MR-FNC: A Fake News Classification Model to Mitigate Racism

Muhammad Kamran[1], Ahmad S. Alghamdi[2], Ammar Saeed[3], Faisal S. Alsubaei[4]
Department of Cyber Security, College of Computer Science and Engineering, University of Jeddah, 21959, Saudi Arabia[1, 2, 4]
Department of Computer Science, COMSATS University Islamabad, Wah Cantt, Wah Cantt, 47010, Pakistan[3]

***Abstract*—One of the most challenging tasks while processing natural language text is to authenticate the correctness of the provided information particularly for classification of fake news. Fake news is a growing source of apprehension in recent times for hate speech as well. For instance, the followers of various beliefs face constant discrimination and receive negative perspectives directed at them. Fake news is one of the most prominent reasons for various kinds of racism and stands at par with individual, interpersonal, and structural racism types observed worldwide yet it does not get much importance and remains to be neglected. In this paper, to mitigate racism, we address the fake news regarding beliefs related to Islam as a case study. Though fake news remained to be a concerning factor since the beginning of Islam, a significant increase has been noticed in it for the last three years. Additionally, the accessibility of social media platforms and the growth in their use have helped to propagate misinformation, hate speech, and unfavorable views about Islam. Based on these deductions, this study intends to categorize such anti-Islamic content and misinformation found in Twitter posts. Several preprocessing and data enhancement steps were employed on retrieved data. Word2vec and GloVe were implemented to derive deep features while TF-IDF and BOW were applied to derive textual features from the data respectively. Finally, the classification phase was performed using four Machine-based predictive analysis (ML) algorithms Random Forest (RF), Naïve Bayes (NB), Logistic Regression (LR), Support Vector Machine (SVM), and a custom deep CNN. The results when compared with certain performance evaluation measures show that on average, ML-models perform better than the CNN for the utilized dataset.**

Keywords*—Machine learning; deep learning; fake news detection; social media

I. INTRODUCTION

One of the most challenging tasks while processing natural language text is to authenticate the correctness of the provided information particularly for classification of fake news. Fake news is a growing source of apprehension in recent times for hate speech as well. For instance, the followers of various beliefs face constant discrimination and receive negative perspectives directed at them. Fake news is one of the most prominent kinds of racism and stands at par with individual, interpersonal, and structural racism types observed worldwide yet it does not get much importance and remains to be neglected. In this paper, we address the fake news regarding beliefs related to Islam as a case study. Fake news is concerned with the type of racism done against Islam or Muslims which can be in the form of speech, text, news, attitude, behavior, or emotions. Any negative mention of Islam, Muslims, their mosques, rituals, religious practices, and holy books indicates the origin of fake news and recently United Nations Organization (UNO) has adopted a resolution to observe March 15 as international day to combat fake news [1]. Although it is one of the most prevalent forms of racism currently being noticed, no action has been made to confront or eradicate it. Web and social media are the major sources of spreading fake news-related content worldwide. As per Belgium's statement on fake news, political discussions and actions that were sanctioned addressing Muslim women's headscarves, the production of acceptable meat, and other Islamic traditions increased in 2017 [2]. Anti-Muslim acts have also increased significantly in China. As per a poll conducted in China during the year 2018, the acts of intolerance have been noticed against Muslims in terms of job opportunities, indoctrination availability, medical facilitation. They also face biasness in their social and electronic media representations [3]. According to a report distributed by 2019 in Europe, incidents like the Christchurch terrorist attack, Philip Manshaus' attempted attack on the Baerum mosque, and some physical assaults that took place in the United Kingdom post Christchurch incident are all the result of religious discrimination, hate speech, and derogatory social posts against the followers of Islam [4]. Furthermore, hate crime, social harassment, abusive behavior, and other Anti-Muslim activities have amplified over time in France [5], America [6] because of Donald Trump's statement regarding Anti-Islamic culture, Canada [7], Wisconsin [8], India during COVID-19 [9], and Israel [10]. Fake news is a major issue that is on the rise but has not been recognized on higher levels thus making it a matter of concern to figure it out and analyze its progress over the physical and electronic channels [40, [41]. Apart from this, the followers of Islam vary across different countries in terms of number which also impacts the level of fake news activities occurring in a particular region and there is no specific way to analyze it generically.

Some of the aspects that contribute to propagating fake news include lack of Islamic knowledge, a superficial and abstract understanding of its enactments, a non-acceptance behavior, extremist nature, and irresponsibility concerning other beliefs. A study done in USA during 1993 indicates that most people had negative ideas about Islam despite having little to no understanding of the religion. The balance of the population had positive impressions of Islam and knew something about it. This is because there is a deluge of inaccurate, unfriendly, and arsonist-related news about Islam and its followers on electronic channels. When a novice attempts to learn or research something, they only ever receive

these inaccurate feedbacks [11]. Even while most communal channels have developed hand-made strategies for the brisk rectification of extreme content, there is still a long way to go in terms of automating this process. It is necessary to move away from the labored process of going through each article, tweet, or any electronic post, in favor of an automated system that can input a corpus of textual data, identify the content, and categorize it. The developments in the field of processing semantics language, machine-driven understanding of stuff, and layered models have assisted in the development of such systems that can orderly perform such tasks with utmost accuracy [12]. These tools are being widely utilized in the latest social media platforms and have provided great results till now.

In the proposed work, we contribute to propose a fake news classifier (FNC) model as follows: fake news data instances retrieved from Twitter is analyzed and classified using a variety of ML frameworks with the addition of a DL schema; the labeled tweet data is dispensed as an input to the constituted model in the form of single and multiple combination sets, and the model's operation is divided into various layers; it is then passed through the phase of preprocessing and features extraction based on deep and text features extraction techniques including Word2Vec and GloVe, BoW, and TF-IDF respectively; the extricated textual attributes are then given to several ML models LR, SVM and RF while the features derived from word embedding models are provided to a custom CNN for classification; finally, the results are evaluated based on several execution assessment metrics.

