IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received July 1, 2021, accepted July 16, 2021, date of publication July 26, 2021, date of current version August 4, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3100435

Detecting White Supremacist Hate Speech Using Domain Specific Word Embedding With Deep Learning and BERT

HIND S. ALATAWI, AREEJ M. ALHOTHALI, AND KAWTHAR M. MORIA
Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah 21589, Saudi Arabia

Corresponding author: Hind S. Alatawi (halawti0003@stu.kau.edu.sa)

ABSTRACT White supremacist hate speech is one of the most recently observed harmful content on social media. The critical influence of these radical groups is no longer limited to social media and can negatively affect society by promoting racial hatred and violence. Traditional channels of reporting hate speech have proved inadequate due to the tremendous explosion of information and the implicit nature of hate speech. Therefore, it is necessary to detect such speech automatically and in a timely manner. This research investigates the feasibility of automatically detecting white supremacist hate speech on Twitter using deep learning and natural language processing techniques. Two deep learning models are investigated in this research. The first approach utilizes a bidirectional Long Short-Term Memory (BiLSTM) model along with domain-specific word embeddings extracted from white supremacist corpus to capture the semantic of white supremacist slangs and coded words. The second approach utilizes one of the most recent language models, which is Bidirectional Encoder Representations from Transformers (BERT). The BiLSTM model achieved 0.75 F1-score and BERT reached a 0.80 F1-score. Both models are tested on a balanced dataset combined from Twitter and a Stormfront dataset compiled from white supremacist forum.

INDEX TERMS BERT, deep learning, NLP, white supremacist, hate speech, Twitter.

I. INTRODUCTION

Social media has become an essential element of our society by which people communicate and exchange information on a daily basis. The strong influence of social media on internet users has been of great benefit to many individuals, businesses, and organizations. Many companies and organizations nowadays use social media to reach customers, promote products, and ensure customer satisfaction. Despite the benefits associated with the widespread use of social media, they remain vulnerable to ill-intentioned activities. The openness, anonymity, and informal structure of these platforms have contributed to the spread of harmful and violent content.

Although social media service providers have policies to control these ill-intentioned behaviors, these rules are rarely followed by users. Social media providers also allow users to report any inappropriate content, but unreported content is less likely to be discovered due to the huge volume of data on these platforms. Some countries have restricted social media use, and others have taken legal action regarding violent and harmful content that might target particular individuals or communities. However, these violations might end up unpunished due to the anonymous nature of these platforms, allowing ill-intentioned users to fearlessly share harmful content using nicknames or fake identities.

One of the most-shared harmful content on social media is hate content, which might take different forms such as text, photos, and/or video. Hate speech is any expression that encourages, promotes, or justifies violence, hatred, or discrimination against a person or group of individuals based on characteristics such as color, gender, race, sexual orientation, nationality, religion, or other attributes [1]. Online hate speech is rapidly increasing over the entire world, as nearly 60% of the world's population (≈ 3.8 billion) communicates on social media [2]. Studies have shown that nearly 53% of Americans have experienced online hate and harassment [3]. This result is 12% higher than the results of a comparable questionnaire conducted in 2017 [4]. For younger people, the results show that 21% of teenagers frequently encounter hate speech on social media [5].

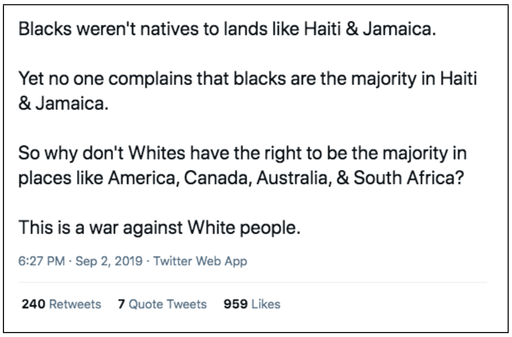
The associate editor coordinating the review of this manuscript and approving it for publication was Juan Wang.

One of the most dangerous and influential forms of online hate speech is led and spread by supporters of extreme ideologies who target other racial groups or minorities. White supremacists are one of the ideological groups who believe that people of the white race are superior and should be dominant over people of other races; this is also referred to as white nationalism in more radical ideologies [6]. White supremacists often claim that they are undermined by dark skin people, Jews, and multicultural Muslims, and they want to restore white people's power, violently if necessary. White supremacist hate speech has become a significant threat to the community, they use social media as a mean for communication and making movements to implement their goals in the real world. A study has also suggested links between hate speech and hate crimes against others (e.g., refugees, blacks, Muslims, or other minorities) [7]. Several recent brutal attacks have also been linked to radical white supremacists supporters who were very active members on social media [8], [9].

From a psychological point of view, any violent attack must be preceded by warning behaviors. This behavior happens prior to the violent attack associated with it and can help predict it in certain situations. Warning behaviors can be either real-world markers (e.g., buying a weapon and make an explosive missile) or linguistic markers or signs (e.g., "I had a lot of killing to do"), which can happen in real life or online [10]. Automatic detection of white supremacist content on social media can be used to predict hate crimes and violent events. Perpetrators can be caught before attacks happen by examining online posts that give strong indications of an intent to make an attack. Predicting violent attacks based on monitoring online behavior would be helpful in crime prevention, and detecting hateful speech on social media will also help to reduce hatred and incivility among social media users, especially younger generations.

Studies have investigated the detection of different kinds of hate speech such as detecting cyberbullying [11]–[13], offensive language [14], [15], or targeted hate speech in general by distinguishing between types of hate speech and neutral expressions [16]–[18]. Others have dealt with the problem by detecting a specific types of hate speech, such as anti-religion [19], [20], jihadist [21]–[24], sexist, and racist [25]–[27]. However, less attention has been given to detecting white supremacist content in particular, with only one study that uses white supremacist data [28].

White supremacist extremists tend to use rhetoric (i.e., the art of effective and compositional techniques for writing and speaking) [29] in their language. They also use specific vocabulary, abbreviations, and coded words to express their beliefs and intent to promote hatred or encourage violence to avoid being detected by traditional detection methods. They also mostly use hate speech against other races and religions or claim that other races undermine them. Figure 1 shows an example of a white supremacist tweet.



Blacks weren't natives to lands like Haiti & Jamaica.

Yet no one complains that blacks are the majority in Haiti & Jamaica.

So why don't Whites have the right to be the majority in places like America, Canada, Australia, & South Africa?

This is a war against White people.

6:27 PM · Sep 2, 2019 · Twitter Web App

240 Retweets **7** Quote Tweets **959** Likes

FIGURE 1. Example of white supremacists claim they are undermined by other race.

A. RESEARCH GOAL AND CONTRIBUTIONS

In this paper, we aim to detect white supremacist tweets based on textual features using deep learning techniques. We collected about $1M$ tweets from white supremacist accounts and hashtags to extract word embeddings, and then we labeled about $2k$ subsets of the data corpus to create a white supremacist dataset. We applied two approaches: the first uses domain-specific word embedding learned from white supremacist corpus and BiLSTM-based deep learning classifier.This approach is evaluated on multiple datasets achieving F1-scores ranged from a 0.49 to a 0.75 F1-score. The second approach uses a pre-trained language model that is fine-tuned on the white supremacist dataset using a neural network dense layer. The BERT language model F1-scores ranged from 0.59 to 0.80.

