Using Deep Learning to Detect Islamophobia on Reddit

Esraa Aldreabi[1], **Justin M. Lee**[2], **Jeremy Blackburn**[1]
Binghamton University[1] Independent Scholar[2]
ealdrea1@binghamton.edu, jblackbu@binghamton.edu

Abstract

Islamophobia, a negative predilection towards the Muslim community, is present on social media platforms. In addition to causing harm to victims, it also hurts the reputation of social media platforms that claim to provide a safe online environment for all users. The volume of social media content is impossible to be manually reviewed, thus, it is important to find automated solutions to combat hate speech on social media platforms. Machine learning approaches have been used in the literature as a way to automate hate speech detection. In this paper, we use deep learning techniques to detect Islamophobia over Reddit and topic modeling to analyze the content and reveal topics from comments identified as Islamophobic. Some topics we identified include the Islamic dress code, religious practices, marriage, and politics.To detect Islamophobia, we used deep learning models. The highest performance was achieved with $\text{BERT}_{\text{base}}$+CNN, with an F1-Score of 0.92.

Introduction

Social media platforms have established rules to prevent hate speech and create a safe online environment for the users (MacAvaney et al. 2019). However, the large content on social media platforms made it impossible to be manually reviewed. This has created a need for automated hate speech detection methods. Machine learning approaches have been used in the literature as a way to automate hate speech detection. Hate speech is commonly defined as communication that insults a person or group based on race, nationality, gender, religion, or other characteristics. Social media platforms' usage policies fail to prevent the dissemination of such content entirely (MacAvaney et al. 2019). This hate speech harms the online communities that experience it and causes damage to the reputations of social media platforms (Burnap and Williams 2015). Moreover, it has been shown that online hate speech provokes violence and serious harm (Yang et al. 2018). This toxic content has motivated researchers to build models to detect it by using advanced machine learning and natural language processing, as well as generating large annotated datasets to train such models. However, several challenges face this research area, including but not limited to: the many types of hate speech, insufficiently labelled datasets, false positive classifications, and inherent biases in the datasets.

In this work, we focus on analyzing hate speech towards Muslim communities on Reddit. (Vidgen and Yasseri 2020) describes Islamophobia as a type of racism, stereotyping, prejudice, fear, and dominance. Furthermore, (Ahmed Khan, Shah, and Ahmad 2020) defines Islamophobia as an unfounded fear or hatred of Islam, Islamic values & traditions, and an unjustified hostility towards Muslims. Moreover, they also found that Islamophobia and violence towards the Muslim community has increased in The United States and Europe since 9/11. Another study by (Mozafari, Farahbakhsh, and Crespi 2020) also reported that there has been increase in hate speech towards immigrant and Muslim communities in The United Kingdom after the Manchester and London terrorist attacks, and in the USA after the election of Trump. From this, it is made apparent that social media platforms require more effective tools to better detect hate speech in order to create safe environments for online communities.

Previous work showed that Islamophobia often increases against Muslim communities after a terrorist attack, linking the entire community to terrorism and violence (Ciftci 2012). Similar increases were observed during the COVID-19 Pandemic, blaming the large gatherings and traditions of Muslims (Chandra et al. 2021)(Ghasiya and Sasahara 2022). In this study we used topic modelling to show that Islamophobia manifests itself in Reddit with different topics and not related specifically to an event. Some revealed topics include the Islamic dress code, religious practises, marriage, and politics. In order to detect Islamophobia, we built deep learning Islamophobic hate speech detection by fine-tuning and integrating BERT with several Neural Networks.

Background and Related Work

In recent years, researchers have used deep learning in various downstream NLP tasks and achieved better performance than traditional machine learning. (Zhang, Robinson, and Tepper 2018) used a CNN and GRU (Gated Recurrent Unit Network) neural network model initialized with pre-trained word2vec embedding to capture both word/character combinations. (Founta et al. 2018) built deep learning architecture that can handle different types of abusive language. (Mozafari, Farahbakhsh, and Crespi 2020) researched the ability of BERT at capturing hate speech on publicly available dataset from Twitter by using fine-tuning methods based on transfer learning. (Chandra et al. 2020) presented a dataset from