The rest of the paper is as follows: Section II provides a literature review of the techniques used by previous works for Islamophobic news classification and text analysis. Section III provides insights of the proposed work of this research study. All the experiments conducted along with their results and performance evaluations are listed in Section IV. Section V provides a discussion of results obtained and finally, Section VI concludes the proposed work.

II. Related Work

In recent years, there has been a rise in the quantity of study looking at different religious organizations in terms of their racial makeup, religion, gender, representation, and degree of equality. Numerous studies have been conducted on the topic of fake news because of the surge in hate speech, online publishing, and social media content that is directed towards Muslims and Islam worldwide [13]. However, a few studies have been conducted on the automatic detection of anti-Islamic content. A brief synopsis of these investigations is provided in the material below.

Mehmood et al. [14] performed the hate speech identification and isolation from 1290 tweets that were part of a publicly available twitter collection. Out of the 1290 tweets gathered, 566 are classified as unfavorable and 724 as positive. Preprocessing processes for the raw data include case folding, tokenization, removing superfluous words, cleaning, and breaking words rectification. 1D-CNN and other RNN variations are created and used to conduct feature ex-traction and categorization. 80% of the data is utilized for model training, while the remaining 20% is used for model testing. Several RNN and CNN combinations are used to get the findings. While using the CNN with Bi-LSTM, which is the most accurate method, the maximum accuracy of 90% is achieved. Chandra et al. [15] presented a tweet based CoronaBias dataset to do the analysis of social media data for Islamophobic content. Between the months of February and March 2020, CoronaBias included 410,990 tweets, and every single one of them had terms relating to Islam or Muslims. BERT and SVM are used to annotate the data. A total of 2000 good and hate tweets are retrieved for model training, and the PELT method is used to do temporal analysis on them. Positive matrix factorization, which increases the model's capabilities, is used for feature derivation. The results are studied, compared, and tracked using graphical representations and a few metrics, and it can be shown that the proposed BERT model provides the best accuracy results with a rate of over 85%, which is higher than the SVM's rate of 79%.

In this work [16], Khan et al. gathered twitter data for the first six months of 2020. The 17,228 tweets that were gathered were annotated by skilled humans after going through necessary pre-processing procedures. The experiments are carried out using deep and textual feature extraction methods. Additionally, ML and DL approaches are used to classify tweets into different polarity categories. Because of embedment of deep and text attributes, the SVM model offers an accuracy more than 95% on the validation data. Alraddadi et al. [17] categorized text utilizing sentiment and text analysis techniques. Arabic dataset obtained for the 3 months of 2021is based on news articles and publications from famous search engines. The Octoparsescraing program is used to assemble the relevant data into an Excel spreadsheet. The input data is subjected to pre-processing methods while for feature selection, n-grams and term frequency computation models are used. Resultant data is sent to multiple ML algorithms for feature creation and data categorization based on labels. Results are derived using the model while considering several performance monitoring standards after data is divided into a 70-30% division. For word-level, balanced, and non-balanced datasets, the TF-IDF and SVM combination yields successful results with accuracy above 97%. Vidgen et al. [18] examined data from social media to categorize the content that contains strong and mild Islamophobic hate content. Tweets more than 100 million are collected from Twitter for the entire year of 2017 and the first half of 2018. The final dataset comprises 1300 tweets after 4000 tweets from this data collection are picked as a training batch and manually annotated. Several features are generated using deep and text extraction methods which are then supplied to ML and DL models for categorization. DL model delivers the best outcomes and perform almost equally, with accuracy of 71.14%. The results are obtained for various data sets, and they are afterwards further evaluated using certain metrics. The authors [19] used AI methods to analyze the data from social channels. The data utilized is based on comments from various writers' and authors' personal blogs and is collected based on several keywords. The findings, which were generated with the use of various execution criteria, demonstrate that the accuracy of the RF and bagging classifiers is practically identical at 0.66%, and pre-processing did not increase the results any further.

The authors in [20] examined tweets about Islamophobia from the time when the Christchurch incident occurred in 2019. The study's data is based on 3100 deleted tweets from the time of mishap. Tokenization, stop word removal, and scraping are a few of the preparation and refining processes the data goes through. NB and SVM models are used in the classification procedure. The results are based on both the unbalanced set and the data labelled using the Valence Aware Dictionary and Sentiment Reasoner. Along with the two ML models already described, the synthetic minority oversampling method is employed to derive and compare results. In the work [21], the authors also examined tweets for the trace of Islamophobic material. They gathered 150,000 tweets from 2018 and professional annotators manually categorized them into polarity ratings. Before sending the data to the ML-models, pre-processing is performed. Both in terms of accuracy (98.1%) and processing time, the RR classifier outperforms the Bayesian classifier. Gonzalez-Pizarro et al. [22] used contrastive learning to study the hostile attitudes on political data acquired from Papasavva. 134,5 million political postings from June 2016 to November 2019 are included in the data. In addition to this data, a collection of 5,859, 439 pictures were also collected from Zannettou. The data is given ratings, and after going through a few pre-processing steps, TF-IDF is utilized to look for hate speech content. For the extraction of features, all photos with a cosine similarity index of at least 0.3 are chosen and compared with the textual data using several API's. With precision up to 80%, the results show 69,000 antisemitic and 100,000 hate content from the entire data collection.

From the above discussion, it can be observed that some works have utilized embedding models, others have made use of n-grams for the derivation of useful data attributes. Also, there is a huge gap in Islamophobic content detection because to date only the above-mentioned studies have been conducted. Moreover, the results have been concluded based on either by using ML or DL models. Taking the lead from this, this study's conducted work focuses on implementing the combination of all for a better comparison of each model's performance on the currently utilized Islamophobic news data.