Thus, the research contribution can be summarized as follow: 1) This research, to the best of our knowledge, is the first attempt to explicitly target white supremacy (white supremacist vs. non-white supremacist) or right-wing hate detection in the English language. Only one previous study has investigated automatic detection of hate speech (hate vs. non-hate) on a dataset collected from a white supremacist forum (Stormfront) [28]. 2) This research is the first study to build a domain Specific word embedding from white supremacist content. 3) This research is the first approach that examines BiLSTM with domain-specific embedding in detecting white supremacy (Stormfront dataset), showing 2.0 points improvement over [28]'s approach. 4) In this research, a new dataset is built of 2000 English tweets consisting of the most recent white supremacist posts on Twitter. 5) This research is the first study to examines the BERT language model on white supremacist detection providing 6 points improvements compared to de Gibert [28] result, accordingly providing an important baseline for the future work comparison.

The rest of the paper proceeds as follows: Literature Review section (Section II) covers the related studies in hate speech detection, Background section (Section III provides information on the utilized methodology in this study, the Methodology section (Section IV) gives a detailed description of the proposed methods, Dataset section

(Section V) provides details of the used datasets, Experiment and Results section (Section VI) presents specifications of the experiment and the results, the Discussion section (Section VII) provides an analysis of the performance of each the proposed approaches, and finally, the Conclusion and Future Work section (Section VIII).

II. LITERATURE REVIEW

This section covers prominent studies related to hate speech detection, focusing on studies that utilized domain-specific embedding and targeted types of hate speech related to white supremacy. There has been a considerable research effort with regard to hate speech detection, but not much effort into specifically detecting white supremacist hate speech as only one study have looked into detecting hate speech in white supremacy forum.

Liu *et al.* [30] introduced hate speech word embedding to achieve higher accuracy in hate speech detection and achieved 0.78 F1-score by using word embeddings trained on Daily Stormer articles extremist website and high centrality users' tweets. They found that Convolutional Neural Network (CNN) performed better than LSTM on tweets due to the short-term dependency in the tweets. The study was based on 140 characters tweets, but Twitter extended tweet lengths to 280 characters. A comparative study was conducted by Gupta *et al.* [31] to assess the performance of the Word2Vec model to detect hate speech on three datasets and achieved maximum performance of 0.91 F1-score using domain-specific Word2Vec embeddings and Logistic Regression (LR) classifier. They concluded that domain-specific word embedding provides better classification results and is suitable for unbalanced classes.

Nobata *et al.* [32] used a pre-trained word embedding model and a regression model to detect abusive language on different domains. Their approach achieved a 0.60 F1-score on a finance domain and a 0.65 F1-score on a news domain, but Word2Vec domain-specific word embedding provided better performance, with 5% improvements on both domains. The study of Badjatiya *et al.* [26] examined different deep learning (i.e., deep neural networks) and machine learning models (i.e., Logistic Regression (LR)), Random Forest(RF), Gradient Boost Decision Tree (GBDT), Support Vector Machine (SVM), and with different word embedding models for detecting hate speech on a benchmark dataset. Deep learning models like LSTM, CNN, Fast-text were examined to build domain-specific embedding tuned towards hate speech labels. Their best F1-score was 0.93 obtained using GBDT as a classifier and random embeddings tuned using LSTM as features. They reported that domain-specific embeddings learned using deep neural networks expose ''racist'' or ''sexist'' biases for different words. From the above studies, domain-specific-based detection has good performance as it provided more accurate semantic representations of domain-related hate words frequently used by users in a given domain.

Several studies have looked into detecting particular types of online hate speech that target others based on their racial and cultural identities. Hartung *et al.* [33] classified Twitter profiles into either right-wing extremist or not. They used lexical (BOW model), emotional, lexico-syntactic, social identity features, and a combination of all the features with a linear SVM classifier. They reported an F1-score of 0.95 achieved using BOW features that outperformed all other features combined. They reported the most common features for each class by performing a qualitative analysis on the German language (e.g., asylum seekers, citizens' initiative, demonstration, and autumn offensive). They found that the content of tweets is a good indicator for hateful accounts.

Hartung *et al.* [34] in another study identify German right extremist accounts and rank unknown profiles based on their relative proximity to other users in the vector space. They used four feature sets: lexical (word stems), social identity, emotional, and lexico-syntactic (sentence constructions). The proposed model represented each Twitter profile as a point in a high-dimensional vector space based on the account textual content. The classification result was a 0.65 F1-score obtained using an unbalanced discrete decoding model over all the subsamples. The results also showed that the F1-score increased to 0.81 when profiles had greater than 100 tweets. This shows that the proposed ranking model depends heavily on the number of tweets of the profile. However, this condition may not apply to extremist profiles as they often use newly created accounts, as found in other studies [16].

The most recent and related study focuses on detecting hate speech in a white supremacy forum (i.e., Stormfront) [28]. Their proposed model is trained and tested on a balanced subset of a dataset consists of about $2k$ sentences collected from the Stormfront forum. Several machine learning approaches were examined to detect hate speech in this dataset, including SVM [35], Convolutional Neural Network (CNN) [36], and LSTM [37]. The results show that the LSTM outperforms the other models with an accuracy of 0.78 and 0.73 with and without excluding sentences requiring extra context. The main limitation of this study is annotating sentences extracted from paragraphs without providing any additional knowledge that might help to understand the sentences' context to label them accurately.

III. BACKGROUND

This section provides background on the most recent and commonly used word embedding models and language models. Researchers have continuously investigated techniques to represent words semantics such as word2vec [38] and Glove [39]. Also, several pre-trained language models have recently received massive attention in the different NLP tasks, such as Bidirectional Encoder Representations from Transformers (BERT) [40], which provides the state-of-the-art results for many NLP problems.

A. WORD EMBEDDING

Word embedding [41] is one of the most popular recent Natural Language Processing (NLP) trends. It refers to techniques that map words to dense vector representations capture words' semantic/meanings in specific language [30]. The primary purpose of this mapping is to represent linguistic terms in dense vectors to be utilized by machine learning algorithms. Word embedding has proven to be a powerful technique for extracting the most meaningful representations of words based on their context [30]. The evolution of word embedding has resulted in tremendous success in various NLP tasks like text classification [27], [42], document clustering [43], part of speech tagging [44], named entity recognition [45], sentiment analysis [46]–[48], and so on. Many researchers have built models to reach the best meaningful word representations either using neural network models (e.g., Word2Vec [38]) or using co-occurrence statistics and matrix factorization techniques(e.g., GloVe [39]).

1) Word2Vec

Word2Vec was developed by the Google research team [38] to represent words in dense-dimensional space based on its context. Word2Vec is a prediction-based model in which a loss function is used to evaluate the prediction performance. The meaning of words is obtained from surrounding words within a specified window size, the resulting word vectors from the model are considered as features representing the meaning of the word in many NLP problems. Google released pre-trained word embeddings named Google Word2Vec trained using skip-gram and continuous bag of words models on a large news corpus of 100 billion words [49].

2) GloVe

GloVe (Global Vectors for Word Representation) is another word embedding model [39] that obtains a vector representing words' meaning using corpus-based distributional features. The algorithm performs several operations on a constructed word-to-word co-occurrence statistics matrix. Different pre-trained GloVe versions are released that are trained on different datasets such as Wikipedia and Twitter [39].

B. BERT PRE-TRAINED LANGUAGE MODEL

A pre-trained language model can be defined as a black box that has previous knowledge about the natural language and can be applied and fine-tuned to solve various NLP problems. The pre-training process uses inexpensive unlabeled data to learn the initial parameters of a neural network model. BERT [40], is the latest revolution in NLP pre-trained language model trends. It is a deeply bidirectional language model trained on very large datasets (i.e., Books corpus and Wikipedia) based on contextual representations. BERT model can be fine-tuned using a neural network dense layer for different classification tasks. The fine-tuning advantage incorporates the contextual or the problem-specific meaning with the pre-trained generic meaning and trains it for a specific classification problem. BERT provides high performance for NLP tasks and improves the results of traditional models.