Gab posts the data was labeled for abuse presence, target and severity. They experimented with both traditional and deep learning based models and they reported that a BERT based model performed the best. (ElSherief et al. 2018) studied the targets of hate speech if its directed towards a specific person or entity, or towards a group of people sharing a common protected characteristic like religious or nationality groups. Their analyses showed that directed hate speech, is angrier, and often attacks the target using offensive words. And generalized hate speech is dominated by religious hate. There is a lot of research on hate speech detection, but according to (Chandra et al. 2021) Islamophobia has not been researched in depth. Additionally, according to (Belal, Ullah, and Khan 2022), identifying Islamophobic hate speech from other types of offensive language poses a challenge for existing hate speech detection methods. In previous work related to Islamophobia, rather than binary classification, (Vidgen and Yasseri 2020) proposed three classifications of Islamophobic content: 1) non-Islamophobic, 2) weak Islamophobic, and 3) strong Islamophobic. (Ghasiya and Sasahara 2022) and (Vidgen and Yasseri 2020) showed a rise in Islamophobia during the COVID-19 pandemic. (Soral, Liu, and Bilewicz 2020) performed a measurement study of anti-Muslim hate speech and Islamophobia based on where the users consumed news from social media or traditional mass media. They reported that frequent users of social media were exposed to a higher level of of Islamophobic content. (Mehmmod, Kaleem, and Siddiqi 2022) employed deep learning to detect Islamophobic hate speech. (Ahmed Khan, Shah, and Ahmad 2020) used quantitative and qualitative techniques to analyze the Twitter hashtag #stopIslam. (Ghasiya and Sasahara 2022) studied Islamophobic hate on Facebook during the COVID-19 pandemic and showed that anti-Muslim hate groups and individuals had spread misinformation that led to violence against Muslims in India. (Belal, Ullah, and Khan 2022) proposed a transfer learning approach using Universal Language Model Fine Tuning (ULMFIT) to detect Islamophobia over Twitter. (Khan and Phillips 2021) proposed a solution to the multilingual data classification problem by translating the content to English to try to overcome the challenge of detecting Islamophobia over multiple languages.

Methods

To create our dataset, we employed a method of semi-automatic annotation of the comments using Bidirectional Encoder Representations from Transformers (BERT), based on its higher performance as shown in (Chandra et al. 2021).

To create the dataset to train and test our models, we manually collected 2000 comments publicly available on Reddit that are related to Islam without constraints on location or time frame. To collect comments that are related to Islam we used a list of keywords that contained positive, negative, and neutral Islamic-related terms to ensure creating a balanced dataset. The list of keywords: 1) "Muslim," 2) "Islam," 3) "extremist," 4) "terrorists," 5) "sharia," 6) "jihad," 7) "Quran," 8) "Hadith," 9) "Islamophobia," 10) "Taliban," 11) "Hijab," 12) "mosque," 13) "ISIS," 14) "halal," 15) "Jannah," 16) 'Sunni,' 17) "Friday," 18) "Osama," 19) "Mohammad," 20) "Arab," 21) "Masjid," 22) "alhamdulillah," 23) "Madhab." Then we assigned binary labels to the comments based on their propensity for Islamophobia by two annotators as Hateful (1) and non-Hateful (0). The two annotators labelled the data independently. The Cohen's kappa coefficient of the annotation was 0.955, which translates to almost perfect agreement between the annotators. We followed these guidelines to label the data. We labelled comments as hateful if the comment expresses negativity against Muslims such as calling Muslims terrorists or extremists, advocating to ban them from entering countries or practicing their religion, or character assassinating their religious idols. The comments that were labelled as non-hateful were about Islam but non-Islamophobic such as discussing practises, miracles and teachings, and history. Out of the 2000 comments, 1022 were hateful and 977 non-hateful. We used this labelled dataset to train the models. We aimed to build deep learning Islamophobic hate speech detection by fine-tuning and integrating BERT with several NNs due to its recent high performance in hate speech detection task as demonstrated in (Mozafari, Farahbakhsh, and Crespi 2020). In text classification, data needs to be pre-processed by cleaning and representing in an appropriate form for further processing. For cleaning, we converted all comments to lowercase, and we removed punctuation marks, unknown Unicode characters, and delimiters. For the NN implementation, we used the PyTorch-pre-trained-Bert library that contains the pre-trained BERT model, text tokenizer, and pre-trained WordPiece. To integrate NNs with BERT, we compared the results of two models: pre-trained BERT with a Convolutional Neural Networks (CNN) inserted and pre-trained BERT with Long Short-Term Memory Networks (LSTM) inserted.

