions. Commenters engage in conversations about the religious and cultural context surrounding marriage, the role of parental consent, and the implications of early or forced marriages. These discussions reflect the diversity of opinions and perspectives on this topic within the Muslim community.

Furthermore, the model accurately recognizes criticisms pertaining to the persecution of individuals based on their sexual orientation and the penalty of death they face. By acknowledging these discussions as legitimate expressions of concern and critique, the model upholds the principles of freedom of speech while still identifying and addressing instances of Islamophobia.

D. Comparative Toxicity Analysis

Due to the challenges associated with distinguishing Islamophobia from freedom of speech, hate speech detection models

TABLE VI
TOPICS EXTRACTED FROM THE COMMENTS IDENTIFIED AS NON-ISLAMOPHOBIC USING DEFINITION 2.

Seq.No	Tokens	# of Comments
1	women men muslim right cover woman husband treat man equal wive sex	1768
2	wear hijab women cover force clothe woman choice dress muslim ban girls hair	2877
3	islamic law rule follow sharia laws state allow scholars school saudi apply teach	2877
4	convert islam christianity muslim force bear conversion tax didnt become family empire turks	2345
5	marry marriage man woman sex age wife allow husband muslim girl child old family parent	5134
7	say kill islam gay bad wrong someone literally person death murder muslim anyone fuck	5169
8	islam death religion reason exmuslim punishment wing apostasy penalty alone join stay	4193

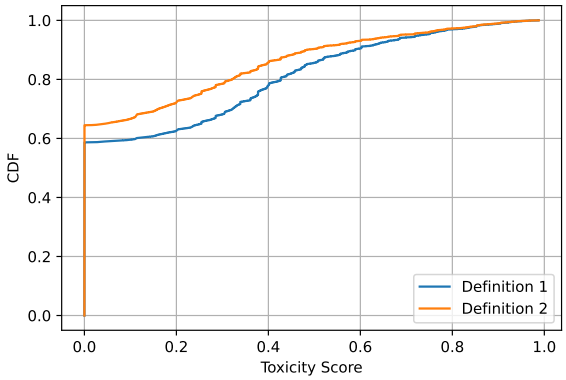

Fig. 4. Comparison of CDFs for Definition-1 and Definition-2 for Islamophobic Comment

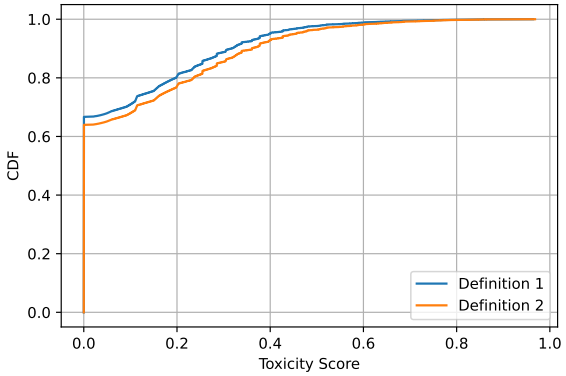

Fig. 5. Comparison of CDFs for Definition-1 and Definition-2 for Not-Islamophobic Comment

might face difficulties in accurately identifying offensive language and emotionally charged tones [4] [5]. It is important to recognize that models should possess a certain level of flexibility and consider multiple factors when dealing with sensitive topics like Islamophobia. To address this issue, our analysis involved leveraging Google's Perspective API [20], specifically the TOXICITY model. This API utilizes machine learning models trained on extensive comment datasets from diverse online sources to assign toxicity scores within a range of 0 to 1 [21].Across both Definition 1 and Definition 2, we observed that both Not-Islamophobic and Islamophobic comments had identical highest and lowest toxicity scores. The highest recorded toxicity score was 0.988, indicating a comment with a high level of toxicity. Conversely, the lowest toxicity score was 0.0, representing a comment with no detected toxicity, for both categories of comments under both definitions. While these highest and lowest toxicity scores provide insight into the range of toxicity levels present in the comments, it is essential to note that toxicity scores alone do not capture the complete context and nuances of Islamophobia. Our comparative analysis of the cumulative distribution function (CDF) curves, as depicted in Figure 4 for Islamophobic comments and Figure 5 for Not-Islamophobic comments, revealed notable differences between Definition 1 and Definition 2. For Islamophobic comments, the area under the CDF curve was 0.801 for Definition 1 and 0.848 for Definition 2, indicating that comments classified under Definition 2 generally exhibit higher levels of toxicity compared to Definition 1. The Kolmogorov-Smirnov test yielded a p-value of less than 0.001, further confirming the significant differences between the CDFs of Definition 1 and Definition 2 for Islamophobic comments. Conversely, when considering Not-Islamophobic comments, the average toxicity level was slightly higher under Definition 1: 0.879 compared to Definition 2: 0.867, suggesting that Definition 1 captures a slightly higher degree of toxicity in Not-Islamophobic comments. The Kolmogorov-Smirnov test also indicated a p-value of less than 0.001, indicating significant differences in the CDFs of Definition 1 and Definition 2 for non-Islamophobic comments.