BERT is deeply bidirectional by jointly conditioning both left and right contexts in all layers [50] in contrast to Word2Vec and GloVe, which generate an embedding in one direction regardless of its contextual differences (context-free models). BERT models have different releases that differ according to model size, cased or uncased alphabet, languages, and the number of layers, and they are all available online.

C. DEEP LEARNING

Deep learning (also known as layered representations learning and hierarchical representations learning) is a sub-field of machine learning which uses successive layers for accurate representations or making decisions [51]. The learning process is performed by feeding training data to the model and estimate model parameters or weight to give the desired target. To control the prediction process, a loss function is used to measure the distance between the predicted and actual targets. The neural networks are structured in layers, and different constructions of layers give different deep learning models. Neural Networks form the basis for deep learning, and one of the common neural network architectures used for deep learning construction is Long Short-Term Memory (LSTM).

1) LONG SHORT-TERM MEMORY (LSTM)

LSTM is a recurrent neural network (RNN) developed to solve the problem of vanishing gradient in RNNs [37]. An RNN is a specific type of neural network which considers the history or context in the computation of the output. To predict the output of the current input, RNN uses computational results from the previous set of hidden units in the network. This design is useful for sequential learning tasks like speech recognition and stock forecasting that need the use of history in the decision-making process. The vanishing gradient problem occurs when the weights of an RNN are lost in a deeper layer of the network, resulting in the failure to capture very lengthy dependencies. To avoid this problem, LSTM replaces each node with a memory cell, which consists of an input gate, forget gate, output gate, and a node connected back to itself. The memory cell in a specific layer uses the hidden state in the previous layer during the current time and the hidden state of the current layer from the previous time. The forget gate decides which information should be ignored in the cell state, and the input gate and hyperbolic tangent (tanh) layer decide which information is stored in the cell state, then using the sigmoid function to determine the final output [30]. The bidirectional Long Short-Term Memory (BiLSTM) is an extension of the traditional LSTM that processes the sequence in both directions [52].

IV. METHODOLOGY

We used two approaches to investigate white supremacist hate speech detection: domain-specific word embedding with a deep model (BiLSTM) and BERT language model. Domain-specific word embedding is able to detect most terms, abbreviations, and intentional misspellings related to the white supremacist hate community, which are not detectable by the general embedding model since it is trained on books and Wikipedia textual data that often do not contain misspellings. However, we also used BERT because it has proved to provide state-of-the-art for most NLP problems showing better performance than some of the traditional domain-specific methods [40]. The following subsection provides more details about the two approaches used in this research.

A. DOMAIN-SPECIFIC WHITE SUPREMACY WORD EMBEDDING AND BiLSTM DEEP MODEL

To obtain domain-specific embedding, we perform the following steps:

1) DATA COLLECTION AND ANALYSIS

Domain-specific word embedding involves word representations constructed from a corpus of a specific domain (e.g., politics, finance, sports). As mentioned earlier, using white supremacist domains to extract embedding helps to identify terms that are commonly used in their community.

To create the domain-specific word embedding, we first collected a corpus consisting of 1, 041, 576 tweets. The tweets were obtained from known white supremacist hashtags such as #white_privilege, and #it_is_ok_to_be_white. We also collected contents from accounts that identified themselves as white supremacist explicitly (e.g., Whit***er) or implicitly (e.g., Na***st), and/or shared supportive phrases for white supremacy in hashtags or tweets encouraging or promoting racial or religious hatred against others. Then, we analyzed the corpus data to have an overall look at the most-used terms in that corpus. Figure 2 shows the terms most commonly used by their community, and they are different from general hate speech terms.

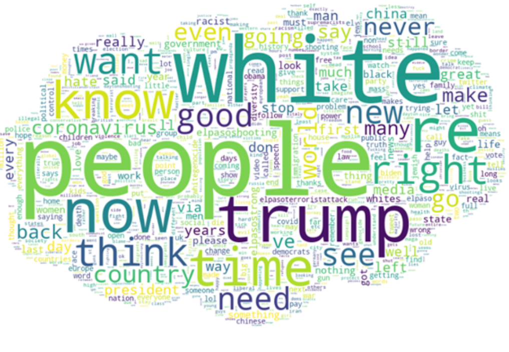


FIGURE 2. Word cloud of the most used terms in the white supremacist corpus.

We also analyzed the influence of using domain-specific word embedding of white supremacist hate speech by using word similarity. Word similarity retrieves the most similar words to the input word according to the cosine distance between their embedding vectors. Table 1 includes the data analysis of our domain-specific pre-trained model and general Google Word2Vec, GloVe models. The results show significant differences between the word embedding models. As can be seen, the words appearing in the domain-specific model tend to be more racist, while the Word2Vec and GloVe models provide the generic meaning of a word; for example, ''Black'' in domain-agnostic models tend to refer to the color of an object, while ''African'' appears to be similar in the domain-specific model which is the most commonly used meaning in the white supremacy context. This observation confirms [26] results.

TABLE 1. Examples of words that commonly appear in white supremacy content and the most similar words in pre-trained word embedding models (Word2vec and Glove) and the domain-specific word embedding.

The word	Google Word2Vec	GloVe	Domain Specific
Black	white, Responded_Letterman_How, blacks, crypt_inscribed, transporting_ petrochemicals, brown	white, dark, blue, brown, red, colored	blacks, white, asian, Hispanic, latino, latinx, african
Muslim	muslims, Muslim, Moslem, islamic, moslem, christian	muslims, moslem, Islamic, sunni, shiite, moslems	Muslims, Islamic, somali, pakistani, hindu, islamist
Race	races, Race, racing, sprint, Rain_postpones_Martinsville, Races	races, racing, winner, finish, event, runner	races, ethnicity, racial, existence, ethnicities, tribalism

2) PRE-PROCESSING

Pre-processing techniques are often used to remove noise and exclude unrelated words. In this study, we removed URL links (http or www), user names (@user_name), numbers, symbols. We also lowered the case of the text alphabet and handle the negation abbreviation such as (''can't'': ''cannot''). Text normalization such as stemming and handling misspellings were excluded from pre-processing as the hate community intentionally used misspellings to avoid being detected. Stemming was also excluded as it aims to remove prefixes and suffixes of the word, but some words (e.g., blacks) are used more frequently as hate words.

3) FEATURE EXTRACTION (Word2Vec)

To build our domain-specific embedding from the white supremacy textual data, we trained the Word2Vec model on the collected white supremacy tweets. The training was performed using the Gensim library with the Continuous Bag of Word (CBOW) model, a window size of five words, which is the number of surrounding words, and a 300-vector size representing the dimension of the output vector. The CBOW model aims to predict a target word from its neighboring words. The result of this stage is word embeddings of the corpus words (i.e., the vocabulary). The domain-specific embedding method in this study will be referred to as White Supremacy Word2Vec (WSW2V).

4) DEEP LEARNING MODEL

A number of experiments were carried out to choose the best deep learning model by testing different structures, depth, or parameters of models. Based on these experiments, and to the best of our knowledge, we found that sequential learning models are the most suitable approaches for this problem. Thus, we investigated two sequence models, namely, Bidirectional GRU (BiGUR) and Bidirectional LSTM (BiLSTM). We implemented the algorithms on the same structure of layers, starting with the embedding layer, sequence models layer, and two dense layers. We compare the performance of the models on a balanced subset of the StromFront dataset [28](See Section V for further details). The results show that BiLSTM provides the best score which reached up to 0.792 F1-score while BiGRU achieved 0.774 F1-score.

