Identification of Abusive Behavior Towards Religious Beliefs and Practices on Social Media Platforms

Tanvir Ahammad[1]
Department of Computer Science
and Engineering
Jagannath University, Bangladesh

Md. Khabir Uddin[2]
Department of Computer Science
and Engineering
Jagannath University, Bangladesh

Tamanna Yesmin[3]
Department of Computer Science
and Engineering
Uttara University, Bangladesh

Abdul Karim[4]
Department of Computer Science
and Engineering
Jagannath University, Bangladesh

Sajal Halder[5]
Department of Computer Science
RMIT University
Melbourne, Australia

Md. Mahmudul Hasan[6]
Department of Computer Science
and Engineering
Dhaka International University, Bangladesh

***Abstract*—The ubiquitous use of social media has enabled many people, including religious scholars and priests, to share their religious views. Unfortunately, exploiting people's religious beliefs and practices, some extremist groups intentionally or unintentionally spread religious hatred among different communities and thus hamper social stability. This paper aims to propose an abusive behavior detection approach to identify hatred, violence, harassment, and extremist expressions against people of any religious belief on social media. For this, first religious posts from social media users' activities are captured and then the abusive behaviors are identified through a number of sequential processing steps. In the experiment, Twitter has been chosen as an example of social media for collecting dataset of six major religions in English Twittersphere. In order to show the performance of the proposed approach, five classic classifiers on n-gram TF-IDF model have been used. Besides, Long Short-term Memory (LSTM) and Gated Recurrent Unit (GRU) classifiers on trained embedding and pre-trained GloVe word embedding models have been used. The experimental result showed 85% accuracy in terms of precision. However, to the best of our knowledge, this is the first work that will be able to distinguish between hateful and non-hateful contents in other application domains on social media in addition to religious context.**

***Keywords*—Social media; religious abuse detection; religious keywords; religious hatred; feature extraction; classifier**

I. Introduction

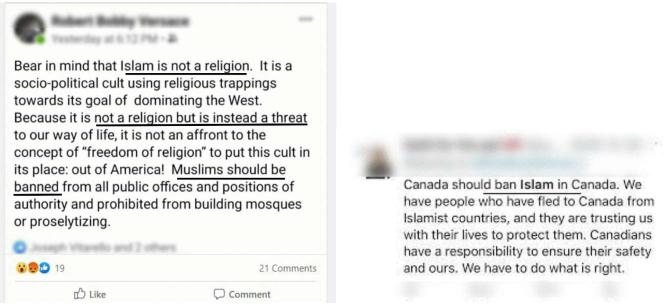
In the modern age of Information and Communication Technology (ICT), social media platforms seem to be an indispensable part of human lifestyle. Everyday, millions of individuals share their opinions, ideas, thoughts, feelings, experiences through social media services such as Twitter, Facebook, YouTube, Instagram and so on [1], [2]. As the people are not only sharing opinions but also connecting with new people, creating groups and making friendships, so social media has become an important source of information with variety of communities [3]. Moreover, different types of people, even many people who would be speechless previously, now use social media for different purposes in response to the fulfillment of the desires or goals of the mind [4]. Despite of having such a positive effect on freedom of expression on social media, we can not ignore its negative impacts. Some people either intentionally or unintentionally involved in abusive activities such as spreading false news and videos, trolling, cyberbullying and many more [5].

The people concerned with religion find the social media as useful platforms for sharing religious beliefs, rules and rituals to large audiences. They want to increase affiliations and trust on their religion through different types of activities on social media [6], [7]. During COVID-19 pandemic time, this scenario has increased significantly [8]. However, some people use hate speech against other religions to justify their religion. Furthermore, exploiting people's religious beliefs and practices, some extremist groups spread provocative and hateful content on social media. Consequently, the falsified and hateful information creates instability among the religious communities since social media news spread so fast by liking and sharing [9], [10]. Moreover, all these religious narrowness increase the likelihood of militant propaganda. As the people are very sensitive regarding religion, so it is essential to restrict the abusers who spread religiously offensive or hate speech on social media.

As religion refers to the ideological identity of a person [11], so expressing hatred by hurting a religion is like hurting a person's identity. In the literature, there is limited research on detecting religious abuse on social media. In our previous work [12], we addressed the concept of identification of religious abusers on Twitter. We only considered verifying whether abusive activities originated from spamming sources. Besides, there have been many works on social media about hate speech, such as identifying hate speech on Arabic social media [13], [14], which do not mention religious abuses. There is also another similar research in the literature demonstrating hate speech identification on vulnerable communities [15]. Detecting spamming activities in social media [16] is another type of research that is somewhat related to religious abuse detection. However, it should be admitted that not all abusers may use spamming techniques in spreading offensive speech regarding religion. There has been some more researches on identification of spreading propaganda on social media regarding different issues, like jihadist propaganda [17], [18], [19] and propaganda of COVID-19 deadly virus [20], [21], [22], [23] and then showing how the propaganda is analyzed [24]. Most of these are focused on Arabic social media context and targeted to single community. Therefore, there is a need for identifying abusive/offensive expressions from the posts and comments of every user who expresses religious views on social media.

Each social media has its own policy regarding different abusive activities, but it is still difficult to find out those activities in every user's posts. More specifically, identifying religiously abusive contents on social media are hard due to the noisy data structure. In other words, social media data consists of misspelled user-generated data (as users express freely anything they want without following any rules in writing), jargon, heterogeneity in data with the mixture of texts, urls, videos, emojis and so on [25].

In this paper, we focus on identifying abusive attitudes towards religion on social media. We refer to those activities religiously abusive that represents hatred, violence, harassment and extremist expressions against people of any religion or community, as depicted in Fig. 1 for an example. However, to find abusive contents, first we retrieved social media posts using a predefined set of religious keywords for six major religious beliefs and filtered them in order to remove unwanted information. Then, we labeled the filtered data with lexicon and rule based approaches. After that, we extracted features using traditional and advanced deep learning strategies from the texts after preprocessing. Finally, we fed the dataset into classifiers to identify abusive speech/content. In our experiments, we used Twitter as a social media platform for collecting the dataset. We compared the classical classification models, including Naïve Byes, Support Vector Machine (SVM), Random Forest, Logistic Regression, and MLP (Multilayer Perceptron) on TF-IDF feature extraction method. We also compared trained word embedding with pre-trained GloVe embedding in terms of LSTM and GRU models. The experimental result shows that we can obtain 85% overall accuracy in state-of-the-art performance in terms of precision.



Bear in mind that Islam is not a religion. It is a socio-political cult using religious trappings towards its goal of dominating the West. Because it is not a religion but is instead a threat to our way of life, it is not an affront to the concept of "freedom of religion" to put this cult in its place: out of America! Muslims should be banned from all public offices and positions of authority and prohibited from building mosques or proselytizing.

(a) Facebook post

Canada should ban **Islam** in Canada. We have people who have fled to Canada from Islamist countries, and they are trusting us with their lives to protect them. Canadians have a responsibility to ensure their safety and ours. We have to do what is right.

(b) Abusive Tweet

Fig. 1. An Example of Abusive Speech on Social Media Against Religious Communities.

To the best of our knowledge, this is first work that represents an approach to identify the abusive attitudes towards religion on social media platforms. That is, we can detect religiously abusive activities on any social media platform. In sum, our contributions are as follows:

- Our proposed approach examines the application of identifying religious abusive expressions on social media platforms.
- We identity abusive behaviour from social media users' posts for major religious beliefs in the world.
- We can use the proposed approach to detect hate/offensive speech on other application domains in addition to religious context.

The rest of the article is organized as follows: Section II represents recent studies in the literature related to this paper. In Section III, we have demonstrated and explained every module of the proposed approach. Then, Section IV represents the experimental procedures including data collection, feature extraction, building classification models and so on. Next, we have shown and discussed experimental results elaborately in Section V. Finally, we have concluded the work with future directions in Section VI.

