Social Media and Online Islamophobia: A Hate Behavior Detection Model

Abdulwahab Ali Almazroi^{1†}, Asad Ali Shah^{1††} and Fathey Mohammed^{2†††}

1: College of Computing and Information Technology at Khulais, Department of Information Technology, University of Jeddah, Jeddah, Saudi Arabia
2: School of Computing, Universiti Utara Malaysia (UUM), 06010 Sintok , Kedah, Malaysia

Abstract
Muslim community has faced a lot of hatred towards them due to the rise in Islamophobia. Taking no measures to control Islamophobia can create fear among the Muslim community while at the same time giving others an open hand to spread hate and toxic remarks toward Muslims. While Muslim leaders and countries are taking measures to stop Islamophobia through awareness and building content to share Islam's true peaceful and moderate image, it does not help in controlling the spread of Islamophobia on social media platforms. In this regard, this research proposes a framework capable of detecting Islamophobic content. The proposed solution achieves this using natural language and artificial intelligence techniques such as keyword detection, tone analyzer, machine learning, impartiality ratio, and more. The proposed model is also capable of categorizing comments based on their severity and context. The research is hopeful that the proposed framework would allow experts to detect such posts causing Islamophobia early and report them so they can be taken down timely before being widespread. The successful completion of this research will not only have positive implications for the Muslim community but will also allow experts and researchers from other areas to use the same model in combating hateful and toxic speech on other platforms.

Keywords:
Social Media; Detection; Islamophobia; Hate

1. Introduction

The web has grown into one of the largest repositories of data and Social media platforms are one of the most used services on it. The reason for its popularity is the ability for others to share their opinion, thoughts, and reviews without going through an editorial process. According to one report, the number of users using social media platforms was 2.86 Billion and it is expected to reach 4.41 Billion by 2025 [1]. This shows that a huge number of users use these platforms. This is why many companies and new agencies have started to shift towards social media as well for communication and advertisement. While these social media platforms are a great way to share your opinion, they can be used negatively as well [2-5]. Hate speech is one of the major problems that the world is facing due to the rise of Social media. Due to freedom of speech, one cannot be apprehended just for sharing his opinion [6, 7]. However, these platforms are being used to spread hate, toxic remarks, and shaming targeted hosts which have resulted in these people conducting social boycotts and avoiding people, but in extreme cases, it results in people getting hurt or worse being murdered as a result. One example of such an incident is the murder of Mashal Khan where an angry mob of students instigated a false claim towards a Muslim student for committing blasphemy which result in his murder [8]. While the perpetrators were apprehended but it was too little too late. Similarly, events like these are occurring throughout the globe which is why some automatic tool is necessary for detecting hate speech.

Detecting hate speech is a complex problem [9, 10]. This is because that while hateful speech should be stopped and action should be taken against but it should not falsely accuse a valid opinion. A non-hateful comment may also be tagged as one as a result if not checked carefully. Moreover, it also goes against freedom of speech. Thus, its implementation should be handled with care and consideration.

Adding an extra layer to his problem is the detection of Islamophobic comments [3, 10]. Islamophobia is defined as an opinion that shows sign of dislike or hatred towards the religion Islam. Thus, it falls under a special case of hate speech that is targeted towards Islamic only. Moreover, religion is a sensitive topic and careful classification between Islamophobic content and non-Islamophobic content is important.

Researchers have come up with solutions that provide insights like the tone of a sentence or a person's impartiality [11-14]. These factors have

Manuscript received March 5, 2025
Manuscript revised March 20, 2025
https://doi.org/10.22937/IJCSNS.2025.25.3.9

helped companies in judging reviews or analyzing how customers are reacting to their new product and how they can improve it [15]. The same is being used to detect online hate speech and threats being made toward others. However, this is a challenging problem and it may be difficult to decide whether a comment is targeted toward a certain group of people due to their race or religion. This is why more factors are needed to be incorporated to make it successful.

While much work has been done on the detection of hateful speech, there is limited research done on detecting Islamophobic content. Simply applying hateful speech detection is not sufficient and requires additional factors to be added for its detection. Thus, this research wishes to propose a framework that will allow researchers and experts to build their own Islamophobic text detection systems.

