International Journal of Engineering Trends and Technology
ISSN: 2231–5381 / https://doi.org/10.14445/22315381/IJETT-V71I11P203

Volume 71 Issue 11, 27-32, November 2023

Original Article

Social Media and Online Islamophobia: A Hate Behavior Detection Model

Abdulwahab A. Almazroi[1*], Asad A. Shah[1], Fathey Mohammed[2]

[1]*College of Computing and Information Technology at Khulais, Department of Information Technology, University of Jeddah, Jeddah, Saudi Arabia.*
[2]*Sunway Business School, Sunway University, Selangor, Malaysia.*

[*]*Corresponding Author : aalmazroi@uj.edu.sa*

Received: 26 April 2023 Revised: 12 June 2023 Accepted: 25 September 2023 Published: 04 November 2023

Abstract *- Since 9/11, the 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 control 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 data repositories, and Social media platforms are one of the most used services. 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, expected to reach 4.41 billion by 2025 (STATISTICA, 2022). 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 (Alkomah, Salati, & Ma, 2022; MacAvaney et al., 2019; Shah, Ravana, Hamid, & Ismail, 2020; Bertie Vidgen & Yasseri, 2020). 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 (Howard, 2019; Mutanga, Naicker, & Olugbara, 2022). However, these platforms are being used to spread hate, toxic remarks, and shame 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 resulted in his murder (Kermani, 2017). While the perpetrators were apprehended, 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 (Ahmad, Rodzi, Shapiei, Yusop, & Ismail, 2019; Vega, Reyes-Magaña, Gómez-Adorno, & Bel-Enguix, 2019). This is because while hateful speech should be stopped and action should be taken against it, 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 detecting Islamophobic comments (Ahmad et al., 2019; Bertie Vidgen & Yasseri, 2020). Islamophobia is defined as an opinion that shows signs of dislike or hatred towards the religion Islam. Thus, it falls under a special case of hate speech targeted toward Islam 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 into the tone of a sentence or a person's impartiality (Almazroi, 2011; Almazroi, Mohamed, Shamim, & Ahsan, 2020; Almazroi, Mohammed, et al., 2022; Almazroi, Shah, Almazroi, Mohammed, & Al-Kumaim, 2022). These factors have helped companies in judging reviews or analyzing how customers are reacting to their new product and how they can improve it (Al Marouf, Hossain, Sarker, Pandey, & Siddiquee, 2019). 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 Islamophobic text detection systems.

Many researchers are actively working on this to detect comments containing Islamophobia (Althobaiti, 2022; Khan & Phillips, 2021; MacAvaney et al., 2019; Mehmood, Kaleem, & Siddiqi, 2022; Bertie Vidgen & Yasseri, 2020; Yin & Zubiaga, 2021). 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 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 more.

In this regard, this research makes the following contributions:

- Features identification: The first contribution made by this paper is summarizing different features identified in the literature that researchers and experts can utilize in building a framework for combating Islamophobia. These factors can be used to build systems to combat Islamophobia and train classifiers accordingly. Moreover, further research can be done on these features to see which one of these features has more weightage over others.
- 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 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, along with future directions.

2. Related Work

Work on hate speech detection dates back to 1990, since the advent of the internet (Massaro, 1990). While some state-of-the-art hate speech detection systems have been looked into to understand their design and characteristics that are similar to an Islamophobia detection system, the focus of this research will be on the latter mostly.

Bertie Vidgen and Yasseri (2020) proposed using a multi-class classifier instead of treating it as a binary task. This was achieved using the gloVe word embedding model and 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.

Duwairi, Hayajneh, and Quwaider (2021) proposed a framework for detecting hateful speech in Arabic tweets. The research evaluated the capabilities of various deep learning models on their capability in successfully categorizing and detecting hateful content on Twitter. The system has various options, including binary classification, Ternary classification, and multi-class classification. The system was evaluated on an ArHS dataset containing 23,678 tweets. The results showed that BiLSTM-CNN performed the best on binary and ternary classification, while both CNN-LSTM and BiLSTM-CNN achieved better results on multi-class classification.

Alraddadi and Ghembaza (2021) proposed a system capable of detecting and classifying Arabic text for anti-Islamic text. The system makes use of machine learning models, including Support Vector Machines and

Multinomial Naive Bayes for classification. The results showed that Support Vector Machines combined with term frequency feature extraction achieved a result of 97% in detection and classification.