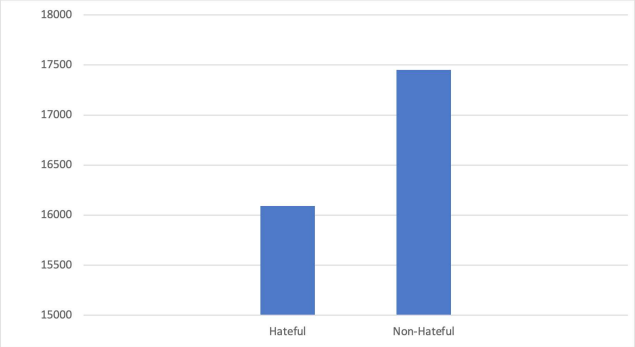

Figure 1: The distribution of comments after labeling

For our experiment, we trained the classifiers for 10 epochs with a batch size of 32. We used Adam optimizer with a learning rate of 10^{-6} and the dropout probability set to 0.2. For input, we used a BERT tokenizer to tokenize each comment. Based on the original BERT (Devlin et al. 2019), we used WordPiece tokenization to split words into sub-word units. We split the dataset to 80% for training, 10% for validation, and 10% for testing. We used stratified sampling to select 0.8, 0.1, and 0.1 to avoid overfitting. For evaluation, we used the test dataset, and considered three differ-

Table 1: Evaluation results

Method	Precision	Recall	F1-Score
$BERT_{base}$+CNN	0.92	0.92	0.92
$BERT_{base}$+LSTM	0.895	0.90	0.895

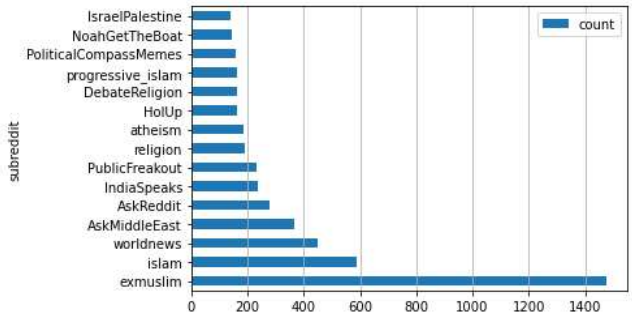

Figure 2: Distribution of comments over subreddits

ent metrics: F1-scores, precision, and recall. The results are summarized in Table 1.

We collected 40K comments from Reddit that contains 'Islam' or 'Muslim' keywords for further analysis without constraints on location or time frame. We ran a detect language algorithm and kept the comments in English language only. To classify the comments,we considered F1-score as the most robust metric for the model performance. The F1-score of $BERT_{base}$+CNN is 92% and The F1-score of $BERT_{base}$+LSTM is 90%. Based on the higher F1-score by $BERT_{base}$+CNN we used this model to classify the 40K comments and use them for further analysis. The distribution after classification is shown in Figure 1.

Topic Modeling and Content Analysis

The comments that were classified as Islamophobic were spread out over 1,793 subreddits. Figure 2 shows the top 15 subreddits. The "ex-Muslim" subreddit has the highest number of Islamophobic comments followed by "Islam" and "world news."

To analyze the content difference between the hateful and non-hateful comments we performed topic modeling and compared the results.

To reveal the prevalent topics in our dataset, we performed topic modeling, we used Non-Negative Matrix Factorization(NMF). We first pre-processed the comments by removing stop words, white spaces, and punctuation, and then lower-casing. To determine the number of topics to best represent the data, we used the Coherence Score. We iterated through the number of topics from 5 to 50 with a step size of 5. The vectors that we used in NMF were created by Term Frequency-Inverse Document Frequency(TF-IDF)-based vectorization method. When creating the topics, we separated the datasets into two subsets based on the label: 1) hateful and 2) non-hateful. In Table 2, we see tokens from comments that are not Islamophobic. They represent discussions without offensive words. The topics regard the Islamic dress code, diet restrictions, Abrahamic religious teachings, Arab and Islamic cultural and marriage traditions. In Table 3, we see tokens from the comments that are Islamophobic.Topic 1 relates to marriage traditions and rules in Islam and uses words like "rape" and "slavery." Some topics are close to each other like (Topics: 2, 3, and 6) which are related to terrorist groups and violence. Topic 2 is about terrorist attacks and groups involved. Topic 3 describing Islam as a cult with barbaric actions that force others to convert to Islam. Topic 4 has a political aspect related to the Israel-Palestine conflict. Topic 5 criticizes women's clothing in Islam. Topics 7 and 8 concern Christianity, Islam, atheism, and anti-Abrahamic sentiment. Topic 9 has an abundance of generally offensive language intermixed with age-related criticisms of the Muslim community, including admonitions of pedophilia. While some topics appear only in the hateful comments, there is some crossover with non-hateful comments. For example, the discussions of women being forced to wear Hijab and the discussions of women who wear it by choice have similar tokens, as these topics tend to use similar vocabulary. Overall, we observed that the Reddit comments identified as hateful had a tendency to be directed towards the Muslim community in general and not towards other users. Identity attacks on the Muslim community at large were more common than those against specific individuals. We see in comments blaming the Muslim community for terrorist attacks (such as those committed by groups like ISIS) and forcing non-believers to convert to Islam a tendency to infer that Islam is fundamentally a threat to others. Islamophobic comments laden with such rhetoric could incite the spread of irrational fear, discrimination, and physical attack toward Muslims. We see accusations of allowing child marriage, forcing women to wear the Hijab, and condoning rape in Islam which engenders a negative impression that Islam is an anti-feminist ideology which in turn may provoke aggression toward Muslim women who wear the Hijab. The framing of Islam as a threatening ideology on social media could negatively influence the greater community and facilitate the development of racist attitudes, negative perceptions, and even violence towards Muslims.