Many researchers are actively working on this to detect comments containing Islamophobia [2, 3, 16-19]. While some of the researchers are using trivial techniques while others are using complex ones. However, there is still limited research in the area which is why this research wishes to work more on the area. This research wishes to explore the different methods and techniques that can be useful in hate and toxic speech detection. For this, the research wishes to investigate natural language processing techniques such as regular expression, keyword detection, tone analyzer, and impartiality rating. In addition to this, the research wishes to categorize the comments based on different categories. For example, whether the comment is related to a certain topic, or an event, what is the level of severity and much.

In this regard, this research makes the following contributions:

1) *Features identification: The first contribution made by this paper is summarizing different features identified in literature that can be utilized by researchers and experts in building a framework for combating islamophobia. These factors can be used to building systems to combat Islamophobia and train classifiers accordingly. Moreover, further research can be done on these features to see which one of these features have more weightage over others.*
2) *Islamophobia detection framework: The second contribution made by this paper is proposing a framework for identifying islamophobia content. This framework will help researchers in building their own systems from scratch as per their research requirements.*

The rest of the paper is structured as follows. Chapter 2 covers existing research done in the area. Chapter 3 covers the methodology and framework of the proposed system. Finally, the paper is concluded in chapter 4 and along with future directions.

.

2. Related Work

Work on hate speech detection dates back to 1990 since the advent of the internet [20]. While some state-of-the-art hate speech detection systems have been looked into to understand their design and characteristic that are similar in nature to a Islamophobia detection system, the focus of this research will be towards the latter mostly.

Vidgen and Yasseri [3] proposed the use of a multi-class classifier instead of treating it as a binary task. This was achieved by using gloVe word embedding model and using multiple classifiers from which Support Vector Machines produced the best results. This allowed the system to categorize classify Twitter tweets (dataset having 109,488 tweets) into three categories including 1) non-Islamophobic content, 2) weak-Islamophobic content, and 3) strong-Islamophobic content. The system achieved an accuracy of 77.6% and a balanced accuracy of 83% on 1-fold classification and an accuracy of 73.5% and 79.8% respectively on 10 fold classification.

Mehmood, et al. [17] proposed the use of deep learning for detecting Islamophobic content on Twitter. The proposed model makes use of a one-dimensional Convolution Neutral Network to perform feature extraction. These features are then used to perform classification using a Bi-directional Long Short-Term Memory network classifier. While other variations of Convolution Neutral Network and recurrent layers were evaluation, but the combination of Convolution Neutral Network and Bi-direction Long Short-Term Memory network produced the best results. The proposed system was able to achieve training accuracy of 92.39 and a test accuracy of 90.13 using the proposed approach.

González-Pizarro and Zannettou [21] suggested the use of contrastive learning to detect Antisemitism and Islamophobic content. It should be noted that contrastive learning differs from Machine Learning and Deep Learning classification as the form relied on self-learning and self-supervision whereas the latter need supervision for training the dataset before they can start predicting successfully. The proposed system makes use of Google's Perspective API and finding specific keywords phrases that make them Antisemitism or Anti-Islamic content. Moreover, OpenAI's contrastive language-image Pretraining was also used to extract text from images and find connections between them. This system was tested on 66 million posts and 4.8 Million images and achieved an accuracy score of 81%.

Chandra, et al. [22] analyzed the behavior and reasoning behind the sudden increase in Islamophobic content during the COVID-19 outbreak. The research used longitudinal analysis to find links to Islamophobic content of users with kind of religious events that they attended. Moreover, the system also performs content analysis to find features that can help in classifying the content as Islamophobic. Moreover, the system also analyzed the user profile making the comment/tweet by analyzing his tweet history, people followed, etc.

Vidgen [23] PhD thesis also investigates this topic deeply and evaluated various techniques for detecting Islamophobic content. The research explores features that can help in detecting Islamophobic content easily. This includes categorizing Islamophobic content including "Fear and Anxiety", "Threat", "Negativity", "Difference", "Stereotyping" and more. The research also uses surface and derived features, language syntax and word embedding to help in better identification of Islamophobic content. The research used various algorithms for evaluating accuracy including Naïve-Bayes, Random Forests, Support Vector Machines, Logistic Regression, Decision Trees, and Deep Learning with epochs. Among the different classifiers the system achieved the best results with Support Vector Machines achieving an accuracy of 72.17%. Moreover, instead of adding all possible features, the research suggested that the combination of six features produced better results over adding more or less than the suggested number of features.