III. Proposed Methodology

The proposed research focuses on identifying and categorizing social media corpora that are associated with Islamophobia. The dataset of extreme and hateful tweets against Muslims and the Islamic faith was gathered from Twitter and other internet sources, pre-processed, cleaned, and subjected to several word embedding and n-gram algorithms, such as Word2Vec, Glove, TF-IDF, and BOW, for analysis. RF, SVM, LR, and a deep model CNN are some of the existing ML-algorithms that are used in the final step of data categorization. Accuracy, K-fold cross-validation, and F1-score are a few assessment metrics used to analyze the outcomes. Fig. 1 represents a detailed flow diagram of the proposed model.

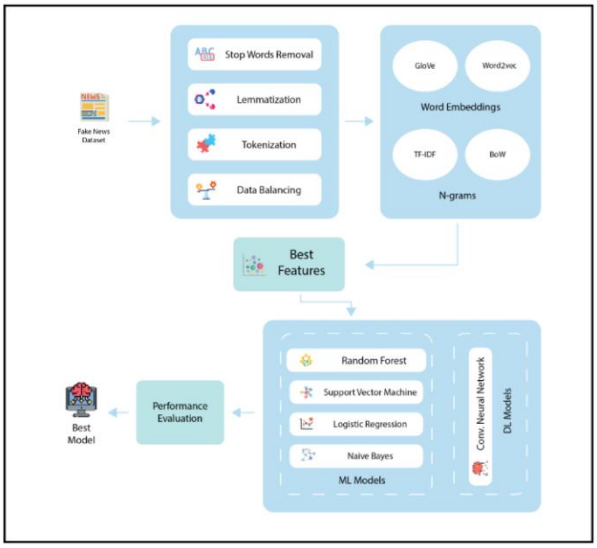

Fig. 1. Proposed framework.

A. Dataset

Tweet dataset exploited in the study is based on tweets and is not focused on a specific country to track the spread and effects of Islamophobia globally. Based on lexicons from Hate-base, some pre-defined hashtags were utilized for data retrieval during first six months of 2020 [23]. The data is scattered since it doesn't concentrate on user accounts, but it is nevertheless retrieved using an impartial technique. The 8438 English-language tweets in the study's dataset were pre-annotated by three annotators who are fluent in the language. The tweets were labelled into one of the three pre-defined categories namely Anti-Islamic, about Islam but having positive sentiment, and neither about Islam nor having any bad sentiments about Islam. The editors worked with data that was completely devoid of user and tweet identities. With considerable care, the annotators assigned the labels, and in cases of a tie, consensus casting ballots assignation was also employed. Table I provides an overview of data statistics.

B. Dataset Preprocessing and Balancing

Data preparation serves as the foundational stage for every classification work since it prepares, cleans, and removes ambiguities from the data [24]. Conversion of alphabets into smaller notations, end word rectification, hyperlink removal, removal of false full stops and half-sentences, tokenization and lemma generation are some of the preprocessing techniques used in this study. As a result of the utilized unbalanced data, created randomly using a variety of sources, the proposed model may not perform well [25]. To address this problem, the class with the greatest number of tweets is chosen, and those from the other two classes are replicated at random to keep the frequency of tweets across all classes under consideration equal. The experiments and findings derivation use the balanced data from this point forward. Following data balancing, the total number of tweets in each class is shown in Table II.

TABLE I. OVERVIEW OF DATASETS ATTRIBUTES

Attribute	Value	Attribute
Total tweets	8438	Total tweets
Tweets containing Islamophobic content	2485	Tweets containing Islamophobic content
Tweets about Islam but not Islamophobic	2398	Tweets about Islam but not Islamophobic
Tweets neither Islamophobic nor Islamic	3555	Tweets neither Islamophobic nor Islamic
Tweets language	English	Tweets language

TABLE II. TWEET COUNT FOR EACH TWEET CLASS

Label	Tweets
Islamophobic	3554
About Islam Not Islamophobic	3554
Neither Islamophobic nor About Islam	3555

The created vocabulary magnitude for the balanced data after pre-processing seems to be 17861 unigrams with the distribution of tweet-length set at 14 words each tweet. After pre-processing the data with eight words per tweet, the same magnitude drops to 16580 unigrams.

C. Feature Extraction

After preprocessing the data and balancing it, the next phase to be performed is feature extraction in which data is converted into vector attributes for the ML and DL models to interpret it. Two types of features are derived from dataset including word embedding based deep features and textual features which are described in next sections.

1) Word embedding: Word embedding is used to convert and represent textual data comprised of words into a vector and mathematical representation [26], [27]. Many models are available for this purpose, but in this study, we utilized the pre-trained GloVe from Stanford NLP and Wor2vec from Google news vectors. The GloVe is an unattended learning algorithm utilized for extracting word embeddings from the input data corpus based on the global word co-occurrence matrix. It is trained on global statistics of words included in a huge corpus compiled from online sources and when applied to any data, it directly obtains information about the words occurring frequently in that data and maps the words into vector spaces [28]. It has been widely utilized in text classification problems to derive features [29]– [31] and pass them on to classification models. It is based on the Log Bilinear (LBL) model that operates on the principle of weighted least squares [32] as Eq. (1) depicts.

$$d_a . d_b = logP\left(\frac{a}{b}\right) \quad (1)$$

Here, $d_a . d_b$ represent the weight density that any two data points carry within the corpus.Prepresents the co-occurrence probability of both the points. The complete working logic behind GloVe is represented in Eq. (2).

$$l = \sum_{a,b=1}^{n} f\left(G_{a,b}\right)(d_a^t d_b - \log G_{a,b})^2 \quad (2)$$

where, l represents loss function, $f(G_{a,b})$ is the function that maps least-squares between both the points starting from 1 to onwards, $d_a^t d_b$ is the density of both the data points concerning time t, and $log\, G_{a,b}$ is the log of the function containing the square computation of data points. Word2vec is also a word embedding technique that works based on shallow neural networks and utilizes the skip-gram method to achieve this functionality [33]. It creates vectors of textual data included in the corpus based on the frequency of documents and their co-occurrence matrix. Eq. (3) demonstrates how Word2vec uses the skip-gram approach to do computation.

$$\frac{1}{T}\sum_{p=1}^{D}\sum_{-s\le a\le s, a\neq 0} log\ prob(word_{p+1}|word_t) \quad (3)$$

where, D is the corpus proportionality, p is the position of $word_t$ in data, $\log prob(word_{t+1}|word_t)$ indicates the logarithm of $word_t$ concerning incrementing positions and co-occurrences within the document [34]. In the proposed work also, the preprocessed data is delivered to both GloVe and Word2vec models, and the features derived by them are later given a customized CNN and the results are evaluated.