II. Related Work

People use social media to express their feelings and find emotional gratifications for various reasons [26]. In religious point of view, the use of social media is making a significant contribution to the development of religious values. That is, the linkage between religion and social media has attracted many people over the years, including religious scholars and priests [27]. Despite of these advantages, abusive activities on religion can not be ignored. So, we need to pay attention to the negative impact of using social media in the context of religion. However, research on how to automatically detect offensive remarks from every user's posts and comments is limited in literature. Our previous work [12] highlighted the concept of detecting religious abusers on Twitter. It only showed that the spamming sources were identified as religiously offensive.

Detecting hate speech on Arabic twittersphere is very promising research tread in literature nowadays. Albadi et al. [28] published first publicly available annotated dataset and three lexicons with hate scores in order to detect religious hate speech from Arabic tweets. They classified extracted tweets as hate and not hate speech in terms of lexicon based, n-gram, GRU plus word embedding based, and GRU word embedding with handcrafted features including temporal, user and content features. The experimented result showed that feature based GRU model gave the best accuracy in terms of recall. However, their approach makes a pathway to find hate speech in Arabic tweets but it can not be applicable in many religious contexts such as Hinduism, Buddhism and so on. Besides, the annotated lexicons are unavailable for English contents.

Spreading offensive expressions in many contexts, such as Islamophobia, anti-Africa, and anti-Arab, on social media against people of different groups is not a good sign. Z. Mossie and J. H. Wang [15] were the first to find hatred against minor communities in Ethiopia from Amharic texts on Facebook. They annotated the dataset by the handcrafted method and then clustered the hate words using Word2Vec method. Although their approach successfully identified hate speech against vulnerable groups, but handcrafted features and annotation tend to the possibility of biasing on any cultural group. There is another research on communal hatred detection that was published by B. Vidgen and T. Yasseri [29]. They identified hate speech in their work on Muslims in terms of non-Islamophobic, weak and strong Islamophobic speech on Twitter. They considered the names of Muslims and Mosques in their analysis. It is not the case that only name of Muslim or Mosque spread Islamophobia, rather the hateful behavior, threats and online harassment against Muslims incite Islamophobia. So, these are also essential to consider.

G. Jain et al. [16] proposed a novel deep learning based approach (combination of CNN and LSTM) to identify spam detection in social media. They considered spamming behavior of user and short text messages in their proposed model. Similar type of research was conducted by N. Sun et al. [30] to find near teal-time spam on Twitter. They identified spams in terms of number of tweets issued by a user, number of retweeted, and fake accounts using traditional machine learning models. In articles [13] and [14] we found how deep learning models were used to identify hate speech in Arabic tweets that inspired us to do the work. Although these researches (on spam and hate speech detection) are somewhat related to our work, but it is acknowledged that all abusers may not use spamming tactics or use Arabic language in spreading offensive speech regarding religion. Therefore, we need to establish a model to analyse (lexically, syntactically, and/or semantically) the abusive/offensive behavior of each user when sharing religious views on social media platforms.

III. Methodology

This section represents our proposed approach that processes religious posts from social media users' activity and then identifies the abusers who spread hatred, violence, harassment, and extremist expressions among different religious communities. It consists of several segments with different functionalities, namely, Information Retrieval, Filtering, Labelling Class Attribute, Text Normalization, Text Vectorization, and Building and Evaluating Classifiers, represented in Fig. 2.

Information Retrieval (IR) is the method of capturing and extracting relevant users' activity from social media platforms. It follows three sequential processing steps. First of all, the social media from which we want to retrieve data needs to provide the required credentials (e.g., access token, consumer key, page id or post id) in compliance with all the terms and conditions. Then, it is necessary to identify the type of data (content language preferences, e.g., English, Japanese, or Arabic) we want to collect through religious search patterns or keywords. Finally, the collected data is transformed into tabular form by selecting different attributes.

Since not all users' activities or social media posts are related to religiously abusive, so only interested of these are considered for further processing. In other words, users' posts that are likely to have an impact on the religious community are filtered out. However, the filtering process follows three phase filtering schemes. In phase one, the information of the best possible impacted posts (or status) are considered, including keeping all unique posts, and posts that have length greater than a threshold. For example, the best impacted post length is between 70-100 in Twitter [31]. The second phase filters information based on the reputation of the user account, calculated as

$$R = \frac{Number\ of\ followers}{Number\ of\ friends + Number\ of\ followers}$$

If the R (stands for reputation) is small and close to zero, then it is probable to be an abusive account, as the abusers generally tend to get more followers [32]. Phase third of the filtering represents users' activity including the duration of account, date of posts, number of sharing (or retweeting) posts, number of posts issued, and number of likes or dislikes. Thus, the information of the higher activity is kept for future processing steps.

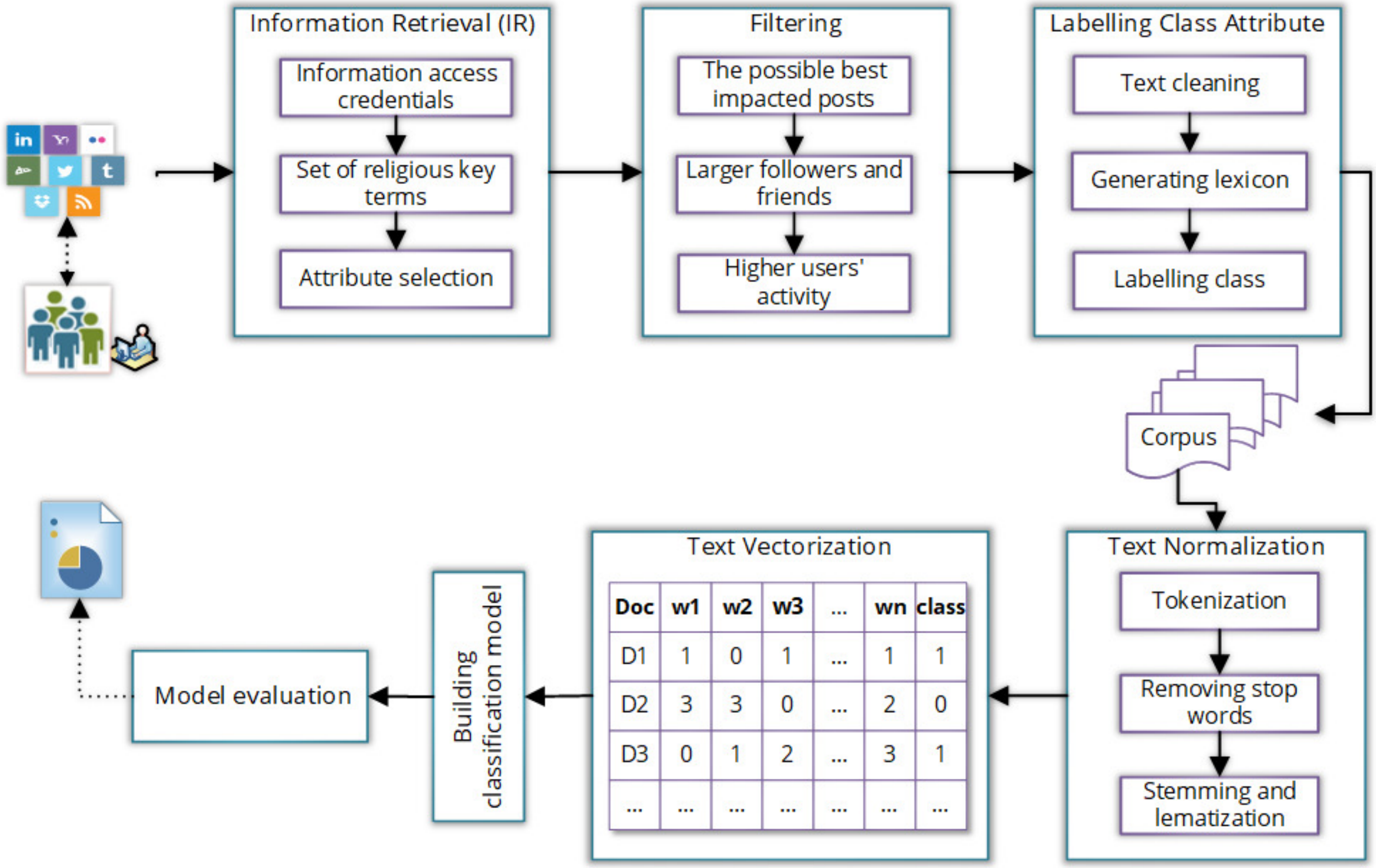
Labelling class attribute defines introducing the class attribute in the filtered dataset. As dataset consists of different types of attributes, so labelling class attribute is created on text data (i.e., religious posts or status of users). At first, text data is cleaned in order to remove unwanted texts or symbols, such as URLs, hashtags, special symbols, and punctuation. Then, generate religious lexicon containing words and phrases with their own polarity scores to be abusive, non-abusive, or neutral. After that, the cleaned texts are attributed as abusive or non-abusive class label using rule-based and generated lexicon-based approaches.