From the various systems reviewed, regardless of the classification method chosen, performing pre-processing and feature extraction is key to achieving high accuracy. In addition to conventional hate detection systems, some additional steps need to be added to feature extraction to assist the classifiers in identifying Islamophobic content correctly. In the next section this research presents the proposed framework to accomplish this task.

3. Framework

This section will cover the proposed framework for identifying Islamophobic content. From literature various systems were reviewed highlight steps that can be taken to classify the content accurately. Moreover, different systems have highlighted different and unique features for performing classification. The framework proposed by this research summarizes these features in the framework to allow researchers and experts to select the features they prefer or the classifier suitable for their use case.

The proposed framework is divided into four major modules including 1) Data Pre-processing, 2) Feature Extraction, 3) Classification, 4) Evaluation. These modules are also shown in Figure 1.

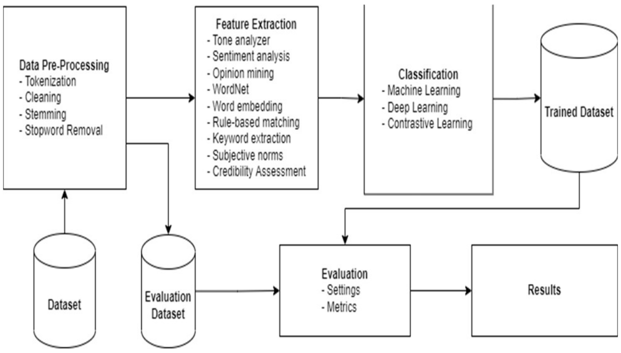

Fig. 1. Islamophobic content detection framework

3.1 Data Pre-Processing

Usually, the datasets selected that not in an organized format or have unnecessary information attached to them. For smoother processing it is required that this data is cleaned before it can be used for feature extraction. In this regard, several methods can be used as highlighted in Figure 1.

Some of the methods that are performed in the pre-processing module are discussed briefly. Cleaning involved removing or filtering out unwanted information. This may involve information that is unwanted whose inclusion may increase the processing time or lead to abnormal results. Thus, cleaning is performed to remove unwanted data. Tokenization is the process to divide the text/content into smaller units for better organization and easier processing. This can be achieved using a natural language processing unit that can utilize part-of-speech tagger, named entity recognizer, and normalization. A better organization can allow additional information to be extracted while allow easier processing in the later stages. Stopword removal is applied to removed words that have little to no significance towards further processing. All of these steps and other pre-processing steps help in bringing the raw dataset to a state in which it becomes more useful, and on which feature extraction can be performed easily and will produce meaningful results.

3.2 Feature Extraction

Feature extraction refers to mining meaningful data from the pre-processed data. This is done so that the features can be used to find patterns and trends that can help in distinguishing between non-Islamophobic and Islamophobic content easily. In some cases, additional methods are applied on the pre-processed data to extract additional information that can also act as features.

The research will highlight some of the features and methods that can be used in this module. Tone Analyzer can be applied to highlight the feeling and attitude of the sentence. This can help in classification by finding whether the sentence contains anger, sadness, aggression, neutrality, etc. This feature can also help further classification in Islamophobic content based on the intensity and category of the tone analyzed by the tone analyzer. Sentiment analysis is used to check the biasness of the sentence. This can highlight whether the sentence is positive, negative, or neutral. In most cases sentences have extreme sentiment values highlight warning flags towards being Islamophobic content. Similarly, opinion mining can also be applied to filter certain types of opinions that may be flagged as Islamophobic. Opinion mining can also be if the focus is towards a certain type of Islamophobic content only. WordNet is useful for finding alternative words that help in understanding the semantic of the content better.

Word embedding also adds additional details to words in a sentence that can help in finding association between words and thus understanding the semantics of the sentence. Rule-based matching is useful to shortlist Islamophobic content quickly if they match a certain rule. These can include sentences and combination of words that are often used by people who are trying to instill anger and hatred among Muslims by writing Islamophobic content. Keyword extraction is a weaker form of rule-based matching in which the method highlights sentences contain a particular keyword. These sentences that be explored further to check whether they contain Islamophobic content or not. Subjective norms include rules that apply to a certain scenario only. Considering this research is focused towards identifying Islamophobia only then its important to write methods or reasoners that are able to understand some of the norms that are used in Islam which may be unrecognizable by a normal language parser. This may include expanding the vocabulary, or adding rules to a rule-based language, etc. Lastly, credibility assessment is also vital to see the credibility of the author, or the credibility of the content written by checking it logically and factually. These are some of the features and methods identified by our research but is not limited to these only.