Mehmood et al. (2022) proposed using 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 evaluated, combining the Convolution Neutral Network and Bi-direction Long Short-Term Memory network produced the best results. The proposed system was able to achieve a training accuracy of 92.39 and a test accuracy of 90.13 using the proposed approach.

González-Pizarro and Zannettou (2022) 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 relies on self-learning and self-supervision, whereas the latter needs supervision for training the dataset before they can start predicting successfully. The proposed system makes use of Google's Perspective API and finds specific keyword 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. (2021) 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 the Islamophobic content of users with the kind of religious events they attended. Moreover, the system also performs content analysis to find features that can help classify the content as Islamophobic. Moreover, the system also analyzed the user profile comments and tweets by analyzing his tweet history, and people followed, etc.

B. Vidgen (2019) PhD thesis also investigates this topic deeply and evaluates 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 better identify Islamophobic content. The research used various algorithms to evaluate 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 combining six features produced better results over adding more or less than the suggested number of features.

From the various systems reviewed, performing pre-processing and feature extraction is key to achieving high accuracy regardless of the classification method chosen. In addition to conventional hate detection systems, some steps need to be added to feature extraction to assist the classifiers in correctly identifying Islamophobic content. 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 the literature, various systems were reviewed to 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, and 4) Evaluation. These modules are also shown in Figure 1.

3.1. Data Pre-Processing

Usually, the datasets selected are 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 involves removing or filtering out unwanted information. This may involve unwanted information whose inclusion may increase the processing time or lead to abnormal results.

Thus, cleaning is performed to remove unwanted data. Tokenization is the process of dividing 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 allowing easier processing in the later stages. Stopword removal is applied to remove words with little to no significance for further processing.