The findings highlight that Definition 2 shows higher toxicity levels for Islamophobic comments compared to Definition 1, while Definition 1 captures slightly higher toxicity levels for Not-Islamophobic comments. This suggests that Definition 2 is more considerate of freedom of speech and avoids classifying comments as Islamophobic based solely on criticism of religion. It emphasizes the need to consider contextual factors in addressing Islamophobia and upholding freedom of speech online. Taking a holistic approach and considering multiple dimensions enables machine learning models to better assess and classify Islamophobic comments, providing a nuanced understanding of the issue while preserving freedom of speech.

V. LIMITATIONS AND FUTURE RESEARCH

In our study, we've pinpointed some key limitations that offer valuable insights for future research. Firstly, the inherent complexity of distinguishing between Islamophobia and valid criticism presents a significant challenge. Despite our

meticulous crafting of two distinct definitions of Islamophobia, the fine line between criticism and discrimination can lead to ambiguities in classification. This underscores the need for ongoing research efforts to refine classification methodologies and enhance accuracy in identifying Islamophobic content. Secondly, our study focused exclusively on the social media platform Reddit, selected for its diverse communities. However, it is crucial to acknowledge that Islamophobic content and its variations may manifest differently on other platforms. Therefore, future research should broaden its scope to investigate Islamophobia across various social media platforms, capturing the subtle differences that may exist. Furthermore, questions arise concerning the generalizability of our refined definition. It is essential to consider whether this definition is universally applicable across different platforms, regions, and cultural contexts. While our study provides valuable insights, its effectiveness and applicability may vary in diverse settings. Further research is required to explore the adaptability of the definition in different contexts to better understand its limitations. Expanding the study to encompass multiple social media platforms, regions, and cultural contexts represents a promising avenue for future research.

VI. Conclusion

In conclusion, this paper highlights the significance of defining and detecting Islamophobia on social media platforms. The lack of a clear and comprehensive understanding of Islamophobia poses challenges in accurately identifying and addressing this form of discrimination while upholding the principles of freedom of speech. It is crucial for automated hate speech detection models to consider not only offensive language but also the underlying context and sentiment to distinguish between valid criticism and expressions of discrimination or hatred. Our study has shed light on the complexities of defining and classifying Islamophobia, but we also acknowledge its limitations. Distinguishing between Islamophobia and valid criticism remains intricate, and the effectiveness of our refined definitions across various platforms, regions, and cultural contexts requires further investigation. Nonetheless, by proposing refined definitions of Islamophobia and leveraging advanced deep learning models, this research aims to enhance the accuracy of identifying genuine instances of Islamophobia while preserving the principles of free expression and open dialogue. This contribution is significant as it addresses a pressing issue in the digital age, where online spaces should foster inclusivity and respect for all individuals, irrespective of their religious beliefs or identities.

References

[1] P. Burnap and M. L. Williams, "Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making," *Policy & Internet*, vol. 7, no. 2, pp. 223–242, 2015. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/poi3.85

[2] M. Schmitz, G. Muric, and K. Burghardt, "Quantifying how hateful communities radicalize online users," in *2022 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE, nov 2022. [Online]. Available: https://doi.org/10.1109%2Fasonam55673.2022.10068644

[3] T. Sealy, "Islamophobia: With or without islam?" *Religions*, vol. 12, no. 6, 2021. [Online]. Available: https://www.mdpi.com/2077-1444/12/6/369

[4] M. Belal, G. Ullah, and A. A. Khan, "Islamophobic tweet detection using transfer learning," in *2022 International Conference on Connected Systems & Intelligence (CSI)*, 2022, pp. 1–9.