Additional features may also be added to improve the accuracy of the system. These features and additional data extract is forward to the classification module for further processing.

3.3 Classification

This module is responsible for using the data provided by features and methods and use them to find trends and patterns in the data. Depending upon the classifier used, the system can be trained through supervision or not. Once done a trained dataset is produced using which new incoming dataset can be predicted whether it Islamophobic or not.

3.4 Evaluation

The last module is used for testing purpose to see the performance of the overall system. Here the results are compared against ground truths to see how the system performed. For evaluation several metrics are available including accuracy, F-1 measure, precision and recall.

4. Conclusion

Combating hateful and toxic comments on the internet has been a major issue for a long time. A special case of this includes combating Islamophobic content which is growing at an alarming rate. While many researchers have proposed systems for combating hate and toxic speech, there is little to no research done in building systems that can detect Islamophobic content. Thus, this paper focuses on proposing a framework that can identify Islamophobic content. This is done by reviewing existing literature in the area and highlighting the features used in such systems as well as proposing a system that is capable of identifying Islamophobic content.

This research believes that there are many positive implications of this research, and it will certainly help in stopping or minimizing islamophobia. Moreover, it can also help in promoting more content that shows the moderate image of Islam by encouraging scholars or people who talk with reason and logic. Moreover, this research will also open the door for other researchers who wish to pursue more complex methods to improve the system further.

Acknowledgment

This work was funded by the University of Jeddah, Jeddah, Saudi Arabia, under grant No. (UJ-22-MRI-3). The authors, therefore, acknowledge with thanks the University of Jeddah for its technical and financial support.

References

[1] STATISTICA, "Number of social network users worldwide from 2017 to 2025," 2022.

[2] S. MacAvaney, H.-R. Yao, E. Yang, K. Russell, N. Goharian, and O. Frieder, "Hate speech detection: Challenges and solutions," *PloS one,* vol. 14, p. e0221152, 2019.

[3] B. Vidgen and T. Yasseri, "Detecting weak and strong Islamophobic hate speech on social media," *Journal of Information Technology & Politics,* vol. 17, pp. 66-78, 2020.

[4] A. A. Shah, S. D. Ravana, S. Hamid, and M. A. Ismail, "Web pages credibility scores for improving accuracy of answers in web-based question answering systems," *IEEE Access,* vol. 8, pp. 141456-141471, 2020.

[5] F. Alkomah, S. Salati, and X. Ma, "A New Hate Speech Detection System based on Textual and Psychological Features," *International Journal of Advanced Computer Science and Applications(IJACSA),* vol. 13, 2022.

[6] J. W. Howard, "Free speech and hate speech," *Annual Review of Political Science,* vol. 22, pp. 93-109, 2019.

[7] R. T. Mutanga, N. Naicker, and O. O. Olugbara, "Detecting Hate Speech on Twitter Network using Ensemble Machine Learning," *International Journal of Advanced Computer Science and Applications,* vol. 13, 2022.

[8] S. Kermani. (2017, Could a student's death change Pakistan's blasphemy laws? Available: https://www.bbc.com/news/world-asia-39665102

[9] L. E. A. Vega, J. C. Reyes-Magaña, H. Gómez-Adorno, and G. Bel-Enguix, "MineriaUNAM at SemEval-2019 task 5: Detecting hate speech in Twitter using multiple features in a combinatorial framework," in *Proceedings of the 13th international workshop on semantic evaluation*, 2019, pp. 447-452.

[10] S. R. Ahmad, M. Z. M. Rodzi, N. S. Shapiei, N. M. M. Yusop, and S. Ismail, "A review of feature selection and sentiment analysis technique in issues of propaganda," *International Journal of Advanced Computer Science and Applications,* vol. 10, 2019.

[11] A. A. Almazroi, "A fast hybrid algorithm approach for the exact string matching problem via berry ravindran and alpha skip search algorithms," *Journal of Computer Science,* vol. 7, p. 644, 2011.