Before feeding the corpus into machine learning models, it is essential to transform the texts into a standard form or normalized form. At the beginning of the normalization, the texts are fragmented into smaller units called tokens, e.g., 'this is religiously abusive' →'this', 'is', 'religiously', 'abusive'. Then, remove all unwanted words such as 'about', 'above', 'across' from texts with predefined stop word list that is particularly applicable for social media text analytics. At the end of normalization, identical or near identical words (or tokens) are mapped into its base form, called steaming and lemmatization. That is, steaming and lemmatization help to convert the root form of inflected words. For example, the word "followers" and "followings" are transformed to "follow", its root form. Another example is representing of near similar words such as "keywords", "key-words" and "key words" to just "keywords".

After normalization, the textual data in the corpus is mapped into real valued vector form, called text vectorization or feature extraction from texts. That is, it is the process of making textual document into numerical vector form. In Fig. 2, text vectorization is illustrated by feature matrix which is defined as

- **w1, w2, w3,...,wn** represents features (n-grams or words);
- **Doc** represents textual documents (D1, D2, D3,...,Dn) in the corpus where each document indicates a social media post;
- **Class** represents class label attribute; and
- **Each row** represents a text data of a religious post containing feature values in it.

Building and evaluating classifiers is the final stage of the proposed approach where a classification model is built by feeding the feature matrix into it and then evaluate the model


Information Retrieval (IR)
Information access credentials
Set of religious key terms
Attribute selection
Filtering
The possible best impacted posts
Larger followers and friends
Higher users' activity
Labelling Class Attribute
Text cleaning
Generating lexicon
Labelling class
Corpus
Text Normalization
Tokenization
Removing stop words
Stemming and lematization
Text Vectorization
Building classification model
Model evaluation

Doc	w1	w2	w3	...	wn	class
D1	1	0	1	...	1	1
D2	3	3	0	...	2	0
D3	0	1	2	...	3	1
...	...	...	...	...	...	...

Fig. 2. Overall Architectural Diagram of the Proposed Approach.

with different performance metrics for showing how effectively the machine learning models classify religiously abusive or non-abusive contents.

IV. Experimental Setup

In this paper, we focused on six major religious beliefs in worldwide[1], including Christianity (31.2%), Islam (24.1%), Secular/Nonreligious/Agnostic/Atheist (16%), Hinduism (15.1%), Buddhism (6.9%), and Judaism (0.2%). We selected Twitter as a social media platform for collecting data in our experimental purposes. We used TF-IDF, trained Word Embedding, and pre-trained GloVe model for feature extraction. For classification, we used different classifiers in the state-of-the-art. However, the overall experimental design flow is demonstrated in Table I.

A. Data Collection

We retrieved total 9,787 publicly available users' Tweet related to English language and letters using Twitter's search API[2] method with Python Tweepy[3] library based on a predefined list of keywords as illustrated in Table II. As the Twitter API search approach returns object with a mix of root-level attributes, so we highlighted the most fundamental attributes for our experiment. We then removed duplicates and filtered the tweets based on attributes, such as friends, followers, like count, retweets, and total tweet issued. These are demonstrated in the Algorithm 1 and Algorithm 2, respectively. However, after filtering process, we obtained 4,903 tweets for subsequent processing steps.

B. Labelling Class to Tweet Dataset

Before labelling class in filtered tweets, we cleaned the tweets to eliminate hash tags, mentions, and urls because these don't make any significant impact in detecting religious abusive tweets. We then applied ruled based and lexicon based analysis on tweet texts. In order to this, we considered various techniques including position of words, surrounding of words, contexts, parts of speech, phrases, religious slangs (e.g., bible thumper), punctuations (e.g., good!!!), and degree of modifier (e.g., very, kind of). We used VADER (Valence Aware Dictionary and sEntiment Reasoner) Lexicon Tool[4] which is very effective in finding motifs from texts on social media platforms, especially for microblogging contexts. Using this tool, we calculated different polarity scores of words of a tweet text and then aggregated the scores. We then decided how much close the score for being abusive with a predefined threshold value. However, the Algorithm 3 demonstrates how

[1] https://www.pewforum.org/2015/04/02/religious-projections-2010-2050/
[2] https://developer.twitter.com/en/docs/twitter-api/v1/tweets/search/api-reference/get-search-tweets
[3] https://github.com/tweepy/tweepy
[4] https://github.com/cjhutto/vaderSentiment

TABLE I. SUMMARY OF THE EXPERIMENTAL PROCEDURES

Experiment steps	Description
Dataset	Contains tweets of six Religion communities/trusts/beliefs
Total tweets retrieved	9,787
Unique tweets	5,235
Filtered tweets	4,903
Total class labels	3 labels- • 0: non-abusive • 1: abusive • 2: neutral
Removed neutral class label samples	762
Total data samples for classification	4,141 samples for two class labels, i.e., • non-abusive (0): 2074 • abusive(1): 2067
Train and test split for classification	70% (2,898 samples) for training and 30% (1,243 samples) for testing
Feature extraction with classification models	• TF-IDF: Naive Bayes, SVM, Random Forest, Logistic Regression, MLP • Trained Word Embedding model: LSTM, GRU • Pre-trained GloVe model: LSTM, GRU
Performance evaluation metrics for classification models	• Accuracy • F1-score • Precision score • Recall score • Jaccard Similarity score • ROC-AUC score • Matthews Correlation Coefficient (MCC) • Zero-one Loss