[5] N. Albadi, M. Kurdi, and S. Mishra, "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)*, 2018, pp. 69–76.

[6] P. Patel, "The appg, islamophobia and anti-muslim racism," *Feminist Dissent*, no. 6, pp. 205–229, 2022.

[7] R. Imhoff and J. Recker, "Differentiating islamophobia: Introducing a new scale to measure islamoprejudice and secular islam critique," *Political Psychology*, vol. 33, pp. 811–824, 12 2012.

[8] B. Vidgen and T. Yasseri, "Detecting weak and strong islamophobic hate speech on social media," *Journal of Information Technology & Politics*, vol. 17, no. 1, pp. 66–78, 2020. [Online]. Available: https://doi.org/10.1080/19331681.2019.1702607

[9] P. Ghasiya and K. Sasahara, "Rapid sharing of islamophobic hate on facebook: The case of the tablighi jamaat controversy," *Social Media + Society*, vol. 8, no. 4, p. 20563051221129151, 2022. [Online]. Available: https://doi.org/10.1177/20563051221129151

[10] M. Chandra, M. Reddy, S. Sehgal, S. Gupta, A. B. Buduru, and P. Kumaraguru, ""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*, ser. HT '21. New York, NY, USA: Association for Computing Machinery, 2021, p. 67–77. [Online]. Available: https://doi.org/10.1145/3465336.3475111

[11] Q. Mehmmod, A. Kaleem, and I. Siddiqi, "Islamophobic hate speech detection from electronic media using deep learning," *Mediterranean conference on pattern recognition and artificial intelligence*, vol. V, pp. 187–200, 01 2022.

[12] H. Khan and J. L. Phillips, "Language agnostic model: Detecting islamophobic content on social media," in *Proceedings of the 2021 ACM Southeast Conference*, ser. ACM SE '21. New York, NY, USA: Association for Computing Machinery, 2021, p. 229–233. [Online]. Available: https://doi.org/10.1145/3409334.3452077

[13] E. Aldreabi, J. M. Lee, and J. Blackburn, "Using deep learning to detect islamophobia on reddit," *The International FLAIRS Conference Proceedings*, vol. 36, May 2023. [Online]. Available: https://journals.flvc.org/FLAIRS/article/view/133324

[14] J. Baumgartner, S. Zannettou, B. Keegan, M. Squire, and J. Blackburn, "The pushshift reddit dataset," *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 14, no. 1, pp. 830–839, May 2020. [Online]. Available: https://ojs.aaai.org/index.php/ICWSM/article/view/7347

[15] "U.s. constitution," https://constitution.congress.gov/constitution/amendment-1/, amendment I.

[16] B. Clarke, "Freedom of speech and criticism of religion: What are the limits?" *MURDOCH UNIVERSITY E LAW JOURNAL*, vol. 14, no. 2, pp. 94–121, 05 2007.

[17] M. Bailey, D. Dittrich, E. Kenneally, and D. Maughan, "The menlo report," *IEEE Security and Privacy*, vol. 10, no. 2, pp. 71–75, Mar. 2012.

[18] C. Rivers and B. Lewis, "Ethical research standards in a world of big data [version 2; peer review: 3 approved with reservations]," *F1000Research*, vol. 3, no. 38, 2014.

[19] U. Balci, C. Ling, E. De Cristofaro, M. Squire, G. Stringhini, and J. Blackburn, "Beyond fish and bicycles: Exploring the varieties of online women's ideological spaces," in *Proceedings of the 15th ACM Web Science Conference 2023*, ser. WebSci '23. New York, NY, USA: Association for Computing Machinery, 2023, p. 43–54. [Online]. Available: https://doi.org/10.1145/3578503.3583618

[20] "Perspective api," 2018. [Online]. Available: https://www.perspectiveapi.com/

[21] B. Rieder and Y. Skop, "The fabrics of machine moderation: Studying the technical, normative, and organizational structure of perspective api," *Big Data & Society*, vol. 8, p. 205395172110461, 07 2021.