2) N-gram methods: N-grams are any sequence of word tokens in each data where n = 1 denotes unigram, n = 2 denotes bigram, and so on. An n-gram model can compute the probability of n-grams within a data corpus and provide a prediction. The use of such models becomes useful in text classification tasks where there is a need to count the number of specific words included in the vocabulary from the corpus [35]. Such a metric is the TF-IDF, which assesses how closely a word in a catalogue is connected to its mood or meaning. It determines the frequency of each relevant text and generates phrases with an inverse frequency of those that appear often throughout multiple articles [36]. TF-IDF analyzes document terms frequency in each document [37] represented in Eq. (4).

$$weight_{a,b} = freq_{a,b}^t x \log\left(\frac{N}{freq_a}\right) \quad (4)$$

where, $weight_{a,b}$ indicates the total weightage carried by the data two points, $freq_{a,b}^t$ calculates the appearance ratio of data point a in b, N shows the total documents count included in the corpus, $\log(\frac{N}{freq_a})$ computes the log of all included documents with the frequency of data point a. The textual material to be categorized can also be used by BOW to extract valuable properties. It operates using a specified vocabulary and uses that vocabulary to seek for the frequency of specific terms in the relevant material. The model only deals with whether familiar words occur in the document and has no concern about where they occur and it provides the histogram of given words within the data which can be easily given the classifiers [38]. BOW performs the creation of bags based on words based on Eq. (5).

$$doc_b = \sum_{a=1}^{N} weight_a^b\, x weight_a \quad (5)$$

where, doc_b indicates the document housing the targeted data point b. $weight_a^b$ shows the numerical weights of the repeating word for concerned feature point b included in the document. $weight_a$ indicates the weight of frequent word a that we are looking for in this scenario [39]. In the proposed

work, both TF-IDF and BOW are used for the derivation of features from the preprocessed dataset. The extracted features from both these models are tested and classified using a set of four ML classifiers for classification.

D. Fake News Classification

After completing all the stages, the word embedding techniques' feature sets are fed into a DL algorithm called CNN, and feature sets derived from N-grams are directed to ML distributors. To perform the categorization against the derived attributes, the proposed work uses four ML-Classifiers: RF, SVM, LR, and NB well as a DL-based CNN that includes embedding, convolutional, max-pooling, and SoftMax layers.

IV. Experiments and Results

The suggested methodology uses word embedding and n-gram techniques to extract valuable characteristics from the input textual data based on Islamophobic news from social media, before performing classification using four ML algorithms and a CNN model. Word embedding features are first identified using a deep CNN, and then n-gram method-based features are sent to the four ML-algorithms for classification in a series of experiments based on word embedding, n-grams, ML, and DL model combinations. After balancing the dataset, each experiment is run, and the results are analyzed using a variety of performance analysis standards, such as recall, f1-score, 10-fold accuracy, and precision. In the first experiment, SVM is used to assess n-grams-based features. Table III mentions the results of SVM-TF-IDF and SVM-BOW with assessment metrics.

SVM is a ML classifier that is employed for the high-dimensional feature mapping process. Most frequently, it is used to categories and transforms data so that it may be used to sort records into their correct classifications. Using a renowned sklearn linear model package and n-gram based textual feature extraction techniques; we applied it to our categorical islamophobia data in the Python programming environment. 90% and 10% of the dataset are used, respectively, for training and testing the model. The number of folds is set to 10, and the maximum number of iterations is equal to 10000, for the k fold cross-validation procedure to test the model. As can be seen through Fig. 2, upon maintaining cross substantiation of 10-folds, it can be observed that the SVM in combination with the BOW technique obtains an accuracy of 97.3% as opposed to the 97.1% obtained by the SVM-TF-IDF.

In the following experiment, the RF classifier is given the same n-gram-based characteristics for the identification of Islamophobic material. Since the three used ML models are all the standard variety, we additionally used an ensemble model called Random Forest to further investigate the outcomes. Since Decision Tree is not an ensemble approach and produces almost identical hyperparameters as RF, we opted against using the most popular ML model. The library used to integrate the model into our environment is called sklearn ensemble, and the experiments employing this model are carried out using the Python programming language. 90% of the dataset is used to train the model, and 10% of the dataset is used to test it using k fold cross-validation with a fold size of 10 and 200 estimators.

The outcomes of the RF-TF-IDF and RF-BOW, with the identical assessment metrics, are shown in Table IV. Additionally, it can be noted that in this instance, when used in conjunction with the BOW model, the RF obtains an accuracy of 94.1% as opposed to the 91.4% obtained by the RF when utilizing TF-IDF when 10-fold cross-validation is maintained. As shown in Fig. 3, the RF and BOW combination also obtains greater f1-score, recall, and precision values than the RF-TF-IDF model.

TABLE III. Results of TF-IDF and BOW Features with SVM

PEM	SVM – TFIDF	SVM - BoW
10-Fold Accuracy	0.97	0.97
Precision	0.97	0.97
F1 Score	0.96	0.97
Recall	0.96	0.97

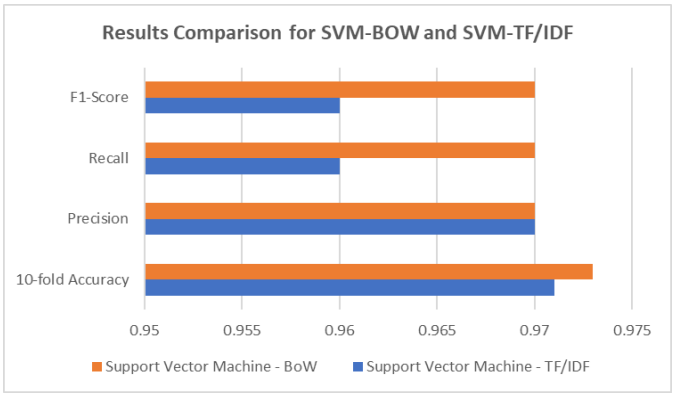

Fig. 2. Results comparison for SVM-BOW and SVM-TF / IDF.