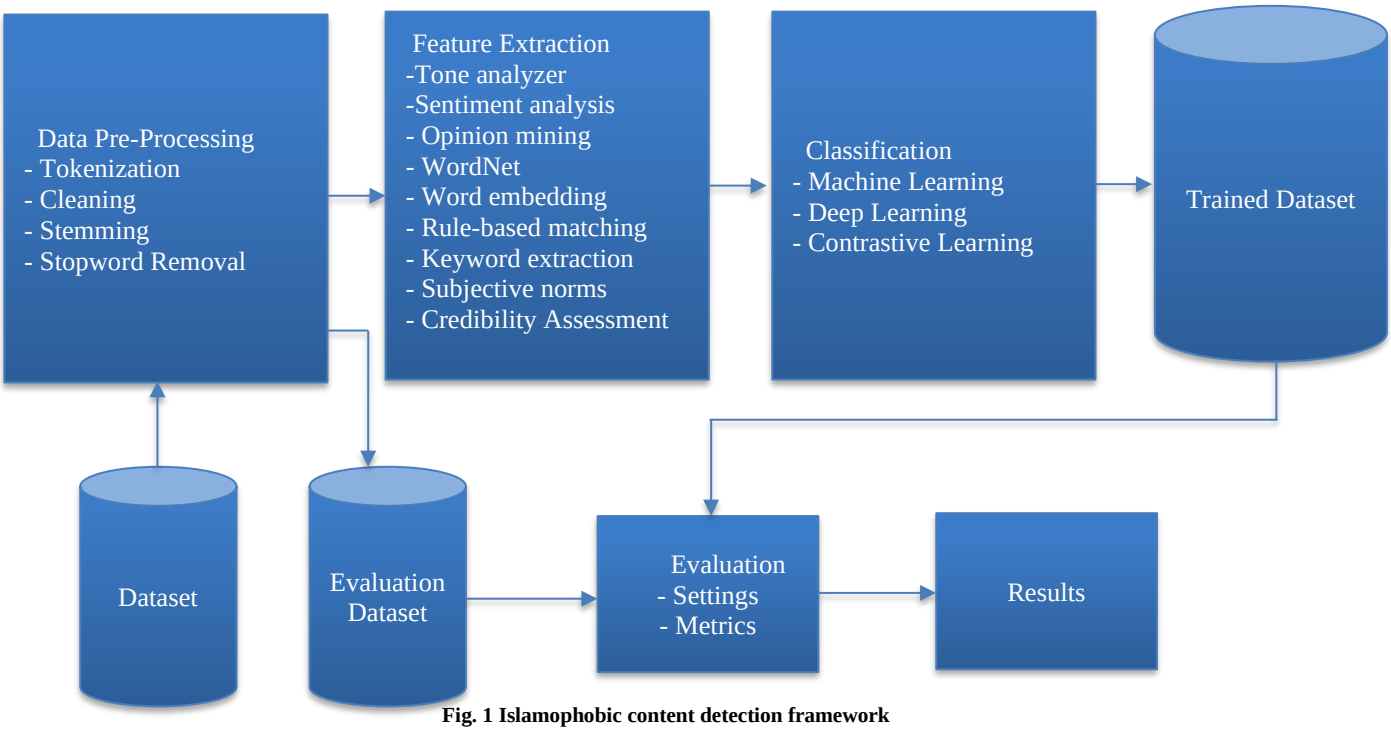

Fig. 1 Islamophobic content detection framework

All these steps and other pre-processing steps help bring the raw dataset to a state where it becomes more useful, and feature extraction can be performed easily and 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 distinguish between non-Islamophobic and Islamophobic content easily. In some cases, additional methods are applied to the pre-processed data to extract additional information that can act as features.

The research will highlight some of the features and methods that can be used in this module. A 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 classify Islamophobic content based on the intensity and category of the tone analyzed by the tone analyzer. Sentiment analysis is used to check the biases of the sentence. This can highlight whether the sentence is positive, negative, or neutral. In most cases, sentences with extreme sentiment values highlight warning flags towards 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 occur if the focus is only on a certain type of Islamophobic content. WordNet is useful for finding alternative words that help in understanding the semantics of the content better. Word embedding also adds details to words in a sentence that can help find associations between words and thus understand the sentence's semantics. Rule-based matching is useful to shortlist Islamophobic content quickly if they matches a certain rule. These can include sentences and combinations of words often used by people trying to instil 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 containing a particular keyword. These sentences can 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 on identifying Islamophobia, it is important to write methods or reasoners that can understand some of the norms used in Islam, which may be unrecognizable by a normal language parser. This may include expanding the vocabulary, 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 this research, but are not limited to these only.

Additional features may also be added to improve the accuracy of the system. These features and additional data extracts are forwarded to the classification module for further processing.

3.3. Classification

This module is responsible for using the data provided by features and methods and uses 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 is Islamophobic or not.

3.4. Evaluation

The last module is used for testing purposes to see the overall system's performance. Here, the results are compared against ground truths to see how the system performed. Several metrics are available for evaluation, 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, little to no research has been done on building systems that 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, highlighting the features used in such systems, and proposing a system 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 promote 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.

Author Contribution Statement

Abdulwahab A. Almazroi (AAA), Asad A. Shah (AAS), and Fathey Muhammad (F.M.) are the authors of the paper. AAA and AAS worked on the introduction and related work section. AAA also worked on the framework and conclusion. F.M. improved sentence structure, paper formatting, citations, and referencing.

References

[1] Siti Rohaidah Ahmad et al., "A Review of Feature Selection and Sentiment Analysis Technique in Issues of Propaganda," *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 11, pp. 240-245, 2019. [Google Scholar] [Publisher Link]

[2] Ahmed Al Marouf et al., "Recognizing Language and Emotional Tone from Music Lyrics Using IBM Watson Tone Analyzer," *IEEE International Conference on Electrical, Computer and Communication Technologies*, pp. 1-6, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[3] Fatimah Alkomah, Sanaz Salati, and Xiaogang Ma, "A New Hate Speech Detection System based on Textual and Psychological Features," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 8, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[4] Francisca Onaolapo Oladipo, Ogunsanya Funmilayo Blessing, and Ezendu Ariwa, "Terrorism Detection Model using Naive Bayes Classifier," *SSRG International Journal of Computer Science and Engineering*, vol. 7, no. 12, pp. 9-15, 2020. [CrossRef] [Publisher Link]

[5] V. Uma Maheswari, and R. Priya, "Analysis of Offensive Data over Multi-Source Social Media Environment Using Modified Random Forest Algorithm," *SSRG International Journal of Electronics and Communication Engineering*, vol. 10, no. 9, pp. 63-71, 2023. [CrossRef] [Publisher Link]

[6] Qasim Mehmood, Anum Kaleem, and Imran Siddiqi, "Islamophobic Hate Speech Detection from Electronic Media Using Deep Learning," *Mediterranean Conference on Pattern Recognition and Artificial Intelligence*, pp. 187-200, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[7] Nirali Arora et al., "Hinglish Profanity Filter and Hate Speech Detection," *International Journal of Computer Trends and Technology*, vol. 71, no. 2, pp. 1-7, 2023. [CrossRef] [Publisher Link]