External URL

Some of the comments in the dataset contain URLs referencing external content. In order to analyze this content, we extracted URLs from the comments and counted the frequency of each web domain. The most frequently referenced domains were Wikipedia and YouTube. There were also URLs to posts on social media platforms such as Reddit, Twitter, and Facebook, news media outlets like CNN, BBC, and Al Jazeera, and Islamic Websites like Sunnah.com and Quran.com.

We grouped URLs by the label of the comment referencing them and performed a word frequency calculation on their content. With corresponding sets of Islamophobic and non-Islamophobic word frequencies from linked content, we focused on two domains in particular: Wikipedia and YouTube. We used the Request library to get the title for YouTube videos and Wikipedia articles. We pre-processed the titles by removing stop words, lower casing, and using lemmatizer. Then To get the topic of the referenced URL's

Table 2: Tokens from the comments that were classified as non-hateful

Seq.No	Tokens	No. of Comments
1	wear hijab women force muslim cover choice clothe woman dress girls school	1083
2	muslim halal sunni mosque nonmuslim country christian every hindu become majority	2269
3	allah prophet may messenger upon ibn peace bless hadith sin swt muhammad	1590
4	god quran believe jesus read prophet book bible verse word muhammad hadith true	2141
5	islam sunni convert christianity leave shia quran islamic slave follow allow slavery	2498
6	women men marry muslim wive husband man woman marriage right wife allow slave	1154
7	religion christianity islam believe practice follow culture force beliefs law	1042
8	arab muslim culture arabic israel jews country jewish name language state speak	1099
9	muslims kill law sharia india majority country non christians islamic isis pakistan	1978

Table 3: Tokens from the comments that were classified as Hateful

Seq.No	Tokens	No. of comments
1	women men sex rape slave marry muslim husband wive wife allow marriage	760
2	muslims kill terrorists india attack terrorist christians support hate islamic isis	1257
3	islam cult christianity extremist slavery barbaric isis leave convert death taliban	1524
4	arab jews israel arabs jewish palestinians state palestine land muslim war	1040
5	wear hijab women cover force muslim choice clothe woman dress hair religious	906
6	muslim terrorists extremist christian halal hindu guy rape isis taliban country	1490
7	god allah quran believe prophet jesus muhammad verse read hadith say bible word sin	954
8	religion christianity islam believe judaism beliefs follow church atheist practice	1288
9	fuck shit islam pedophile old ass racist bullshit child year get piece idiot man	729