Thus, BiLSTM model was chosen in this task. The BiLSTM deep model structure consists of four layers the first layer is an embedding layer, with 300 dimensions. The second layer is Bidirectional CuDNNLSTM, a fast LSTM implementation backed by CuDNN, which is a library by NVIDIA CUDA described as a GPU-accelerated library of primitives for deep neural networks [53], [54]. The main advantage of using the bidirectional model of LSTM is to preserve both forward and backward data. This will allow understanding more information about the context, but will also require more computation time, and for this reason, we used CuDNNLSTM to reduce the processing time. The third and fourth layers are dense layers with linear and sigmoid activation functions, respectively. The activation functions were chosen after examining different activation functions like rectified linear (relu), sigmoid, hyperbolic tangent (tanh), and a linear function. The BiLSTM deep model structure is shown in Figure 3.

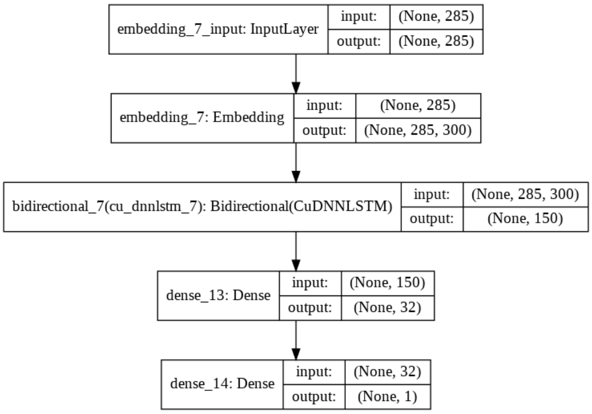


FIGURE 3. BiLSTM deep model structure starting from the embedding layer to the classification layer.

The loss is calculated using binary cross-entropy loss function, and the model is optimized using the Adam optimizer. We tested the model over different values of epochs to eliminate the gap between the validation and training accuracy and prevent over-fitting. While analyzing the results, we discovered that the ideal number of epochs is between 5 and 10, resulting in a validation and training accuracy gap of 0.13 to 0.17 points, with no degradation in validation accuracy and the gap attributable to an increase in training accuracy. We chose 10 epochs which gave a training accuracy of 96.0% and validation accuracy 79.0%. We used a 256 batch size to classify each tweet. We also divided the data sets into 20% for testing, 20% for validation and 60% for training.

B. BERT LANGUAGE MODEL

Our second approach employs the pre-trained language model BERT, which is used to encode the input text based on its own embedding strategy. We used the Bert For Sequence Classification [55] model, in which the last layer is a classification neural-network layer.

We used the BERT-Base model which contains 12 transformer layers, for each transformer,12 self-attention heads and hidden states size is 768. In comparison, the BERT-Large model contains 24 transformer layers and 16 self-attention heads, and the hidden states size is 1024. The model specifications are: LEARNING_RATE $= 2e - 5$, NUM_TRAIN_EPOCHS $= 3.0$ and BATCH_SIZE $= 16, 8$ for training testing respectively. The parameter are chosen according to the literature recommendations for similar problem.

V. DATASETS

This section describes the datasets we used in the experiments. We experimented on two datasets: a Stormfront dataset collected from white supremacist extremist content, which is available online (the Stormfront forum was later shut down because of its support for racial hate), and Twitter dataset (Twitter White Supremacy Dataset) to assess the performance on the recent white supremacist content on social media.

A. AVAILABLE DATASET (STORMFRONT DATASET)

To the best of our knowledge, there is no dataset available for white supremacy content other than the Stormfront dataset [28], which is a dataset collected from the Stormfront white supremacist forum. The dataset consists of a set of sentences that were extracted from posts that have been randomly sampled from several sub-forums in the Stormfront forum and were manually labeled as ‘containing hate speech or not’, ‘skip’, or ‘relation’ (relation means it needs extra context to be annotated), according to certain annotation guidelines. The average Cohen’s kappa annotator agreement score for a batch of $1,144$ sentences of the dataset is 61.4 for three classes (i.e., hate, non-hate, skip) and 62.7 for four categories (i.e., hate, non-hate, skip, and relation) for $1,018$ sentences of the dataset. The Cohen’ kappa percentage does not represent the entire published dataset and is calculated for three or four classes. Their classification experiment was performed on a balanced subset of the dataset, which included only hate and non-hate and excluded other classes.

B. TWITTER WHITE SUPREMACY DATASET

The aforementioned dataset was obtained from the Stormfront website, which has been taken offline and no longer available for research purposes. Thus, and to assemble white supremacist posts from different platforms, we collected a dataset from Twitter by randomly selecting subsets of tweets from the white supremacist corpus. The dataset consists of 1, 999 tweets that were annotated by three judges through Amazon Mechanical Turk (AMT) [56]. The judges have to be located in North America and have a hit approval greater than 80%. The location criterion was chosen to ensure that the reader/annotators fully understood common cultural terminologies, events, figures, and coded words.

The annotation procedure initially consisted of four labels: explicit white supremacy (EWS), implicit white supremacy (IWS), other hate speech (O), and neutral(N). Explicit white supremacy content refers to hate speech/tweets that express either racial or religious hatred towards others or claims of being undermined by other racial or religious groups (e.g., ‘‘These people do not want solution, They only want you dead White man’’. Implicit white supremacy content refers to expressing racial or religious hatred either indirectly or implicitly without using explicit hate terms (e.g., ‘‘we own our diversity, leave our country’’). Other hate speech is any hateful text other than white supremacist hate text, such as misogyny (i.e., hatred of women), homophobia (i.e., hatred of LGBT people), or sexism (i.e., discrimination based on gender). Neutral text, on the other hand, is any content that expresses positive subjective content (e.g., ‘‘Always brother’’), factual text (e.g., weather situation), or any other content not intended to promote or encourage hatred. Neutral also includes textual information that is challenging and cannot be annotated as hate speech or non-hate speech due to ambiguous intentions or contexts, e.g., ‘‘to be against immigration does not mean to kill people’’ or ‘‘die for them’’. Also, any factual text that includes hate terms with no hate intent is considered as Neutral (e.g., ‘‘Christchurch mosque shooter to be sentenced on August 24’’).

The annotators' agreements for the four labels were very low, with a 0.070 Cohen's kappa score. This is because there were large numbers of disagreements between the annotators, especially regarding neutral and implicit white supremacism. The disagreements between the judges were analyzed by counting the number of conflicts between every pairs of annotators. We found that the disagreements between annotators often occurred when the intention of the writer was inexplicit. For example, in the first example in Table 2 two annotators considered a tweet as implicit white supremacist content, while the third annotator considered it as neutral. The annotators disagreed on whether the writer had intended to discuss some statistical and factual information about immigration or to promote hatred against immigrants. This ambiguous content can reduce the agreement between annotators, increases the difficulty of detecting harmful content based on the content of the tweet, and requires the entire user profile to give an additional indication about the user's intent. Other examples of tweets that were found to be challenging for the annotators can be found in Table 2.

TABLE 2. Examples of the annotators' disagreements (A: annotator, N: Neutral, IWS: Implicit White supremacy).