TABLE II. RELIGIOUS BELIEFS WITH RELATED KEYWORDS USED TO COLLECT TWEETS

Religious Beliefs	Keywords
Christianity	Christian, Roman Catholic, Christianity, Apocalypticism, Catholic Church, Baptism, Bible, Bishop
Islam	Quran, Islam, Muslims, Islamic State, Kurdish, Shia, Sunni, Jihad, Wahhabi, Islamphobia
Atheist	Atheist, Atheists, Atheism
Hinduism	Hindu, Bhagavad-Gita, Brahman, Mahabharata, Ma Kali, Ramayana, Durga, Saraswati, Jai Hanuman
Buddhism	Buddhism, Gautama Buddha, Bodhisattva, Buddha, Dalai Lama, Mahayana, Nirvana
Judaism	Judaism, Jews, Jew, Berit

```
Algorithm 1: Tweet Filtering ($Tweets_{keywords}$)
  Data: $Tweets_{keywords}$: All tweets retrieved from
        religious keywords search approach
  Result: Filtered tweets after satisfying different
          conditions
  /* Initially set filtered tweet list to
     null                                         */
  $filtered_{Tweets} \leftarrow \phi$
  $unique_{tweets} \leftarrow UniqueTweets(Tweet_{keywords})$
  for $\forall\ tweet \in unique_{tweets}$ do
      if $tweet.followers \geq 100$ then
          $filtered_{Tweets} = tweet$
      else if $tweet.friends \geq 100$ then
          $filtered_{Tweets} = tweet$
      else if $tweet.retweet \geq 50$ then
          $filtered_{Tweets} = tweet$
      else if $tweet.tweet_{like\_count} \geq 1$ then
          $filtered_{Tweets} = tweet$
      else if $tweet.total_{tweet\_issued} \geq 50$ then
          $filtered_{Tweets} = tweet$
      else
          $filtered_{Tweets} \notin unique_{tweets}$
  return $filtered_{Tweets}$
```

```
Algorithm 2: UniqueTweets ($Tweets$)
  Data: $Tweets$: Set of all tweets retrieved from
        Tweet API search
  Result: Unique tweets after removing duplicates
  /* Initializing first tweet from Tweets
     */
  $uniqueTweets \leftarrow tweet_1 \in Tweets$
  for $\forall\ tweet \in Tweets$ do
      for $\forall\ temp \in uniqueTweets$ do
          if $tweet \neq temp$ then
              $uniqueTweets = tweet$
  return $uniqueTweets$
```

we can achieve labelling class attributes in the dataset. After including class attributes in filtered dataset, we obtained our corpus for classification.

C. Feature Extraction

In feature extraction, we transformed the texts (from corpus) into feature vector. Before doing that, we preprocessed or normalized the texts in a few consecutive ways. That is, we first removed special characters, numbers and punctuations; then expanded the contractions, lowering case, tokenization and removing stop words with our predefined stop word list; and finally mapped the words to their root form with lemmatization. For feature extraction, we used three different models, i.e., n-gram TF-IDF, trained deep learning based word embedding model, and pre-trained GloVe model. The intention of choosing these three methods is to verify whether the accuracy of the classification improves with feature extraction. However, we obtained different feature matrices for each model and then fed them in the classifiers.

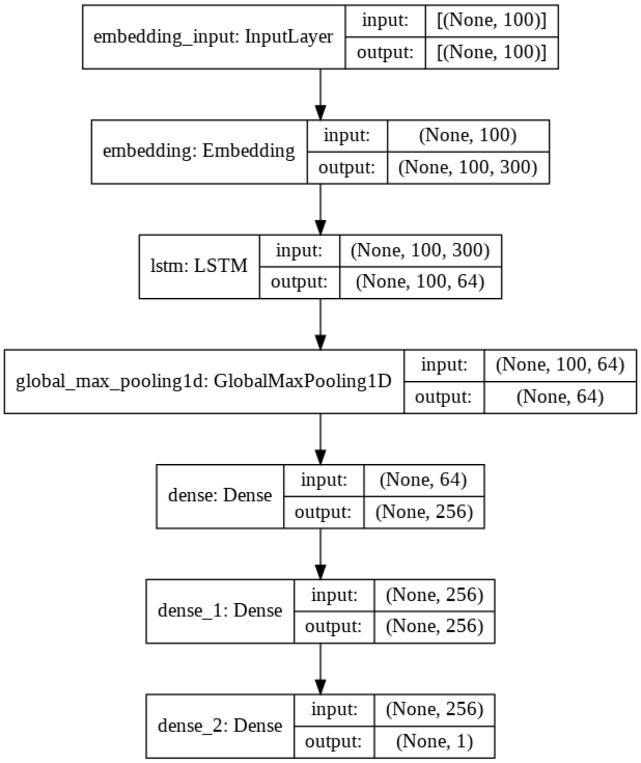


Fig. 3. Visualization of LSTM Model Configuration.

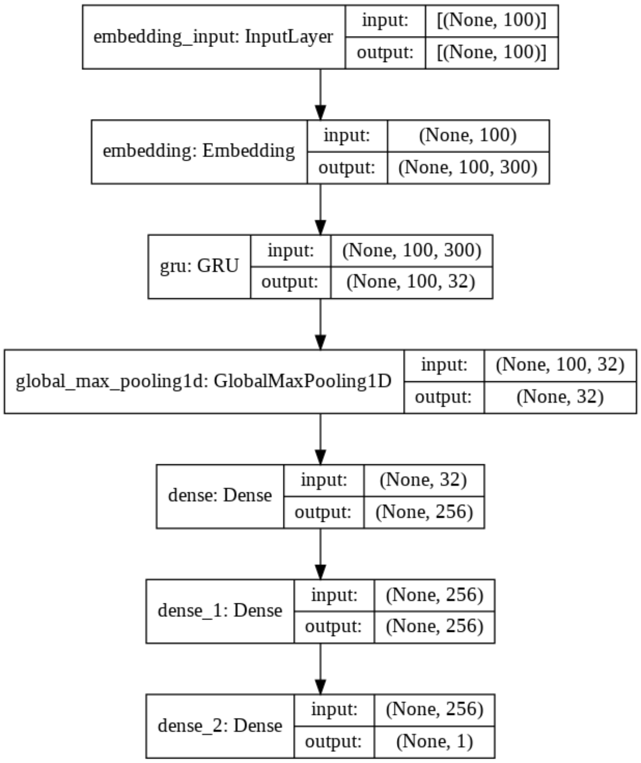


Fig. 4. Visualization of GRU Model Configuration.

```
Algorithm 3: Labelling Classes($Tweets_{Filtered}$)
  Data: $Tweets_{Filtered}$: Set of all tweet texts after
        applying tweet filtering process
  Result: List of class labels
          $(0 : non - abusive, 1 : abusive, 2 : neutral)$
          associated with each tweet
  /* Initialize class label list to null
     */
  $label_{class} \leftarrow \phi$
  for $\forall\ texts \in Tweets_{Filtered}$ do
      $scores_{aggregated} = VADER\_Lexicon(texts)$;
      if $scores_{aggregated} \geq 0.05$ then
          $label_{class}$ = 0;
      else if $scores_{aggregated} \leq 0.05$ then
          $label_{class}$ = 1;
      else
          $label_{class}$ = 2;
  return $label_{class}$
```

D. Building Classifiers

As we focused on detecting religiously abusive contents from texts, so we skipped neutral class labels and considered only abusive and non-abusive classes only. That is, we considered 4,141 (out of 4,903) data samples containing abusive and non-abusive attributes for classification purposes. However, we split the dataset into train and test subsets, where 70% (2,898 samples) were for training and 30% (1,243 samples) for testing. For classification, we used Naïve Bayes, SVM, Random Forest, Logistic Regression, and MLP classifiers on TF-IDF feature matrix. On the other hand, LSTM, and GRU classification models were used on both trained and pre-trained GloVe embedding models because these classifiers are more efficient than the traditional machine learning based classifiers. We have trained all the classifiers more than 100 times with different parameters and then selected the best configuration for the final classification model building. MLP, LSTM, and GRU are neural network based classifiers that may not perform well with predefined parameters for all datasets, rather, depending upon the use case and problem statement, so we emphasized on choosing training parameters that show best performance on test data. The parameters of MLP include: input layer size=total features (i.e., more than 10,000), 3 hidden layers with neurons (125, 125, 125), optimization='adam', learning rate=0.0001, maximum iteration is between 2000-5000, and 1 output layer for deciding whether abusive or non-abusive; on the other hand, the best configuration of LSTM, and GRU networks are shown in Fig. 3 and Fig. 4, respectively. However, we then evaluated the classifiers with various performance metrics (see in Table I).

V. Result Analysis and Discussions

As we focused on detecting abusive activities among major religious beliefs of worldwide in social media, so we haven't mentioned how much abusive activities are identified of a particular religion in our experiments. In this section, we have discussed the obtained results in different perspectives.