TABLE IV. Results of TF-IDF and BOW Features with RF

PEM	RF – TFIDF	RF -BoW
10-Fold Accuracy	0.91	0.94
Precision	0.92	0.94
F1 Score	0.92	0.93
Recall	0.92	0.93

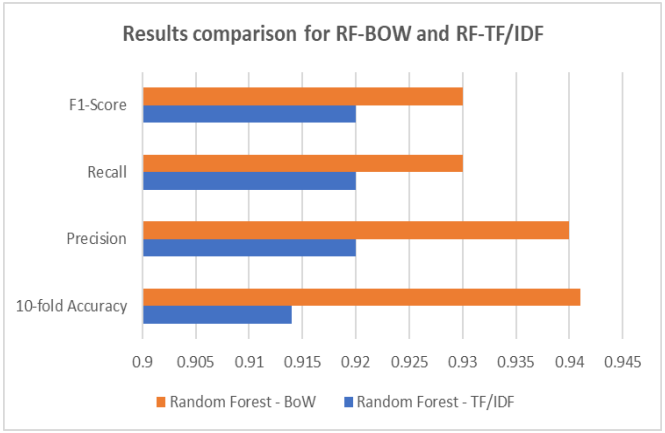

Fig. 3. Results comparison for RF-BOW and RF-TF / IDF.

TABLE V. RESULTS OF TF-IDF AND BOW FEATURES WITH LR

PEM	LR-TF-IDF	LR-BOW
10-Fold Accuracy	0.96	0.97
Precision	0.97	0.98
F1 Score	0.97	0.98
Recall	0.97	0.98

The experiment that follows uses an LR classifier to conduct classification based on the same n-gram characteristics as the ML models discussed before. Categorical data categorization in ML also employs LR algorithm. The finding of connections between probabilities and the outcome of the anticipated record is the first step in this model's major operation. We used this Python-based model to train and evaluate itself against our categorical data. 90/10 ratio is maintained for model's training and evaluation. Sklearn is the name of the library that was used to import this model into the experimental workspace. The outcomes of LR-TF-IDF and LR-BOW using the identical execution standards are shown in Table V

When TF-IDF linked model is maintained, BOW beats it by obtaining superior accuracy of 97.3 percent as opposed to 96.6 percent for the latter when coupled with an ML-classifier, in this instance LR. In addition, as shown in Fig. 4, LR-BOW outperforms LR-TF-IDF in all other performance metrics.

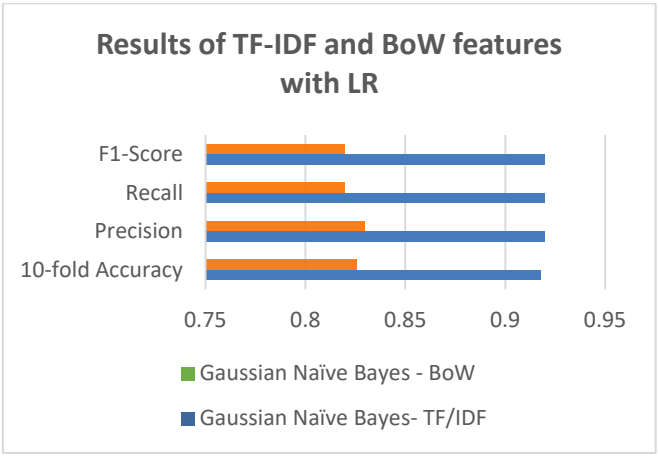

Fig. 4. Results comparison of TF-IDF and BOW with LR.

The GNB is used to classify with the features as an input, like in prior experiments, in this last experiment while utilizing an ML-classifier. Through its several iterations, which treat each input as an independent variable and forecast its likelihood, it aids in the rapid categorization of data. This technique is implemented in our codebase using the sklearn naive bayes package. For training and testing, the algorithm's Gaussian variant is employed, with data splits of 90% and 10%, respectively. The GNB-TF-IDF and GNB-BoW findings are displayed in Table VI.

When 10-fold cross-validation is maintained, TF-IDF outperforms its counterpart in this instance and obtains an accuracy of 91.8 percent as opposed to the 82.6 percent attained by the BOW-based model. In comparison to GNB-BOW, GNB-TF-IDF also outperforms it in all other performance metrics, as shown in Fig. 5.

The features retrieved by the word embedding models GloVe and Word2vec are further evaluated using a bespoke CNN after the implementation of four ML models with derived n-gram features. This model, which is a kind of deep neural networks, is primarily utilized for the accurate and quick categorization of vectorial data. The same data split utilized by ML algorithms is used to train and test this model when it is integrated into our software. The CNN is first trained and tested using Word2vec features with a batch size of 10 for model training and 32 epochs for testing. The number of epochs is kept at 5, and the batch size is kept at 32 for the k-fold cross-validation. Fig. 6 shows the training and validation loss for CNN using Word2vec.

TABLE VI. RESULTS OF TF-IDF AND BOW FEATURES WITH GNB

PEM	GNB-TF-IDF	GNB-BOW
10-Fold Accuracy	0.91	0.82
Precision	0.92	0.83
F1 Score	0.92	0.82
Recall	0.92	0.82

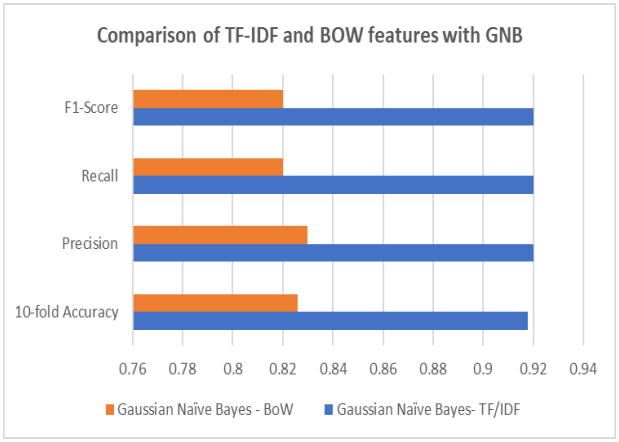

Fig. 5. Results comparison of TF-IDF and BOW with GNB.