No.	Tweet	A1	A2	A3
1.	“Carrying capacity is irrelevant our immigration rate is basically the highest in the OECD 5x the historical NOM rate, 100s of per year is radically changing the demographic composition which is nothing but an insane destructive experiment"	N	IWS	IWS
2.	“wonder if any of the African Dictators have her penciled in”	IWS	IWS	N
3.	“This really confuses me not seeing these claims she speaks of but am seeing Christians being murdered”	IWS	N	N

In this research, we treated the problem as a binary classification problem; thus, we collapse the four labels to binary labels (white supremacy or non-white supremacy). Explicit and implicit white supremacy were collapsed into a ‘white supremacy’ label, and the ‘neutral’ and other hate categories were collapsed into ‘non-white supremacy’ because our goal was to detect white supremacists in particular. We calculated the Cohen's kappa coefficient for the two labels (0 for non-white supremacy, 1 for white supremacy) for all the annotators, and the agreement score was 0.11 Cohen's kappa score. The disagreement is due to the difficulty in detecting implicit white supremacy hate speech. Schmidt and Wiegand [57] recognized from previous studies that the hate speech annotation process is reasonably ambiguous, which results into low agreement scores. To handle the annotators' disagreements, we used a voting strategy by choosing the most common label among the three annotators, so if at least two annotators agreed on one label, either 0 or 1, this label will be used as the final tweet label. The dataset is available on GitHub.[1]

C. BALANCED COMBINED DATASET

We created a balanced combined dataset from the datasets used (Twitter and Stormfront) to train the model on the largest possible diversity of the data from different social platforms of white supremacist hate speech. We combined the Stormfront and Twitter datasets by aggregating them into one CSV file, and then balanced them according to the number of class with lower frequency, and randomly selecting other class examples. Table 3 show the details of the white supremacist datasets.

VI. EXPERIMENTS AND RESULTS

We applied two different experiments that were evaluated separately. The first experiment uses white supremacist domain-specific embeddings and the BiLSTM deep model, and the second experiment uses the pre-trained language model (BERT). The experiments were conducted on Google Colab to utilize GPU processor for fast execution.

[1] https://github.com/Hind-Saleh-Alatawi/WhiteSupremacistDataset

TABLE 3. Details of white supremacist datasets (EWS: Explicit white supremacy, IWS: Implicit white supremacy, O:other hate speech, and N: Neutral).

Dataset	Original labels	#white supremacy	#nonwhite supremacy	Total
Storm front dataset [28]	Hate, non-hate	1,196	9,748	10,944
Twitter white supremacist datasets	EWS, IWS,O,N	1,100	899	1,999
Combined balanced dataset	Hate, non-hate	2,294	2,294	4,588

A. DOMAIN-SPECIFIC WHITE SUPREMACY WORD EMBEDDING AND DEEP MODEL EXPERIMENT

In this experiment, we used WSW2V embedding as features and the BiLSTM deep model as a classifier. We also used the domain-agnostic embedding models, both GloVe and Word2Vec embedding models, with the BiLSTM deep model and compared them against domain-specific WSW2V embedding. Table 4 shows descriptions of the embedding models used in this experiment.

TABLE 4. Details description of embedding models (WSW2V:White supremacy Word2Vec).

Models	**Dimension**	**Training set size**	**Pretraining context**
Word2Vec [2]	300	3B words	Google News
GloVe.Wikipedia [3]	300	6B tokens, 400K vocab, uncased	Wikipedia 2014+ English Gigawor Fifth Edition
GloVe.Twitter[3]	200	2B tweets, 27B tokens, 1.2M vocab, uncased	Twitter
WSW2V	300	1,041,576 tweets, 117083 vocab, uncased	Twitter

Table 5 compares the models' performance using different embeddings (domain-specific and domain-agnostic), and different classifiers (LR and BiLSTM deep Model). The first comparison is between the classifiers (LR, BiLSTM deep model) under the same features (WSW2V) identified by **BOLD** in the table to show the maximum F1-score. The results show that LR-WSW2V model outperforms BiLSTM-WSW2V on two datasets, but not on the Stormfront forum dataset which is the largest dataset among others datasets. The second comparison is between the features (WSW2V, GloVe, Word2Vec) under the same classifier (BiLSTM deep model) identified by underline in the table, we used domain-agnostic word embedding and compared it with domain-specific WSW2V embedding. The results show that our pre-trained word embedding WSW2V outperformed the other models except for the balanced combined dataset, GloVe.Wikipedia performs slightly better than WSW2V with a small margin of 0.006 point. It is worth mentioning that the WSW2V model is trained on only 1M tweets and while the domain-agnostic models is trained on at least 2B tweets.

We also evaluated the results of the proposed approach against similar research efforts in the field; however, the only study that had analyzed white supremacy content to detect hate speech is [28] study. The authors released the Stormfront dataset for research use, but they only reported the results for a sample (2, 000 sentences) of the dataset. Thus, we randomly sampled a balanced subset from that Stormfront dataset. The results show that the BiLSTM outperformed [28]'s result with an accuracy of 0.80 (only the accuracy is reported in [28]'s study). This result shows that our proposed model outperforms their model by 2.0 points, given that they used random word embedding for features and LSTM for classification, as shown in Table 6.

B. BERT PRE-TRAINED LANGUAGE MODEL EXPERIMENT

In the second experiment, we evaluate the BERT language model on the task of white supremacy detection. We chose the BERT model since it has shown high performance in many NLP tasks. We used both BERT Base and Large models. The results of the evaluation are reported in Table 7. As shown in the table, BERT provides better performance (F1-score) than the domain-specific model for white supremacist classification. It improved the F1-score by 3 points on the Stormfront dataset and by about 1 points on Twitter dataset and 6 on the balanced datasets. Also, we found that both BERT-Large and BERT-Base models give comparable results, BERT-base performs slightly better than BERT-Large model for some datasets.

In Table 8, we also compare the BERT model's accuracy with [28]'s result. BERT outperformed their accuracy by 6 points using the Base model, and also outperformed the domain-specific model by 4 points (Table 6).

VII. DISCUSSION

The first approach in this study utilizes domain-specific with deep learning model (BiLSTM). The results of evaluating the first approach (Table 6) show that domain-specific embedding with the BiLSTM model outperforms the results of random embedding and LSTM model [28]. Their reported accuracy was 78%, while our accuracy is 80%. This slight improvement shows that that random initialization of word embedding does not perform very badly. It is important to mention that the training corpus of white supremacist word embedding contains only 1 million tweets. Increasing the corpus size would provide a better performance, but we were limited by Twitter API's policies. This experiment also shows that the BiLSTM based deep model provided good performance for tweets classification, contrary to other prior conclusions that LSTM does not give a good performance due

[2] https://code.google.com/archive/p/word2vec/

[3] https://nlp.stanford.edu/projects/glove/

TABLE 5. Classification experiment results of domain-specific white supremacist using Linear Regression (LR) and bidirectional long short-term memory (BiLSTM) in comparison to domain agnostic pre-trained models on different datasets (Twitter white supremacist dataset, Stormfront, and a balanced combination of both datasets).

Methods	Word Embedding	Dataset Name	Precision	Recall	F1-score (hate)	F1-score (non- hate)	F1-score	AUC
LR	WSW2V-300	Stormfront dataset	0.87	0.89	0.238	0.95	0.864	0.848
		Twitter white supremacist dataset	0.67	0.67	0.730	0.623	**0.667**	0.721
		Balanced combined	0.74	0.74	0.749	0.743	**0.746**	0.819
BiLSTM	Word2Vec	Stormfront dataset	0.88	0.89	0.416	0.937	0.883	0.665
		Twitter white supremacist dataset	0.62	0.61	0.625	0.593	0.611	0.612
		Balanced combined	0.75	0.75	0.758	0.729	0.743	0.746
	GloVe.Wikipedia	Stormfront dataset	0.88	0.87	0.439	0.928	0.875	0.690
		Twitter white supremacist dataset	0.68	0.54	0.609	0.620	0.614	0.622
		Balanced combined	0.75	0.75	0.753	0.740	0.747	0.747
	GloVe.Twitter	Stormfront dataset	0.87	0.89	0.395	0.941	0.879	0.637
		Twitter white supremacist dataset	0.64	0.63	0.724	0.639	0.633	0.634
		Balanced combined	0.75	0.74	0.734	0.744	0.740	0.741
	WSW2V-300	Stormfront dataset	0.89	0.89	0.510	0.939	**0.892**	0.726
		Twitter white supremacist dataset	0.66	0.65	0.672	0.629	0.653	0.652
		Balanced combined	0.74	0.74	0.743	0.740	0.741	0.741

to the length of tweets which was limited to 180 characters; but, now it is 280 characters [30].