A. Exploratory Data Analysis

In section IV-A, we demonstrated how we collected tweets with a mix of fundamental attributes using different religious keywords. Among them, the hashtag attribute gives an overview of how religiously relevant datasets we have been able to collect. While retrieving the dataset, we found more than 100 unique hashtags, the most commonly used hashtags are shown in the Fig. 5. It shows that the most occurring hashtags are somehow religiously related, so it can be said that the more occurring religious hashtags, the more likely users have involved in religious tweets.

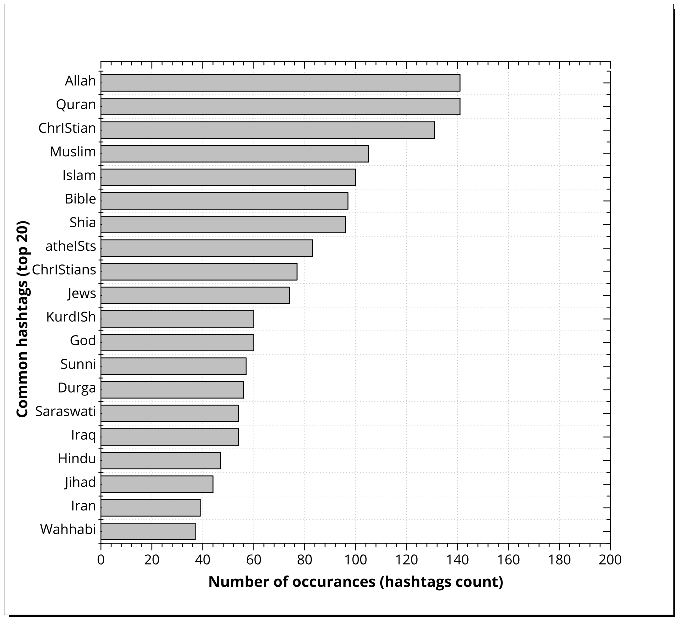


Fig. 5. Most Frequently Occurring Hashtags in the Collected Tweets.

Users in Twitter share their opinions, ideas and thoughts on diverse religious issues with a large number of audiences. So, taking these opportunities, some users may involve in the harassment of someone, sharing offensive posts or contents, spreading hate speech, and then encourage others to do so. In this paper, we collected many users' tweet (4,903 after filtering) containing both abusive and normal or neutral attitudes. To illustrate the behaviour of the users' tweet, we have shown them using the most occurring unigram, bigram and trigram frequencies Fig. 6, Fig. 7 and Fig. 8, respectively. Fig. 6 shows twenty most frequently occurring unigrams (single words) with their respective frequency counts out of 66,730 unique unigrams. Whereas Fig. 7 shows ten most occurring bigrams (2-adjacent words from a sequence of tokens) with the number of times, they appear sequentially in users' tweet out of 33,350 unique bigrams. On the other hand, Fig. 8 depicts ten most commonly occurring trigrams (3 consecutive words from a sequence of tokens) out of 22,243 unique trigrams on different religions.

We also analysed the number of religious abusers based on the location of tweet users. We found more than 1,500 specified locations that the user provided in their accounts profile. However, Fig. 9 shows ten locations where the highest percentage of abusive tweets were found (as there were 2,067 abusers detected out of 4,141 tweets). The locations indicated in the figure are categorized by country to visualize the percentage of abusers.

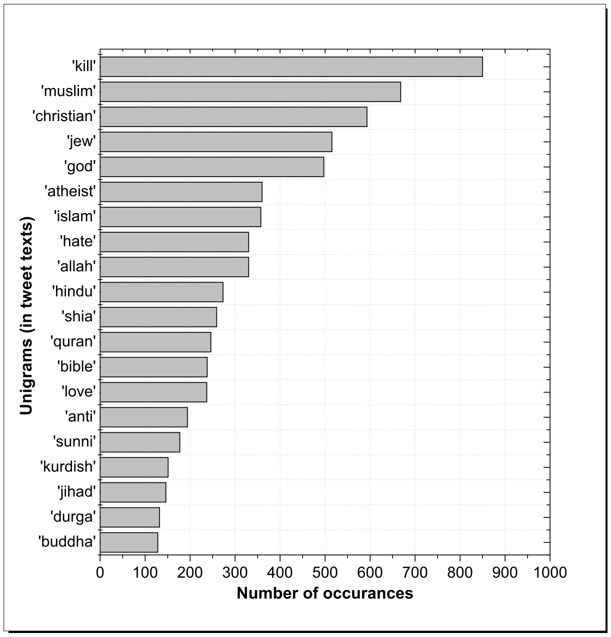


Fig. 6. Most Commonly Occurring Unigrams in users' Tweet on Different Religious Beliefs.

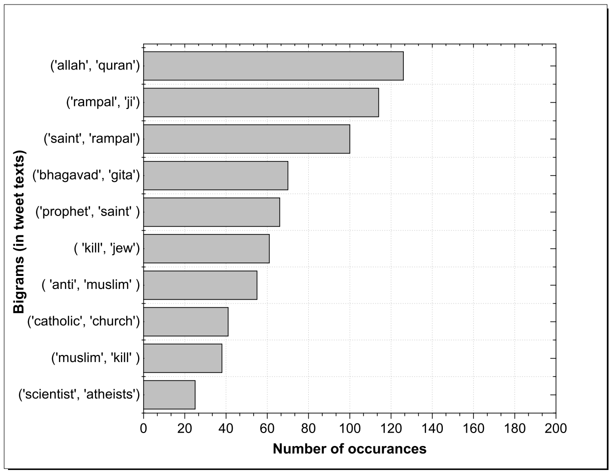


Fig. 7. Ten Most Commonly Occurring Bigrams in users' Tweet on Different Religious Beliefs.

As the text contents on Twitter is limited up to 280 characters, so many users try to attract more audiences to read or engage through posting of different lengths of tweets on it. In our analysis, we considered three different tweet lengths (in total of 4,141 tweets), i.e., '70-140' character limit, '140-280', and less than 70 character limit. These are shown in Fig. 10, where 72.52% of tweets are found within '140-280' character limit, 22% in between '70-140', and then 5.48% in less than 70 character limit. Moreover, the percentage of abusers and normal users in '140-280' length are also higher than that of other character limits. Fig. 10 also shows that the abusive tweets in '140-280'character length are higher than that of other limits, whereas normal or non-abusive tweets in '70-

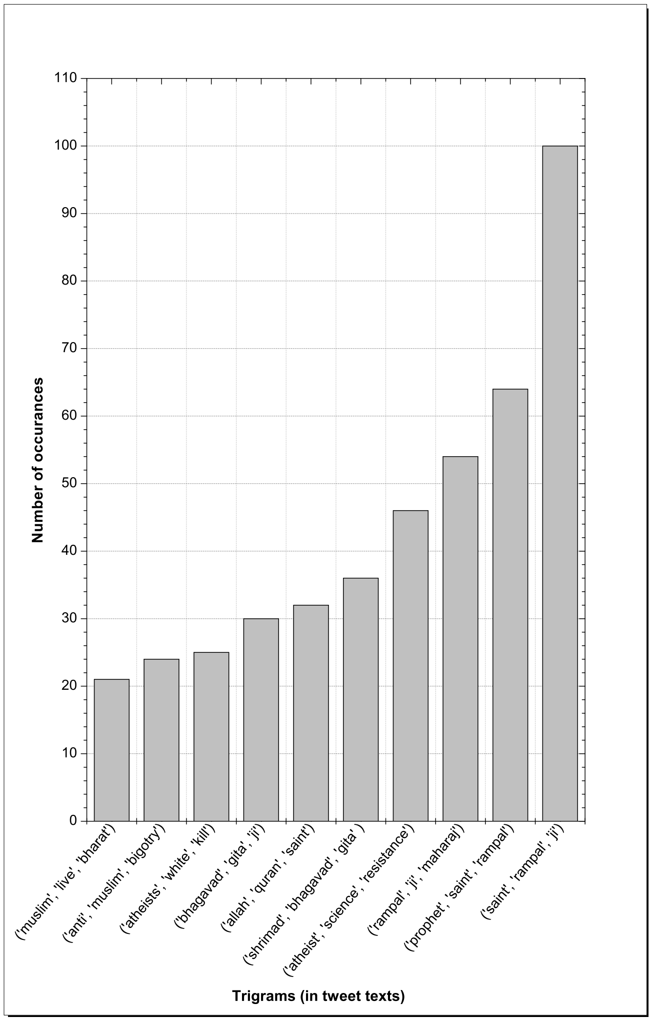


Fig. 8. Ten Most Commonly Occurring Trigrams in users’ Tweet on Different Religious Beliefs.