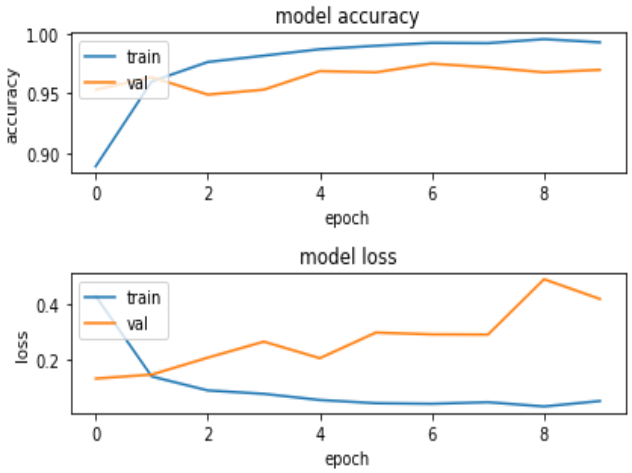

Fig. 6. Training and validation loss for CNN using Word2Vec.

Fig. 7 shows the training and validation loss for CNN using GloVe.

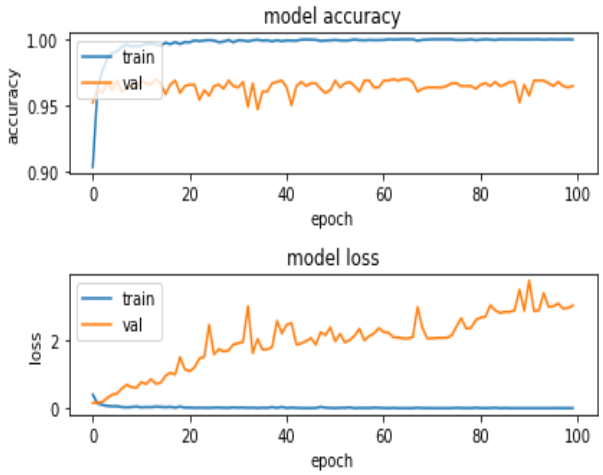

Fig. 7. Training and validation loss for CNN using GloVe.

Results for both embedding models with CNN are displayed in Table VII based on previously applied performance metrics.

TABLE VII. RESULTS OF WORD2VEC AND GLOVE FEATURES WITH CNN

PEM	CNN-Word2Vec	GNB-GloVe
10-Fold Accuracy	0.96	0.96
Precision	0.97	0.96
F1 Score	0.97	0.96
Recall	0.97	0.96

As observed in Fig. 3, Fig. 4, and Table VIII, CNN works a little bit better with Word2vec than GloVe because it obtains a better 10-fold accuracy and retains a higher evaluation rate. Table VIII gives a summary of all the experiments that were done above and gives a more comprehensive perspective of all the outcomes that were inferred.

TABLE VIII. RESULTS OF ML AND DL-MODELS WITH CORRESPONDING WORD EMBEDDINGS AND N-GRAMS

PEM	SVM-TFIDF	SVM-BOW	RF-TFIDF	RF - BOW	LR-TFIDF	LR-BOW	GNB-TFIDF	GNB-BOW	CNN-Word2vec	CNN-Glove
10-Fold Accuracy	0.97	0.97	0.91	0.94	0.96	0.97	0.91	0.82	0.96	0.96
Precision	0.97	0.97	0.92	0.94	0.97	0.98	0.92	0.83	0.97	0.96
F1 Score	0.96	0.97	0.92	0.93	0.97	0.98	0.92	0.82	0.97	0.96
Recall	0.96	0.97	0.92	0.93	0.97	0.98	0.92	0.82	0.97	0.96

V. DISCUSSION ON RESULTS

The findings of each experiment carried out for the planned study are covered in detail in the preceding section. It is clear from the trials on the characteristics that the n-gram models TF-IDF and BOW were able to extract using four ML-models that BOW outperforms TF-IDF in these scenarios. BOW outperformed its rival in terms of accuracy and other performance metrics when it was categorized using SVM, RF, and LR. Only when used in conjunction with GNB models did TF-IDF perform better. This demonstrates why it is preferable to use BOW-based features for the planned task.

The model also produced respectable results when Word2vec and GloVe word embedding features were classified using a custom CNN. The CNN-Word2vec model emerged as the superior one of the two because, as seen in Fig. 8, it performed better across the board.

The performance comparisons for the independently developed ML and DL models with n-gram and word embedding schemas are covered in the section above. However, this study demonstrates that ML-models typically outperform CNN in terms of categorization of the Islamophobic data utilized for this experiment. This shows that both SVM and LR outperform CNN at their maximal performance levels. Using both ML models stated, an average accuracy of 97 percent is attained, which is much higher than the 96.4 percent achieved by CNN.

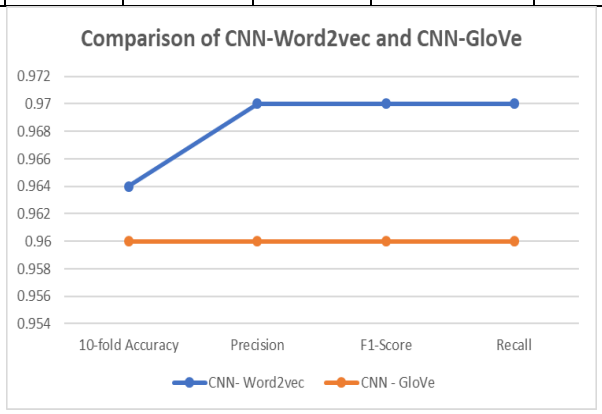

Fig. 8. Performance comparison of Word2vec and GloVe with CNN.