From the feature perspective comparison, Table 5 shows how WSW2V performs in comparison with other domain-agnostic models using the same classifier and datasets; the WSW2V outperforms other models on both the Stormfront and Twitter datasets, but GloVe performs slightly better than WSW2V on the balanced combined dataset. This shows that WSW2V provides a very desirable result despite the big size difference of training corpora (2B) for GloVe Twitter and (1M) for WSW2V. From the classifier perspective comparison, the BiLSTM-based deep model always performs better than LR when trained on larger datasets. Thus, LR outperforms the BiLSTM-based deep model on the Twitter and balanced combined dataset as they have the smallest size.

The second experiment involved using the BERT model on the dataset to assess its performance on the white supremacist hate speech classification task. As shown in Table 7, BERT outperforms all the distributional-based embeddings (Google Word2Vec, GloVe and WSW2V) with the BiLSTM-based deep model in Table 5. This show that the BERT model gives a closer meaningful vector of the words due to its training strategy (deeply bidirectional) and the large corpus trained

TABLE 6. The results of the domain-specific and domain agnostic word embedding with BiLSTM in comparison with [28] on a balanced sample of the Stormfront dataset (2000 sentences).

Methods	Word Embedding	Accuracy (Hate)	Accuracy (non-Hate)	Accuracy	F1-score
de Gibert et al. [28]	Random embedding	0.760	0.800	0.780	-
BiLSTM	Word2Vec	0.824	0.685	0.760	0.787
	GloVe.Wikipedia	0.790	0.780	0.785	0.791
	GloVe.Twitter	0.795	0.775	0.785	0.789
	WSW2V	0.805	0.800	**0.802**	0.792

TABLE 7. BERT models (Base-Large) results in white supremacist classification on the balanced dataset (Twitter white supremacist, Stromfront datasets, and a balanced combination of both).

Methods	Dataset	Precision	Recall	F1-score (hate)	F1-score (non-hate)	F1-score	AUC
BERT Base	Stormfront dataset	0.91	0.92	0.583	0.954	0.912	0.740
	Twitter dataset	0.67	0.67	0.700	0.637	**0.669**	0.669
	Balanced combined	0.8	0.8	0.795	0.809	**0.802**	0.802
BERT Large	Stormfront dataset	0.92	0.93	0.625	0.958	**0.924**	0.784
	Twitter dataset	0.66	0.67	0.715	0.598	0.664	0.655
	Balanced combined	0.80	0.80	0.803	0.799	0.801	0.801

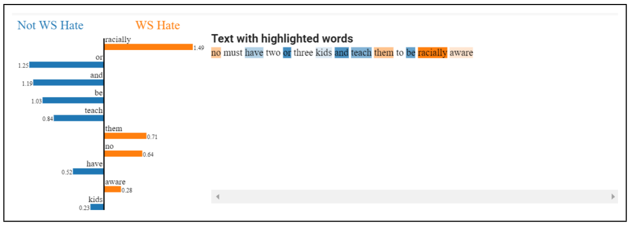
TABLE 8. BERT (Base-Large) model results for white supremacist classification in comparison with [28] study. The evaluation performed on a balanced sampled from Stormfront dataset (2000 sentences).

Source	Methods	Accuracy (Hate)	Accuracy (non- Hate)	Accuracy (All)	F1-score
de Gibert et al. [28]	Random embedding, LSTM	0.76	0.80	0.780	-
Our experiments	BERT BASE	0.82	0.86	**0.840**	0.840
	BERT LARGE	0.80	0.75	0.775	0.777

on. The BERT language model combines the advantages of domain-agnostic and domain-specific embeddings in its training strategy, it is pre-trained on a large corpus and add extra layer for training your specific task.

To better understand the obtained results, we applied LIME which stands for Local Interpretable Model-agnostic Explanations. LIME is an interpretable technique that provides a qualitative representational understanding that reflects the contribution of each feature to a specific label of the data sample. It also shows how changing the features influences the prediction. We presented different situations of sentence predictions for both labels, white supremacist and not white supremacist hate speech. We, in particular, focus on three examples from the testing dataset True Positive (TP), True negative (TN), and False Postive (FP), and False Negative (FN).

Figure 4 shows a bar chart with highlighted texts indicating the importance of the most relevant words. Color denotes which class the word contributes to (blue for "Not WS Hate," orange for "WS Hate"). The words of the document (TN) such as "teach," "have," and "kids" contribute to making the sentence be labeled as not white supremacist tweet. While words such as "racially" and "them" contribute to making the sentence to be classified as white supremacist hate speech.



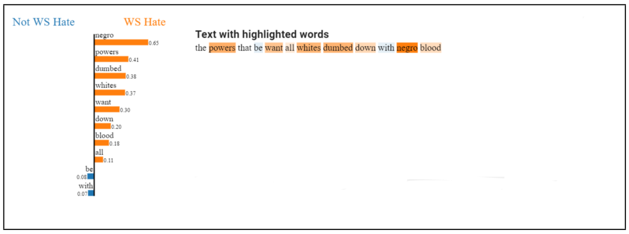
Text with highlighted words

no must have two or three kids and teach them to be racially aware

FIGURE 4. Result of applying LIME on document actual label = 0 and predicted = 0.

While the sentence (TP) shown in Figure 5 provides an example of white supremacist hate speech and most significant words such as "whites", "dumped", "negros" that contributed to classifying this sentence as white supremacy hate speech.

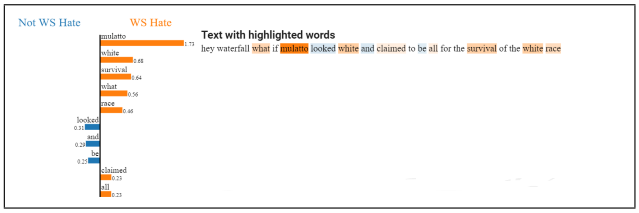
We also analyzed the misclassified examples as shown in Figure 6 and Figure 7, to examine which words contribute in misclassifying that example. It is obvious in Figure 6 that the classifier was more accurate than human annotator. As it is understated from the sentence intent and how the writer expresses hate in an implicit way. Same thing in Figure 7 which shows how this sentence was not hate as the annotators decided the sentence lacks any hate intent.



Text with highlighted words

the powers that be want all whites dumbed down with negro blood

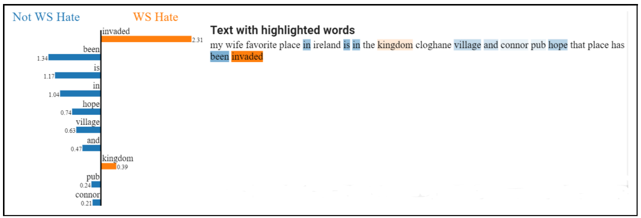
FIGURE 5. Result of applying LIME on a document with actual label = 1 and predicted = 1.



Text with highlighted words

hey waterfall what if mulatto looked white and claimed to be all for the survival of the white race

FIGURE 6. Result of applying LIME on a document with actual label = 0 and predicted = 1.