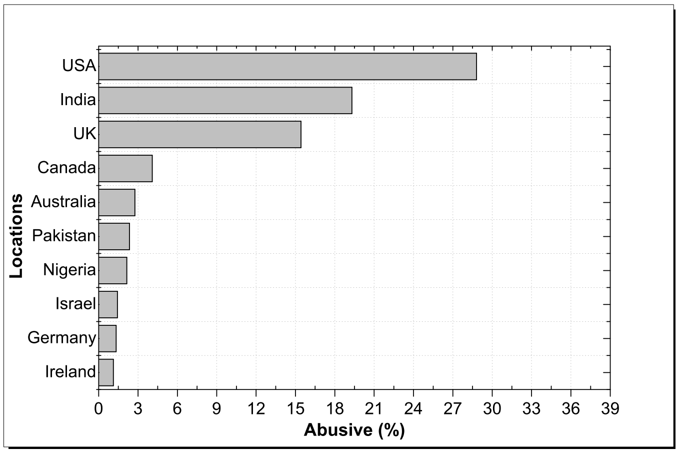


Fig. 9. The Highest Percentage of Abusers Detected in Different Countries.

140’ and ‘¡70’ character limit are higher than that of abusive tweets. So, in sum, the most common length of tweets in our dataset are between 140 and 280 characters, which indicates that the users posted long tweets to express their thoughts on religion.

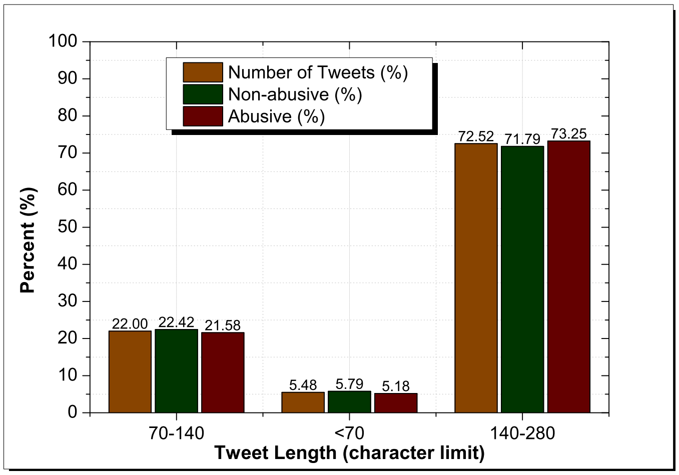


Fig. 10. Percentage users’ Activities in Different Character Limits of Tweets.

B. Evaluating Classifiers

We evaluated the classification models using new unseen testing dataset. As we trained the models more than 100 times with different parameters and selected the best configuration for the final model building, so we evaluated the classifiers with the best performance on testing data (i.e., 1,253 samples with abusive 618, and non-abusive 625). Table III shows a comparative summary of the performance of the classifiers on TF-IDF feature matrix. We can see that SVM shows the best performance compared to other classification models in terms of different classification metrics (except in Jaccard Similarity score). So, the overall accuracy and loss of SVM are 83% and 17%, respectively, which is indeed promising. This means that we have been able to classify 83% correctly of abusive and non-abusive tweets. Table III also indicates that the MLP model shows quite similar performance to SVM except for (0.1-0.4)% marginal difference. However, to illustrate the performance of the models visually, the confusion matrix of each model is presented in Fig. 11. We can see that the number of false positives is higher than that of false negatives in Naïve Bayes and MLP classifiers. On the other hand, false negatives are higher than the false positives in SVM, Random Forest, and Logistic Regression.

As deep learning based feature extraction models are best suited for large features, including millions of parameters [33], so we focused on improving the classification accuracy in terms of trained and pre-trained deep learning models for more than 70,000 n-gram features in our dataset. In Table IV, the performance of LSTM and GRU classifiers are shown on trained word embedding feature extraction model. We can see that the LSTM performs better than that of GRU classifier in all classification metrics. So, we have achieved 84% overall accuracy and 16% loss on trained deep learning word embedding using LSTM. The confusion matrix of the two classifiers is depicted in Fig. 12. It shows that the number of false positives is higher than the false negatives in both LSTM and GRU models.

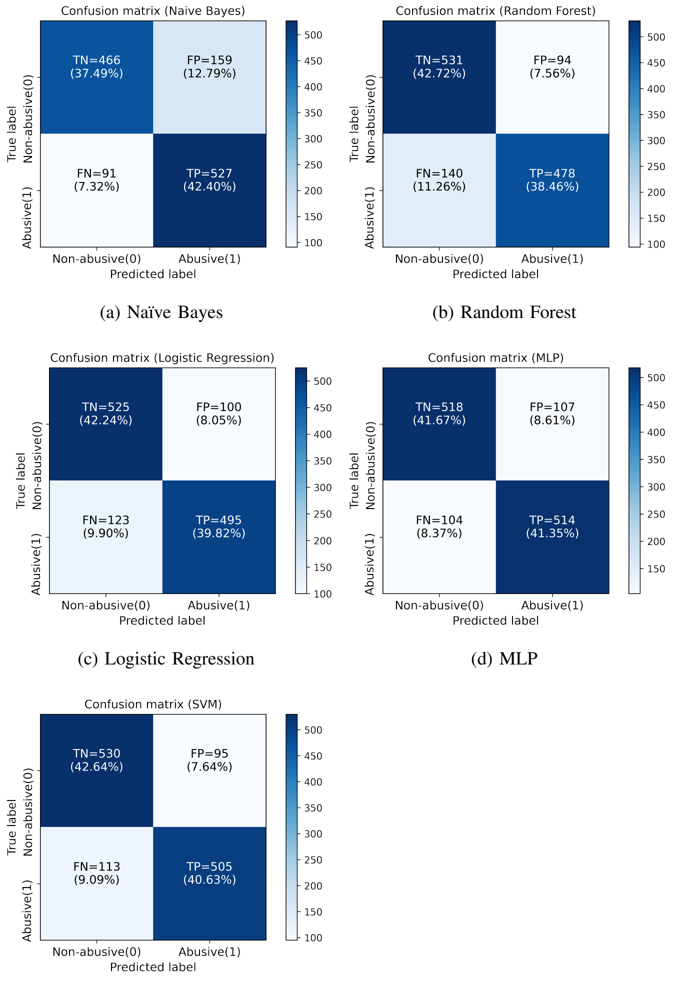
GloVe embedding is learned in one task and used to solve another identical task. We used this pre-trained embedding

TABLE III. Classification Model Evaluation based on TF-IDF Feature Extraction Model

Classifiers	Accuracy	F1-score	Precision	Recall	Jaccard Similarity score	ROC-AUC score	MCC score	Zero-one loss
Naïve Bayes	0.798	0.798	0.802	0.799	0.678	0.799	0.601	0.201
SVM	**0.832**	**0.832**	**0.832**	**0.832**	0.708	**0.832**	**0.665**	**0.167**
Random Forest	0.811	0.811	0.813	0.811	0.671	0.811	0.625	0.188
Logistic Regression	0.820	0.820	0.821	0.820	0.689	0.820	0.641	0.179
MLP	0.830	0.830	0.830	0.830	**0.709**	0.830	0.661	0.170