[12] A. A. Almazroi, O. A. Mohamed, A. Shamim, and M. Ahsan, "Evaluation of State-of-the-Art Classifiers: A Comparative Study," 2020.

[13] A. A. Almazroi, F. Mohammed, M. A. Qureshi, A. A. Shah, I. A. T. Hashim, N. H. Al-Kumaim*, et al.*, "A Hybrid Algorithm for Pattern Matching: An Integration of Berry-Ravindran and Raita Algorithms," in *International Conference of Reliable Information and Communication Technology*, 2022, pp. 160-172.

[14] A. A. Almazroi, A. A. Shah, A. A. Almazroi, F. Mohammed, and N. H. Al-Kumaim, "A Survey of the Hybrid Exact String Matching Algorithms," Cham, 2022, pp. 173-189.

[15] A. Al Marouf, R. Hossain, M. R. K. R. Sarker, B. Pandey, and S. M. T. Siddiquee, "Recognizing language and emotional tone from music lyrics using IBM Watson Tone Analyzer," in *2019 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT)*, 2019, pp. 1-6.

[16] H. Khan and J. L. Phillips, "Language agnostic model: detecting islamophobic content on social media," in *Proceedings of the 2021 ACM Southeast Conference*, 2021, pp. 229-233.

[17] Q. Mehmood, A. Kaleem, and I. Siddiqi, "Islamophobic Hate Speech Detection from Electronic Media Using Deep Learning," in *Mediterranean Conference on Pattern Recognition and Artificial Intelligence*, 2022, pp. 187-200.

[18] W. Yin and A. Zubiaga, "Towards generalisable hate speech detection: a review on obstacles and solutions," *PeerJ Computer Science,* vol. 7, p. e598, 2021.

[19] M. J. Althobaiti, "BERT-based Approach to Arabic Hate Speech and Offensive Language Detection in Twitter: Exploiting Emojis and Sentiment Analysis," *International Journal of Advanced Computer Science and Applications,* vol. 13, 2022.

[20] T. M. Massaro, "Equality and freedom of expression: The hate speech dilemma," *Wm. & Mary L. Rev.,* vol. 32, p. 211, 1990.

[21] F. González-Pizarro and S. Zannettou, "Understanding and Detecting Hateful Content using Contrastive Learning," *arXiv preprint arXiv:2201.08387,* 2022.

[22] 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*, 2021, pp. 67-77.

[23] B. Vidgen, "Tweeting Islamophobia," University of Oxford, 2019.

Abdulwahab Ali Almazroi received the M.Sc. and the Ph.D. in computer science from the University of Science, Malaysia, and Flinders University, Australia, respectively. He is currently serving as an Associate Professor in the Department of Information Technology, College of Computing and Information Technology at Khulais, University of Jeddah, Saudi Arabia. His research interests include parallel computing, cloud computing, wireless communication, and data mining.

Asad Ali Shah received the B.Sc. degree in computer science from COMSATS University, Islamabad, Pakistan, in 2006, the M.Sc. degree in advanced computer science from The University of Manchester, Manchester, U.K., in 2008, and the D.Phil. degree in computer science from the University of Malaya, Kuala Lumpur, Malaysia, in 2017. He is currently serving as an Assistant Professor at Department of Information Technology, College of Computing and Information Technology at Khulais, University of Jeddah, Saudi Arabia. His research interests include information retrieval, information processing, web credibility, and question answering systems. He has produced several articles in the area.

Fathey Muhammad received his B.Sc in Computer Engineering from Esfahan University, Esfahan, Iran in 2003, M.Sc in Information Technology from Tarbiat Modarres, Tehran, Iran in 2005 and Ph.D in Information Systems from Universiti Teknologi Malaysia (UTM), Johor, Malaysia in 2017. His research interests include cloud computing, technology innovation adoption, information system project management, e-government and e-business. He has authored and co-authored over 40 scientific papers in the area of cloud computing services, technology adoption and impact on the performance of organizations, e-government and e-business and others in highly prestigious journals and international conferences, as well as editing 4 books published by Springer. He also appointed as Program Chair and publication committee chair for a number of international conferences, and guest editor for some Scopus journals. Fathey Mohammed is currently, an international senior lecturer in School of Computing, Universiti Utara Malaysia UUM, Malaysia.