[8] Rawan Abdullah Alraddadi, and Moulay Ibrahim El-Khalil Ghembaza, "Anti-Islamic Arabic Text Categorization Using Text Mining and Sentiment Analysis Techniques," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 8, pp. 776-785, 2021. [Google Scholar] [Publisher Link]

[9] Maha Jarallah 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, no. 5, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[10] Mohit Chandra et al., ""A Virus Has No Religion": Analyzing Islamophobia on Twitter during the COVID-19 Outbreak," *Proceedings of the 32nd ACM Conference on Hypertext and Social Media*, pp. 67-77, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[11] Rehab Duwairi, Amena Hayajneh, and Muhannad Quwaider, "A Deep Learning Framework for Automatic Detection of Hate Speech Embedded in Arabic Tweets," *Arabian Journal for Science and Engineering*, vol. 46, no. 4, pp. 4001-4014, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[12] Felipe González-Pizarro, and Savvas Zannettou, "Understanding and Detecting Hateful Content Using Contrastive Learning," *Proceedings of the Seventeenth International AAAI Conference on Web and Social Media*, vol. 17, pp. 257-268, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[13] Jeffrey W. Howard, "Free Speech and Hate Speech," *Annual Review of Political Science*, vol. 22, pp. 93-109, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[14] Secunder Kermani, Could a Student's Death Change Pakistan's Blasphemy Laws?, 2017. [Google Scholar] [Publisher Link]

[15] Heena Khan, and Joshua L. Phillips, "Language Agnostic Model: Detecting Islamophobic Content on Social Medi," *Proceedings of the 2021 ACM Southeast Conference*, pp. 229-233, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[16] Sean MacAvaney et al., "Hate Speech Detection: Challenges and Solutions," *PloS One*, vol. 14, no. 8, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[17] Toni M. Massaro, "Equality and Freedom of Expression: The Hate Speech Dilemma," *William and Mary Law Review*, vol. 32, 1990. [Google Scholar] [Publisher Link]

[18] Qasim Mehmood, Anum Kaleem, and Imran Siddiqi, "Islamophobic Hate Speech Detection from Electronic Media Using Deep Learning," *Mediterranean Conference on Pattern Recognition and Artificial Intelligence*, pp. 187-200, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[19] Raymond T. Mutanga, Nalindren Naicker, and Oludayo O. Olugbara, "Detecting Hate Speech on Twitter Network Using Ensemble Machine Learning," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 3, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[20] Fachrul Kurniawan, Badruddin, and Aji Prasetya Wibawa, "Identification of Islamophobia Sentiment Analysis on Twitter Using Text Mining Language Detection," *Journal of Positive School Psychology*, vol. 6, no. 5, pp. 8286-8294, 2022. [Google Scholar] [Publisher Link]

[21] Statistica, Number of Social Network Users Worldwide from 2017 to 2025, 2022. [Online]. Available: https://www.statista.com/statistics/278414/number-of-worldwide-social-network users/#:~:text=How%20many%20people%20use%20social,almost%204.41%20billion%20in%202025

[22] Luis Enrique Argota Vega et al., "Mineriaunam at Semeval-2019 Task 5: Detecting Hate Speech in Twitter using Multiple Features in a Combinatorial Framework," *Proceedings of the 13th International Workshop on Semantic Evaluation*, pp. 447-452, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[23] B. Vidgen, *Tweeting Islamophobia*, University of Oxford, 2019. [Google Scholar] [Publisher Link]

[24] Bertie Vidgen, and Taha Yasseri, "Detecting Weak and Strong Islamophobic Hate Speech on Social Media," *Journal of Information Technology and Politics*, vol. 17, no. 1, pp. 66-78, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[25] Wenjie Yin, and Arkaitz Zubiaga, "Towards Generalisable Hate Speech Detection: A Review on Obstacles and Solutions," *PeerJ Computer Science*, vol. 7, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[26] Heena Khan, and Joshua L. Phillips, "Language Agnostic Model: Detecting Islamophobic Content on Social Media," *Proceedings of the 2021 ACM Southeast Conference (ACM SE '21)*, pp. 229-233, 2021. [CrossRef] [Google Scholar] [Publisher Link]