TABLE IV. Classification Model Evaluation based on Trained Deep Learning Word Embedding Model

Classifiers	Accuracy	F1-score	Precision	Recall	Jaccard Similarity score	ROC-AUC score	MCC score	Zero-one loss
LSTM	**0.844**	**0.844**	**0.845**	**0.844**	**0.735**	**0.844**	**0.689**	**0.156**
GRU	0.837	0.837	0.838	0.837	0.726	0.837	0.675	0.163



(a) Naïve Bayes

(b) Random Forest

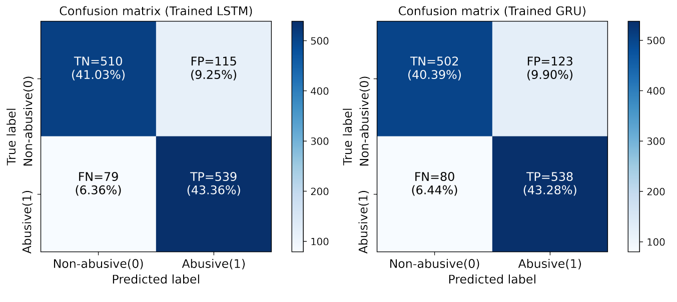
(c) Logistic Regression

(d) MLP

(e) SVM

Fig. 11. Confusion Matrix of Traditional Classifiers on n-gram TF-IDF Feature Method.

model to create the embedding matrix on training dataset and then fed it into LSTM and GRU models for classification.

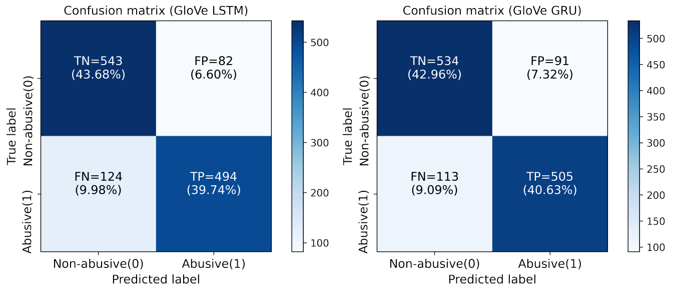


(a) LSTM

(b) GRU

Fig. 12. Confusion Matrix of LSTM and GRU Classifiers on Trained Word Embedding Feature Extraction.

Table V shows the performance of the classifiers on test data. This time GRU model shows better performance than that of LSTM in each classification metric. We can see that the accuracy is obtained 84% whereas the loss is 16%. The confusion matrix of LSTM and GRU model on GloVe embedding is shown in Fig. 13. We can see that false negatives are higher than the false positives in both GRU and LSTM classifiers.



(a) LSTM

(b) GRU

Fig. 13. Confusion Matrix of LSTM and GRU Classifiers on Pre-trained GloVe Embeddings Model.

However, we have seen that SVM shows the best accuracy on TF-IDF, LSTM on trained embedding and GRU on GloVe

TABLE V. Classification Model Evaluation based on Pre-trained GloVe Model. The Bold Numbers Represent the Best Results

Classifiers	Accuracy	F1-score	Precision	Recall	Jaccard Similarity score	ROC-AUC score	MCC score	Zero-one loss
LSTM	0.834	0.834	0.836	0.834	0.706	0.834	0.670	0.166
GRU	**0.836**	**0.836**	**0.836**	**0.836**	**0.712**	**0.836**	**0.672**	**0.164**

TABLE VI. Classifiers Performance Comparison based on TF-IDF, Trained Word Embedding and Pre-trained GloVe Model. The Bold Numbers Represent the best Results.

Feature extraction model	Accuracy	F1-score	Precision	Recall	Jaccard Similarity score	ROC-AUC score	MCC score	Zero-one loss
TF-IDF	0.832	0.832	0.832	0.832	0.709	0.832	0.665	0.167
Trained word embedding	**0.844**	**0.844**	**0.845**	**0.844**	**0.735**	**0.844**	0.689	**0.156**
Pre-rained GloVe emending	0.836	0.836	0.836	0.836	0.712	0.836	**0.672**	0.164

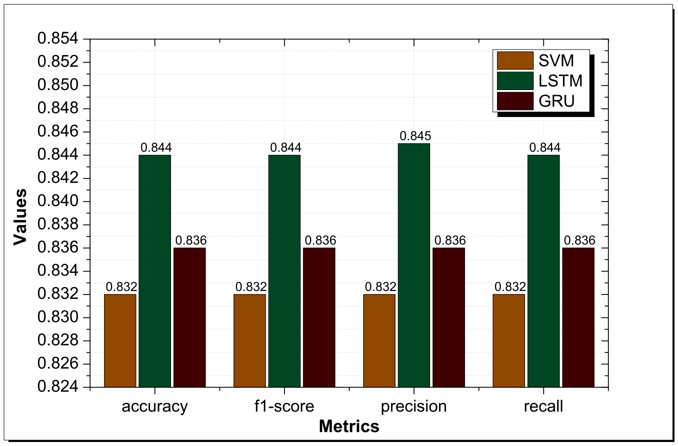


Fig. 14. Classification Performance Evaluation of SVM, LSTM, and GRU Models.

embedding. Now we compare each feature extraction method in terms of classification metrics, as presented in Table VI. The classifier on trained word embedding performs better than the other classification models. That is, LSTM model shows the best classification accuracy (84%) compared to others. The overall accuracy of the classifiers on three feature extraction methods is depicted in Fig. 14. So, if we consider precision, then 85% overall accuracy will be achieved.

VI. Conclusions

This paper focuses on identifying religiously abusive users on social media. It is the first research devoted to analyse and classify user activities regarding diverse religious beliefs and practices on any social media platform. The main contribution of this research is to establish an approach for detecting religiously abusive activities from users' social media posts. For conducting the experiment, Twitter has been selected as a social media data source. There were many users' activities (approx. 10,000) collected using a set of predefined religious keywords in English Twittersphere. Then the tweets that were redundant in nature and contained less user involvement or attraction were filtered. We then labelled the dataset using rule based and lexicon based approaches. The labelled dataset (tweet texts) was fed into classifiers after extracting features using three different methods. The performance of the classifiers were evaluated with various classification metrics on test data. The obtained results indicate that SVM model showed 0.832 (83%) accuracy compared to others on TF-IDF, whereas LSTM gave 0.844 (84%) accuracy on trained embedding, and GRU showed 0.836 (≈84%) accuracy on GloVe model. However, the LSTM model on trained word embedding showed 85% precision in state-of-the-art performance. Finally, we will be able to use the proposed approach for identification of any hatred/offensive speeches on any social media platform besides religious context.

Although this paper shows the identification of abusive behaviors on different religious beliefs with good accuracy, there are few constraints that were needed to be addressed. Firstly, more train and test data were desirable to take into consideration. Secondly, imbalance class distributions should be used to examine the performance of classifiers (since we used almost balanced class distributions) in experiment. However, in future, we will consider these limitations and explore on user activities of different languages. We will also identify whether the abusive contents are spread by human or social robots. In addition, we want to find the semantic similarity of hate/ non-hate speech on any religion with defined abusive key-terms in users' posts.

Acknowledgment

This work was partially supported by Research, Jagannath University, Dhaka, Bangladesh.

References

[1] M. A. Tocoglu, O. Ozturkmenoglu, and A. Alpkocak, "Emotion analysis from turkish tweets using deep neural networks," *IEEE Access*, vol. 7, pp. 183 061–183 069, 2019, doi: 10.1109/ACCESS.2019.2960113.

[2] Y. Qi, F. Aleksandr, and F. Andrey, "I know where you are coming from: On the impact of social media sources on ai model performance (student abstract)," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 10, 2020, pp. 13 971–13 972, doi: 10.1609/aaai.v34i10.7258.