While the results of proposed work are quite encouraging, its performance is highly dependent on the characteristics of the data that can be further investigated in the future research work.

VI. CONCLUSION

It is alarming to see how hateful, exaggerated, extremist, and misunderstood misinformation about fake news and its impact on society is spreading on social media. Such material is accessible to a global audience, and as a result, various communities may be the target of harsh measures. The suggested project focuses on classifying such fake news

content using data from Twitter. It uses several data cleaning and improvement procedures to refine the data. The implementation of Word2vec, GloVe, TF-IDF, and BOW word embedding, and n-grams methods follows to extract key characteristics from the data. Finally, four ML algorithms and a CNN created by the customer are used to classify the data. On average, the ML models outperform CNN and produce superior outcomes. The assessment of DL algorithms on this data might be done in the future using more DL algorithms like LSTM and RNN. Bert is the most recent DL feature extraction model based on a transformer that is becoming increasingly prominent in the field of sentiment analysis on textual data. While the results of proposed work are quite encouraging, its performance is highly dependent on the characteristics of the data that can be further investigated in the future research work.

ACKNOWLEDGMENT

This work was funded by the University of Jeddah, Jeddah, Saudi Arabia, under grant no. (UJ-21-IMT-9). The authors, therefore, acknowledge with thanks the University of Jeddah technical and financial support.

REFERENCES

[1] https://www.arabnews.com/node/2043146/world. (Last visited on 24 August 2022).

[2] A. Easat-Daas, "Islamophobia in Belgium: national report 2017," 2018.

[3] J. Qian, "Historical Ethnic Conflicts and the Rise of Islamophobia in Modern China," Ethnopolitics, p. 1–26, 2021.https://doi.org/10.1080/17449057.2021.2001954.

[4] E. Bayrakli and F. Hafez, "The State of Islamophobia in Europe in 2018," Islamophobia Report, 2018.

[5] M.A. Valfort, "Anti-Muslim discrimination in France: Evidence from a field experiment," World Development, vol. 135, p. 105022, 2020.https://doi.org/10.1016/j.worlddev.2020.105022.

[6] M. H. Khan, H. M. Adnan, S. Kaur, R. A. Khuhro, R. Asghar et al., "Muslims' representation in Donald Trump's anti-Muslim-Islam statement: A critical discourse analysis," Religions, vol. 10, no. 2, p. 115, 2019.https://doi.org/10.3390/rel10020115.

[7] S. Elkassem, R. Csiernik, A. Mantulak, G. Kayssi, Y. Hussain et al., "Growing up Muslim: The impact of Islamophobia on children in a Canadian community," Journal of Muslim Mental. Health, vol. 12, no. 1, pp. 3–18, 2018.https://doi.org/10.3998/jmmh.10381607.0012.101.

[8] A. Mansson McGinty, "Embodied Islamophobia: lived experiences of anti-Muslim discourses and assaults in Milwaukee, Wisconsin," Social & Cultural. Geography., vol. 21, no. 3, pp. 402–420, 2020.https://doi.org/10.1080/14649365.2018.1497192.

[9] S. Banaji and R. Bhat, "How anti-Muslim disinformation campaigns in India have surged during COVID-19," LSE COVID-19 Blog, 2020.

[10] D. Saadi, K. Agay-Shay, E. Tirosh, and I. Schnell, "The effects of crossing ethnic boundaries on the autonomic nervous system in Muslim and Jewish young women in Israel," Scientific Reports, vol. 9, no. 1, pp. 1–9, 2019.https://doi.org/10.1038/s41598-018-38290-z.

[11] K. GhaneaBassiri, "Islamophobia and American history," in Islamophobia in America, Springer, pp. 53–74, 2013.https://doi.org/10.1057/9781137290076_3.

[12] E. M. Mahir, S. Akhter, and M. R. Huq, "Detecting fake news using machine learning and deep learning algorithms," in Proc. 7th International Conference on Smart Computing and Communications, pp. 1–5, 2019.https://doi.org/10.1109/ICSCC.2019.8843612.

[13] T. Mirrlees and T. Ibaid, "The Virtual Killing of Muslims: Digital War Games, Islamophobia, and the Global War on Terror," Islamophobia Studies Journal, vol. 6, no. 1, pp. 33–51, 2021.https://doi.org/10.2307/j50018795.

[14] Q. Mehmood, A. Kaleem, and I. Siddiqi, "Islamophobic Hate Speech Detection from Electronic Media using Deep Learning".https://doi.org/10.1007/978-3-031-04112-9_14.

[15] M. Chandra, M. Reddy, S. Sehgal, S. Gupta, A. B. Buduru et al., "'A Virus Has No Religion': Analyzing Islamophobia on Twitter During the COVID-19 Outbreak," in Proc. 32nd ACM Conference on Hypertext and Social Media, pp. 67–77, 2021.https://doi.org/10.1145/3465336.3475111.

[16] H. Khan and J. L. Phillips, "Language agnostic model: detecting islamophobic content on social media," in Proc. 2021 ACM Southeast Conference, pp. 229–233, 2021.https://doi.org/10.1145/3409334.3452077.

[17] R. A. Alraddadi and M. I. E.-K. Ghembaza, "Anti-Islamic Arabic Text Categorization using Text Mining and Sentiment Analysis Techniques," International Journal of Advanced Computer Science and Applications (IJACSA), vol. 12, no. 8, 2021.

[18] B. Vidgen and T. 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.https://doi.org/10.1080/19331681.2019.1702607.

[19] T. Massey, C. Amrit, and G. C. van Capelleveen, "Analysing the trend of Islamophobia in Blog Communities using Machine Learning and Trend Analysis," in Proc. 28th European Conference on Information Systems, ECIS 2020: Liberty, Equality, and Fraternity in a Digitizing World, 2020.