Text with highlighted words

my wife favorite place in ireland is in the kingdom cloghane village and connor pub hope that place has been invaded

FIGURE 7. Result of applying LIME on a document with actual label = 1 and predicted = 0.

VIII. CONCLUSION AND FUTURE WORK

In this study, we examined domain specific and agnostic word embedding with deep learning (BiLSTM). The results show that this approach performs well for the problem of white supremacist hate speech. BERT model has also proved that it provides the state of the art for this problem. The experiment results show that BERT outperforms domain specific approach with 4 point, however, domain specific approach is able to detect intentionally misspellings and common slang from hate community while BERT model fails to detect as it is trained on Wikipedia and books. Some of the datasets in the experiments are imbalanced to simulate real-world data, and others are balanced to assess the model's performance under an ideal situation. For future work, the corpus size will be maximized in order to generate more meaningful hate word embeddings, and experiments will be done on multiclass problems instead of binary class problems.

ACKNOWLEDGMENT

The authors would like to thank all the researchers who have made their resources available to the research community.

REFERENCES

[1] A. Weber, *Manual on Hate Speech*. Strasbourg, France: du Conseil de lEurope, 2009.

[2] W. A. S. Inc. (Nov. 2020). *Digital 2020*. [Online]. Available: https://wearesocial.com/digital-2020

[3] A.-D. League. (2020). *Online Hate and Harassment: The American Experience*. [Online]. Available: https://www.adl.org/onlineharassment

[4] M. Duggan, *Online Harassment 2017*. Washington, DC, USA: Pew Research Center, 2017.

[5] J. Clement. (Oct. 2019). *U.S. Teens Hate Speech Social Media by Type 2018 L Statistic*. [Online]. Available: https://www.statista.com/statistics/945392/teenagers-who-encounter-hate%-speech-online-social-media-usa/

[6] R. Blazak, "Toward a working definition of hate groups," *Hate Crimes*, vol. 3, no. 1, pp. 133–162, 2009.

[7] M. L. Williams, P. Burnap, A. Javed, H. Liu, and S. Ozalp, "Hate in the machine: Anti-black and anti-muslim social media posts as predictors of offline racially and religiously aggravated crime," *Brit. J. Criminol.*, vol. 60, pp. 93–117, Jul. 2019.

[8] W. Cai and S. Landon. (Apr. 2019). *Attacks by White Extremists are Growing. So are Their Connections*. [Online]. Available: https://www.nytimes.com/interactive/2019/04/03/world/white-extremist-te%rrorism-christchurch.html

[9] B. News. (2020). *Norway Mosque Shooting: Man Opens Fire on Al-Noor Islamic Centre*. [Online]. Available: https://edition.cnn.com/2019/08/09/us/el-paso-shooting-friday/index.htm%l

[10] K. Cohen, F. Johansson, L. Kaati, and J. C. Mork, "Detecting linguistic markers for radical violence in social media," *Terrorism Political Violence*, vol. 26, no. 1, pp. 246–256, Jan. 2014.

[11] M. Dadvar, D. Trieschnigg, R. Ordelman, and F. de Jong, "Improving cyberbullying detection with user context," in *Advances in Information Retrieval*, P. Serdyukov, P. Braslavski, S. O. Kuznetsov, J. Kamps, S. Rüger, E. Agichtein, I. Segalovich, and E. Yilmaz, Eds. Berlin, Germany: Springer, 2013, pp. 693–696, doi: 10.1007/978-3-642-36973-5_62.

[12] Q. Huang, V. K. Singh, and P. K. Atrey, "Cyber bullying detection using social and textual analysis," in *Proc. 3rd Int. Workshop Socially-Aware Multimedia (SAM)*, 2014, pp. 3–6.

[13] B. Haidar, M. Chamoun, and F. Yamout, "Cyberbullying detection: A survey on multilingual techniques," in *Proc. Eur. Model. Symp. (EMS)*, Nov. 2016, pp. 165–171.

[14] Y. Chen, Y. Zhou, S. Zhu, and H. Xu, "Detecting offensive language in social media to protect adolescent online safety," in *Proc. Int. Conf. Privacy, Secur., Risk Trust Int. Conf. Social Comput.*, Sep. 2012, pp. 71–80.

[15] G. K. Pitsilis, H. Ramampiaro, and H. Langseth, "Detecting offensive language in tweets using deep learning," *Appl. Intell.*, vol. 48, no. 12, pp. 4730–4742, Jul. 2018, doi: 10.1007/s10489-018-1242-y.

[16] M. H. Ribeiro, P. H. Calais, Y. A. Santos, V. A. Almeida, and W. Meira, Jr., "'Like sheep among wolves': Characterizing hateful users on Twitter," 2017, *arXiv:1801.00317*. [Online]. Available: https://arxiv.org/abs/1801.00317

[17] M. H. Ribeiro, P. H. Calais, Y. A. Santos, V. A. F. Almeida, and W. Meira, Jr., "Characterizing and detecting hateful users on Twitter," 2018, *arXiv:1803.08977*. [Online]. Available: https://arxiv.org/abs/1803.08977 and https://ui.adsabs.harvard.edu/abs/2018arXiv180308977H

[18] N. Djuric, J. Zhou, R. Morris, M. Grbovic, V. Radosavljevic, and N. Bhamidipati, "Hate speech detection with comment embeddings," in *Proc. 24th Int. Conf. World Wide Web*, May 2015, pp. 29–30.

[19] N. Albadi, M. Kurdi, and S. Mishra, "Are they our brothers? Analysis and detection of religious hate speech in the Arabic twittersphere," in *Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining (ASONAM)*, Aug. 2018, pp. 69–76.

[20] Z. Zhang, D. Robinson, and J. Tepper, "Detecting hate speech on Twitter using a convolution-GRU based deep neural network," in *The Semantic Web*, A. Gangemi, R. Navigli, M.-E. Vidal, P. Hitzler, R. Troncy, L. Hollink, A. Tordai, and M. Alam, Eds. Cham, Switzerland: Springer, 2018, pp. 745–760, doi: 10.1007/978-3-319-93417-4_48.

[21] T. De Smedt, G. De Pauw, and P. Van Ostaeyen, "Automatic detection of online jihadist hate speech," 2018, *arXiv:1803.04596*. [Online]. Available: https://arxiv.org/abs/1803.04596

[22] E. Ferrara, W.-Q. Wang, O. Varol, A. Flammini, and A. Galstyan, "Predicting online extremism, content adopters, and interaction reciprocity," in *Social Informatics*, E. Spiro and Y.-Y. Ahn, Eds. Cham, Switzerland: Springer, 2016, pp. 22–39, doi: 10.1007/978-3-319-47874-6_3.

[23] Y. Wei, L. Singh, and S. Martin, "Identification of extremism on Twitter," in *Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining (ASONAM)*, Aug. 2016, pp. 1251–1255.

[24] I. Gialampoukidis, G. Kalpakis, T. Tsikrika, S. Papadopoulos, S. Vrochidis, and I. Kompatsiaris, "Detection of terrorism-related Twitter communities using centrality scores," in *Proc. 2nd Int. Workshop Multimedia Forensics Secur.*, Jun. 2017, pp. 21–25.

[25] G. K. Pitsilis, H. Ramampiaro, and H. Langseth, "Effective hate-speech detection in Twitter data using recurrent neural networks," *Appl. Intell.*, vol. 48, no. 12, pp. 4730–4742, 2018.