[3] K. S. Sweet, J. K. LeBlanc, L. M. Stough, and N. W. Sweany, "Community building and knowledge sharing by individuals with disabilities using social media," *Journal of computer assisted learning*, vol. 36, no. 1, pp. 1–11, 2020, doi: 10.1111/jcal.12377.

[4] D. L. Hoffman and T. Novak, "Why do people use social media? empirical findings and a new theoretical framework for social media goal pursuit," *Empirical Findings and a New Theoretical Framework for Social Media Goal Pursuit (January 17, 2012)*, 2012, doi: 10.2139/ssrn.1989586.

[5] C. V. Baccarella, T. F. Wagner, J. H. Kietzmann, and I. P. McCarthy, "Social media? it's serious! understanding the dark side of social media," *European Management Journal*, vol. 36, no. 4, pp. 431–438, 2018, doi: 10.1016/j.emj.2018.07.002.

[6] P. J. Brubaker and M. M. Haigh, "The religious facebook experience: Uses and gratifications of faith-based content," *Social Media+ Society*, vol. 3, no. 2, p. 2056305117703723, 2017, doi: 10.1177/2056305117703723.

[7] A. Baraybar-Fernández, S. Arrufat-Martín, and R. Rubira-García, "Religion and social media: Communication strategies by the spanish episcopal conference," *Religions*, vol. 11, no. 5, p. 239, 2020, doi: 10.3390/rel11050239.

[8] F. Lendriyono, "Public's perception on social media towards new normal during covid-19 pandemic in indonesia: Content analysis on religious social media accounts," in *IOP Conference Series: Earth and Environmental Science*, vol. 717, no. 1. IOP Publishing, 2021, p. 012039, doi:10.1088/1755-1315/717/1/012039.

[9] T. Buchanan, "Why do people spread false information online? the effects of message and viewer characteristics on self-reported likelihood of sharing social media disinformation," *Plos one*, vol. 15, no. 10, p. e0239666, 2020, doi: 10.1371/journal.pone.0239666.

[10] S. Vosoughi, D. Roy, and S. Aral, "The spread of true and false news online," *Science*, vol. 359, no. 6380, pp. 1146–1151, 2018, doi: 10.1126/science.aap9559.

[11] L. Peek, "Becoming muslim: The development of a religious identity," *Sociology of religion*, vol. 66, no. 3, pp. 215–242, 2005, doi: 10.2307/4153097.

[12] T. Ahammad, M. K. Uddin, A. Karim, and S. Halder, "A framework for detecting and tracking religious abuse in social media," in *2019 International Conference on Computing, Power and Communication Technologies (GUCON)*. IEEE, 2019, pp. 206–211.

[13] S. Alsafari, S. Sadaoui, and M. Mouhoub, "Hate and offensive speech detection on arabic social media," *Online Social Networks and Media*, vol. 19, p. 100096, 2020, doi: 10.1016/j.osnem.2020.100096.

[14] R. Alshalan and H. Al-Khalifa, "A deep learning approach for automatic hate speech detection in the saudi twittersphere," *Applied Sciences*, vol. 10, no. 23, p. 8614, 2020, doi: 10.3390/app10238614.

[15] Z. Mossie and J.-H. Wang, "Vulnerable community identification using hate speech detection on social media," *Information Processing & Management*, vol. 57, no. 3, p. 102087, 2020, doi:10.1016/j.ipm.2019.102087.

[16] G. Jain, M. Sharma, and B. Agarwal, "Spam detection in social media using convolutional and long short term memory neural network," *Annals of Mathematics and Artificial Intelligence*, vol. 85, no. 1, pp. 21–44, 2019, doi:10.1007/s10472-018-9612-z.

[17] A. Badawy and E. Ferrara, "The rise of jihadist propaganda on social networks," *Journal of Computational Social Science*, vol. 1, no. 2, pp. 453–470, 2018, doi: 10.1007/s42001-018-0015-z.

[18] A. Al-Rawi and J. Groshek, "Jihadist propaganda on social media: An examination of isis related content on twitter," *International Journal of Cyber Warfare and Terrorism (IJCWT)*, vol. 8, no. 4, pp. 1–15, 2018, doi: 10.4018/IJCWT.2018100101.

[19] L. Wakeford and L. Smith, "Islamic state's propaganda and social media: Dissemination, support, and resilience," in *ISIS propaganda: A full-spectrum extremist message*. Oxford University Press, 2020, pp. 155–187, doi:10.1093/oso/9780190932459.003.0006.

[20] A. M. U. D. Khanday, Q. R. Khan, and S. T. Rabani, "Identifying propaganda from online social networks during covid-19 using machine learning techniques," *International Journal of Information Technology*, pp. 1–8, 2020, doi: 10.1007/s41870-020-00550-5.

[21] D. A. Broniatowski, D. Kerchner, F. Farooq, X. Huang, A. M. Jamison, M. Dredze, and S. C. Quinn, "The covid-19 social media infodemic reflects uncertainty and state-sponsored propaganda," *arXiv preprint arXiv:2007.09682*, 2020.

[22] O. D. Apuke and B. Omar, "Fake news and covid-19: modelling the predictors of fake news sharing among social media users," *Telematics and Informatics*, vol. 56, p. 101475, 2021, doi: 10.1016/j.tele.2020.101475.

[23] H. Nguyen and A. Nguyen, "Covid-19 misinformation and the social (media) amplification of risk: A vietnamese perspective," *Media and Communication*, vol. 8, no. 2, pp. 444–447, 2020, doi: 10.17645/mac.v8i2.3227.

[24] D. D. Chaudhari and A. V. Pawar, "Propaganda analysis in social media: a bibliometric review," *Information Discovery and Delivery*, 2021, doi: 10.1108/IDD-06-2020-0065.

[25] Y. K. Dwivedi, G. Kelly, M. Janssen, N. P. Rana, E. L. Slade, and M. Clement, "Social media: The good, the bad, and the ugly," *Information Systems Frontiers*, vol. 20, no. 3, pp. 419–423, 2018, doi: 10.1007/s10796-018-9848-5.

[26] A. Whiting and D. Williams, "Why people use social media: a uses and gratifications approach," *Qualitative Market Research: An International Journal*, 2013, doi: 10.1108/QMR-06-2013-0041.

[27] M. S. Kgatle, "Social media and religion: Missiological perspective on the link between facebook and the emergence of prophetic churches in southern africa," *Verbum et Ecclesia*, vol. 39, no. 1, pp. 1–6, 2018, doi: 10.4102/ve.v39i1.1848.

[28] N. Albadi, M. Kurdi, and S. Mishra, "Investigating the effect of combining gru neural networks with handcrafted features for religious hatred detection on arabic twitter space," *Social Network Analysis and Mining*, vol. 9, no. 1, pp. 1–19, 2019, doi: 10.1007/s13278-019-0587-5.

[29] 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, doi: 10.1080/19331681.2019.1702607.

[30] N. Sun, G. Lin, J. Qiu, and P. Rimba, "Near real-time twitter spam detection with machine learning techniques," *International Journal of Computers and Applications*, pp. 1–11, 2020, doi: 10.1080/1206212X.2020.1751387.

[31] A. B. Boot, E. T. K. Sang, K. Dijkstra, and R. A. Zwaan, "How character limit affects language usage in tweets," *Palgrave Communications*, vol. 5, no. 1, pp. 1–13, 2019, doi: 10.1057/s41599-019-0280-3.

[32] W. Herzallah, H. Faris, and O. Adwan, "Feature engineering for detecting spammers on twitter: Modelling and analysis," *Journal of Information Science*, vol. 44, no. 2, pp. 230–247, 2018, doi: 10.1177/0165551516684296.

[33] H. Liang, X. Sun, Y. Sun, and Y. Gao, "Text feature extraction based on deep learning: a review," *EURASIP journal on wireless communications and networking*, vol. 2017, no. 1, pp. 1–12, 2017, doi: 10.1186/s13638-017-0993-1.