[20] W. Gata and A. Bayhaqy, "Analysis sentiment about islamophobia when Christchurch attack on social media," TELKOMNIKA Telecommunication Computing, Electronics and Control, vol. 18, no. 4, pp. 1819–1827, 2020.http://doi.org/10.12928/telkomnika.v18i4.14179.

[21] B. Ayan, B. Kuyumcu, and B. Ciylan, "Detection of Islamophobic Tweets on Twitter Using Sentiment Analysis," Gazi University Journal of Science Part C, vol. 7, no. 2, pp. 495–502, 2019.https://doi.com/10.29109/gujsc.561806.

[22] F. González-Pizarro and S. Zannettou, "Understanding and Detecting Hateful Content using Contrastive Learning," ArXiv Prepr. ArXiv220108387, 2022.https://doi.org/10.48550/arXiv.2201.08387.

[23] Hatebase, "Hatebase Database." Hatebase. Accessed: Jan. 12, 2022. [Online]. Available: https://hatebase.org/.

[24] S. B. Kotsiantis, D. Kanellopoulos, and P. E. Pintelas, "Data preprocessing for supervised leaning," International Journal of Computer Science, vol. 1, no. 2, pp. 111–117, 2006.

[25] S. Castelo, T. Almeida, A. Elghafari, A. Santos, A. Pham et al., "A topic-agnostic approach for identifying fake news pages," in Proc. Companion Proceedings of the 2019th World Wide Web Conference, pp. 975–980, 2019.https://doi.org/10.1145/3308560.3316739.

[26] M. Chen, "Efficient vector representation for documents through corruption," ArXiv Prepr. ArXiv170702377, 2017.https://doi.org/10.48550/arXiv.1707.02377.

[27] S. Lai, K. Liu, S. He, and J. Zhao, "How to generate a good word embedding," IEEE Intelligent Systems, vol. 31, no. 6, pp. 5–14, 2016.https://doi.com/10.1109/MIS.2016.45.

[28] Y. Sharma, G. Agrawal, P. Jain, and T. Kumar, "Vector representation of words for sentiment analysis using GloVe," in Proc. 2017th International Conference on Intelligent Communication and Computational Techniques, pp. 279–284, 2017.https://doi.com/10.1109/INTELCCT.2017.8324059.

[29] W. K. Sari, D. P. Rini, and R. F. Malik, "Text Classification Using Long Short-Term Memory with GloVe," Jurnal Ilmiah Teknik Elektro Komputer Dan Informatika, vol. 5, no. 2, pp. 85–100, 2019.https://doi.com/10.26555/jiteki.v5i2.15021.

[30] R. A. Stein, P. A. Jaques, and J. F. Valiati, "An analysis of hierarchical text classification using word embeddings," Information Sciences, vol. 471, pp. 216–232, 2019.https://doi.org/10.1016/j.ins.2018.09.001.

[31] A. Mahmoud and M. Zrigui, "Deep neural network models for paraphrased text classification in the Arabic language," in Proc. International Conference on Applications of Natural Language to Information Systems, pp. 3–16, 2019.https://doi.org/10.1007/978-3-030-23281-8_1.

[32] J. Pennington, R. Socher, and C. D. Manning, “Glove: Global vectors for word representation,” in Proc. 2014th Conference on Empirical Methods in Natural Language Processing, pp. 1532–1543, 2014.

[33] K. W. Church, “Word2Vec conventional neural networks for classification of news articles and tweets,” PloS One, vol. 14, no. 8, p. e0220976, 2019.https://doi.org/10.1371/journal.pone.0220976.

[34] D. Herremans and C. H. Chuan, “Modeling musical context with word2vec,” in Proc. First International Conference on Deep learning and Music, pp. 11¬–18, 2017.https://doi.org/10.48550/arXiv.1706.09088.

[35] K. Kowsari, K. Jafari Meimandi, M. Heidarysafa, S. Mendu, L. Barnes et al. “Text classification algorithms: A survay,” Information, vol. 10, no. 24, p. 150, 2019.https://doi.org/10.3390/info10040150.

[36] C. Liu, Y. Sheng, Z. Wei, and Y. Q. Yang, “Research of text classification based on improved TF-IDF algorithm,” in Proc. 2018th IEEE International Conference of Intelligent Robotic and Control Engineering, pp. 218–222, 2018.https://doi.com/10.1109/IRCE.2018.8492945.

[37] S. Qaiser and R. Ali, “Text mining: use of TF-IDF to examine the relevance of words to documents,” International Journal of Computer Applications, vol. 181, no. 1, pp. 25–29, 2018.

[38] H. Peng, J. Li, Y. He, Y. Liu, M. Bao et al., “Large-scale hierarchical text classification with recursively regularized deep graph-cnn,” in Proc. 2018th World Wide Web Conference, pp. 1063–1072, 2018.https://doi.org/10.1145/3178876.3186005.

[39] S. Ma, X. Sun, Y. Wang, and J. Lin, “Bag-of-words as target for neural machine translation,” ArXiv Prepr. ArXiv180504871, 2018.https://doi.org/10.48550/arXiv.1805.04871.

[40] Moreno-Vallejo, Patricio Xavier, Gisel Katerine Bastidas-Guacho, Patricio Rene Moreno-Costales, and Jefferson Jose Chariguaman-Cuji. "Fake News Classification Web Service for Spanish News by using Artificial Neural Networks." International Journal of Advanced Computer Science and Applications, vol. 14, no. 3, pp. 301–306, 2023.

[41] OUASSIL, Mohamed-Amine, Bouchaib CHERRADI, Soufiane HAMIDA, Mouaad ERRAMI, Oussama EL GANNOUR, and Abdelhadi RAIHANI. "A Fake News Detection System based on Combination of Word Embedded Techniques and Hybrid Deep Learning Model." International Journal of Advanced Computer Science and Applications 13, no. 10, pp. 525–534, 2022.