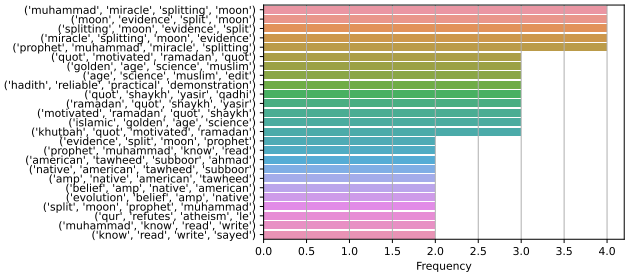

Figure 3: Top 25 non-Islamophobic topics from URLs

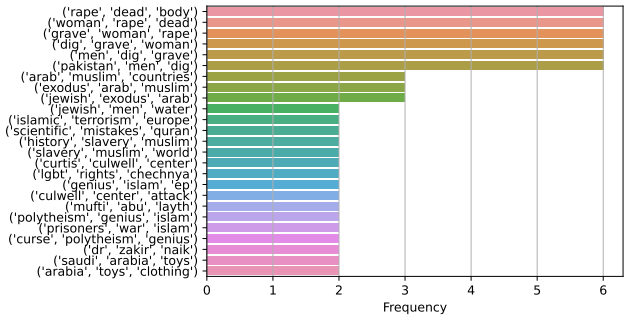

Figure 4: Top 25 Islamophobic topics from URLs

article or video we used Bag of Words (BoW) with Ngrams of 3. In Figure 3 we see the top 25 topics from comments classified as non-Islamophobic and in 4 the top 25 topics from comments classified as Islamophobic. From the topics we can see that non-hateful comments tend to reference articles and videos about Islamic speakers, miracles performed by Mohammad, or Islamic history. In contrast, hateful comments tend to share articles and videos of terrorist attacks, scientific mistakes in the Quran, or disturbing events like Muslims engaging in necrophilia. Based on this, we can see that external URLs may be used to spread Islamophobic content.

Conclusion and Future work

The propagation of hate and the incitement of violence can harm the reputations of social media platforms and potentially lead to user abandonment. Moreover, it could create misleading impressions and discrimination against attacked communities. To prevent this, social media platforms have established rules to prohibit hate speech. However, the sheer amount of content on social media platforms has made it impossible for hate speech to be effectively removed using manual methods. Automated hate speech detection methods provide a possible solution for this problem. This paper proposed deep learning Islamophobic hate speech detection models and used the highest performing model to detect and analyze Islamophobia on Reddit. The classifier had an F1-score of 92%. By performing these content analyses, we were able to identify Islamophobic comments and categorize them within several topics. For future work, data can be collected from additional languages and sources; therein potential differences can be identified between topics where Islamophobia manifests.

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

References

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.

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.

Chandra, M.; Pathak, A.; Dutta, E.; Jain, P.; Gupta, M.; Shrivastava, M.; and Kumaraguru, P. 2020. AbuseAnalyzer: Abuse detection, severity and target prediction for gab posts. In *Proceedings of the 28th International Conference on Computational Linguistics*, 6277–6283. Barcelona, Spain (Online): International Committee on Computational Linguistics.

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.

Ciftci, S. 2012. Islamophobia and threat perceptions: Explaining anti-muslim sentiment in the west. *Journal of Muslim Minority Affairs* 32(3):293–309.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv* abs/1810.04805:4171–4186.

ElSherief, M.; Kulkarni, V.; Nguyen, D.; Wang, W.; and Belding, E. 2018. Hate lingo: A target-based linguistic analysis of hate speech in social media. *Proceedings of the International AAAI Conference on Web and Social Media* 12.

Founta, A.-M.; Chatzakou, D.; Kourtellis, N.; Blackburn, J.; Vakali, A.; and Leontiadis, I. 2018. A unified deep learning architecture for abuse detection. *Proceedings of the 10th ACM Conference on Web Science* 105–114.

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.

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.

MacAvaney, S.; Yao, H.-R.; Yang, E.; Russell, K.; Goharian, N.; and Frieder, O. 2019. Hate speech detection: Challenges and solutions. *PLoS ONE* 14.

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.

Mozafari, M.; Farahbakhsh, R.; and Crespi, N. 2020. A bert-based transfer learning approach for hate speech detection in online social media. In Cherifi, H.; Gaito, S.; Mendes, J. F.; Moro, E.; and Rocha, L. M., eds., *Complex Networks and Their Applications VIII*, 928–940. Cham: Springer International Publishing.

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.

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.

Yang, D.; Heaney, T.; Tonon, A.; Wang, L.; and Cudré-Mauroux, P. 2018. Crimetelescope: crime hotspot prediction based on urban and social media data fusion. *World Wide Web* 21:1323–1347.

Zhang, Z.; Robinson, D.; and Tepper, J. 2018. Detecting hate speech on twitter using a convolution-gru based deep neural network. In Gangemi, A.; Navigli, R.; Vidal, M.-E.; Hitzler, P.; Troncy, R.; Hollink, L.; Tordai, A.; and Alam, M., eds., *The Semantic Web*, 745–760. Cham: Springer International Publishing.