[26] P. Badjatiya, S. Gupta, M. Gupta, and V. Varma, "Deep learning for hate speech detection in tweets," in *Proc. 26th Int. Conf. World Wide Web Companion (WWW) Companion*, 2017, pp. 759–760.

[27] B. Gamb´ack and U. K. Sikdar, "Using convolutional neural networks to classify hate-speech," in *Proc. 1st Workshop Abusive Lang. Online*, 2017, pp. 85–90.

[28] O. de Gibert, N. Perez, A. García-Pablos, and M. Cuadros, "Hate speech dataset from a white supremacy forum," in *Proc. 2nd Workshop Abusive Lang. Online (ALW)*, 2018, pp. 11–20.

[29] A. Brindle, *The Language of Hate: A Corpus Lingusitic Analysis of White Supremacist Language*. Evanston, IL, USA: Routledge, 2016.

[30] A. Liu, "Neural network models for hate speech classification in tweets," Ph.D. dissertation, Dept. Arts Sci., Harvard, Cambridge, MA, USA, 2018.

[31] S. Gupta and Z. Waseem, "A comparative study of embeddings methods for hate speech detection from tweets," Tech. Rep., 2017.

[32] C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad, and Y. Chang, "Abusive language detection in online user content," in *Proc. 25th Int. Conf. World Wide Web*, 2016, pp. 145–153.

[33] M. Hartung, R. Klinger, F. Schmidtke, and L. Vogel, "Identifying right-wing extremism in German Twitter profiles: A classification approach," in *Proc. Int. Conf. Appl. Natural Lang. Inf. Syst.* Springer, 2017, pp. 320–325.

[34] M. Hartung, R. Klinger, F. Schmidtke, and L. Vogel, "Ranking right-wing extremist social media profiles by similarity to democratic and extremist groups," in *Proc. 8th Workshop Comput. Approaches to Subjectivity, Sentiment Social Media Anal.*, 2017, pp. 24–33.

[35] M. A. Hearst, S. T. Dumais, E. Osman, J. Platt, and B. Scholkopf, "Support vector machines," *IEEE Intell. Syst. Appl.*, vol. 13, no. 4, pp. 18–28, Jul./Aug. 2008.

[36] Y. Kim, "Convolutional neural networks for sentence classification," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014.

[37] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[38] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 2013.

[39] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1532–1543.

[40] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2018.

[41] Y. Bengio, R. Ducharme, P. Vincent, and C. Janvin, "A neural probabilistic language model," *J. Mach. Learn. Res.*, vol. 3, pp. 1137–1155, Feb. 2003.

[42] J. Lilleberg, Y. Zhu, and Y. Zhang, "Support vector machines and Word2vec for text classification with semantic features," in *Proc. IEEE 14th Int. Conf. Cognit. Informat. Cognit. Comput. (ICCI*CC)*, Jul. 2015, pp. 136–140.

[43] M. Ailem, A. Salah, and M. Nadif, "Non-negative matrix factorization meets word embedding," in *Proc. 40th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Aug. 2017, pp. 1081–1084.

[44] P. Wang, Y. Qian, F. K. Soong, L. He, and H. Zhao, "Part-of-speech tagging with bidirectional long short-term memory recurrent neural network," 2015.

[45] S. K. Sienčnik, "Adapting word2vec to named entity recognition," in *Proc. 20th Nordic Conf. Comput. Linguistics (NODALIDA 2015)*, 2015, pp. 239–243.

[46] D. Tang, F. Wei, N. Yang, M. Zhou, T. Liu, and B. Qin, "Learning sentiment-specific word embedding for Twitter sentiment classification," in *Proc. 52nd Annu. Meeting Assoc. Comput. Linguistics (Long Papers)*, vol. 1, 2014, pp. 1555–1565.

[47] J. Wang, L.-C. Yu, K. R. Lai, and X. Zhang, "Dimensional sentiment analysis using a regional CNN-LSTM model," in *Proc. 54th Annu. Meeting Assoc. for Comput. Linguistics (Short Papers)*, vol. 2, 2016, pp. 225–230.

[48] S. Al-Azani and E.-S. M. El-Alfy, "Using word embedding and ensemble learning for highly imbalanced data sentiment analysis in short Arabic text," in *Proc. ANT/SEIT*, 2017, pp. 359–366.

[49] Google. (2019). *word2vec*. [Online]. Available: https://drive.google.com/file/d/0B7XkCwpI5KDYNlNUTTlSS21pQmM/edit

[50] J. Devlin and M.-W. Chang. (2018). *Open Sourcing Bert: State-of-the-Art Pre-Training for Natural Language Processing*. Google AI Blog. Weblog.From. Accessed: Dec. 4, 2019. [Online]. Available: https://ai.googleblog.com/2018/11/open-sourcing-bertstate-of-art-pre.html

[51] F. Chollet, *Deep Learning Mit Python Und Keras: Das Praxis-Handbuch vom Entwickler der Keras- Bibliothek*. Bonn, Germany: MITP-Verlags GmbH, 2018.

[52] A. Graves, S. Fernández, and J. Schmidhuber, "Bidirectional lstm networks for improved phoneme classification and recognition," in *Artificial Neural Networks: Formal Models and Their Applications—ICANN*, W. Duch, J. Kacprzyk, E. Oja, and S. Zadrożny, Eds. Berlin, Germany: Springer, 2005, pp. 799–804.

[53] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, and S. Ghemawat. (2020). *Tensorflow: Large-Scale Machine Learning on Heterogeneous Systems*. [Online]. Available: https://www.tensorflow.org/api_docs/python/tf/compat/v1/keras/layers/Cu%DNNLST

[54] N. Developer. (2020). *Nvidia Cudnn*. [Online]. Available: https://developer.nvidia.com/cudnn

[55] H. Inc. (2020). *Bert*. [Online]. Available: https://huggingface.co/transformers/model_doc/bert.html#tfbertforsequen%ceclassification

[56] *Amazon Mechanical Turk*. Accesed: Oct. 6, 2019. [Online]. Available: https://www.mturk.com/

[57] A. Schmidt and M. Wiegand, "A survey on hate speech detection using natural language processing," in *Proc. 5th Int. Workshop Natural Lang. Process. Social Media*, 2017, pp. 1–10.

HIND S. ALATAWI received the B.S. degree in computer science from the University of Tabuk, Tabuk, Saudi Arabia, in 2015, and the master's degree in computer science from King Abdulaziz University, Jeddah, Saudi Arabia, in 2020. Her research interests include artificial intelligence, machine learning, natural language processing, and big data analysis.

AREEJ M. ALHOTHALI received the B.S. degree in computer science from King Abdulaziz University, Jeddah, Saudi Arabia, in 2003, and the Master of Mathematics (MMATH) and Ph.D. degrees in computer science (artificial intelligence) from the University of Waterloo, Waterloo, ON, Canada, in 2011 and 2017, respectively. She joined the Department of Computer Science, Faculty of Computer Science and Information Technology, King Abdulaziz University, as an Assistant Professor, in 2017. Her current research interests include artificial intelligence, machine learning, natural language processing, computer vision, affective computing, and sentiment analysis.

KAWTHAR M. MORIA received the B.Sc. (Hons.) and M.Sc. (Hons.) degrees in computer science from King Abdul Aziz University, and the Ph.D. degree in image and video analysis from the University of Victoria, Canada. She held a number of faculty positions at the Effat College, as a Teaching Assistant, a Lecturer, and an Assistant Professor with the Faculty of Computer Science and Information Systems, King Abdul Aziz University. She is currently acting as a Chair/Supervisor with the Electrical and Computer Engineering Department, King Abdul Aziz University. Her research interests include artificial intelligence related topics, computer vision, and medical image and video analysis.