DETECTION OF HATE SPEECH IN SOCIAL NETWORKS: A SURVEY ON MULTILINGUAL CORPUS

Areej Al-Hassan[1] and Hmood Al-Dossari[2]
[1]Department of Information Systems, King Saud University, Riyadh, Saudi Arabia

[2]Department of Information Systems, King Saud University, Riyadh, Saudi Arabia

ABSTRACT

In social media platforms, hate speech can be a reason of "cyber conflict" which can affect social life in both of individual-level and country-level. Hateful and antagonistic content propagated via social networks has the potential to cause harm and suffering on an individual basis and lead to social tension and disorder beyond cyber space. However, social networks cannot control all the content that users post. For this reason, there is a demand for automatic detection of hate speech. This demand particularly raises when the content is written in complex languages (e.g. Arabic). Arabic text is known with its challenges, complexity and scarcity of its resources. This paper will present a background on hate speech and its related detection approaches. In addition, the recent contributions on hate speech and its related anti-social behaviour topics will be reviewed. Finally, challenges and recommendations for the Arabic hate speech detection problem will be presented.

KEYWORDS

Text Mining, Social Networks, Hate Speech, Natural Language Processing, Arabic NLP

1. INTRODUCTION

Over the last decades, people are getting more engaged with the wide spread of social networks. Microblogging applications opened up the chance for people around the globe to express and share their thoughts extensively and in a real-time manner. Such expressions afford researchers with the ability to investigate the online social emotions in different events. people now have the potential to speak freely, this allowed them to exchange all sorts of thoughts, emotions and knowledge. However, cyberspace is not always safe, it can be a reason for the dissemination of aggressive and harmful content. Hate speech is an online common form for expressing prejudice and aggression. This may convey racist, xenophobic and many forms of verbal aggression. Hate speech is typically defined as the act that disparages a person or people on the basis of a number of characteristics that may include and not limited to: race, ethnicity, sexual orientation, gender, religion and nationality [1]. In social media platforms, there are uncontrollable number of comments and posts issued every second which make it impossible to trace or control the content of such platform. Therefore, social platforms are facing a problem in limiting these posts while balancing the freedom of speech[2]. In addition, the diversity of people and their backgrounds, cultures and believes can ignite the flame of hate speech [3]. In the other hand, each culture has its own different interpretations and characteristics of cyber-hate. So, every culture is assumed to act differently and have their own way of intervention in a manner which best suits the culture.

Dhinaharan Nagamalai et al. (Eds) : COSIT, AIAPP, DMA, SEC - 2019
pp. 83–100, 2019. © CS & IT-CSCP 2019 DOI: 10.5121/csit.2019.90208

For the Arab region, there is a noticeable growth in the usage of social media platforms. According to the Arab social media report [4], the social media penetration in the Arab region reached 90% of the population in some countries. this increase in usage and the openness in speech results in a public concern on existing practices in social networks. A vast amount of posts that is hard and even impossible to control manually by platform owners. Hate and aggression through social networks should be rationed and regulated by policy makers and should be also countered by harnessing the power of artificial intelligence and machine learning algorithms to automate the detection of hate speech in social media.

The rest of the paper is organized as follows: section 2 will include a theoretical background, section 3 will go through the works related to hate speech detection. Then a discussion and recommendations will be presented in section 4. Finally, we conclude this paper by highlighting the future research directions.

2. BACKGROUND

2.1. What is Hate Speech

The case of hate speech and violent communication conducted over the internet can be referred as cyber-hate [5]. It is a narrow and specific form of cyber-bullying and it can be defined as "any use of electronic communications technology to spread racist, religious, extremist or terrorist messages" it is different from cyber-bullying in that hate speech can target not only individuals but it also has implications on whole communities [1]. Brown [6] has also defined hate speech as any textual or verbal practice that implicates issues of discrimination or violence against people in regard to their race, ethnicity, nationality, religion, sexual orientation and gender identity. According to Anis [7] hate speech can occur in different linguistic styles and several acts like insulting, provocation, abusing and aggression. However, according to Chetty and Alathur [8], hate speech can be categorized into the following categories:

2.1.1. Gendered hate speech

This category includes Any form of hostility towards particular gender or any devaluation based on person's gender. This include any post that offense particular gender. Also it includes any form of misogyny. Moreover, Jha and Mamidi [9] clarify that sexism may come in two forms: Hostile (which is an explicit negative attitude) and Benevolent (which is more subtle).

2.1.2. Religious hate speech

This will include any kind of religious discrimination, such as: Islamic sects, calling for atheism, Anti-Christian and their respective denominations or anti-Hinduism and other religions. However, Albadi et al. [10] mentioned that religious hate speech is considered as a motive of crimes in countries with highest social crimes.

2.1.3. Racist hate speech

Lastly, this category includes is Any sort of racial offense or tribalism, regionalism, xenophobia (especially for migrant workers) and nativism (hostility against immigrants and refugees) and any prejudice against particular tribe or region. For instance, offending an individual because he belongs to a particular tribe or region or country or favoritism of a particular tribe. Add to that, offending the appearance and color of individual.

2.2. What Constitutes Hate Speech

Hate speech is hard to comprehend. However, it can be recognized based on specific characteristics that can be distinguished from one culture to another. These characteristics are debatable, some may interpret them as a pure hate and some don't. This problem is considered as a controversial problem that no one can agree upon. Gelashvili and Nowak [11] argued that it is an obstacle for social media platforms owners to regulate hate speech as many questions will raise to their heads such as what constitute hate speech? And what kind of hate speech need to be countered? Only legitimate people who are actively engaged in the same culture and who can be competent enough can give the answers to these questions. Some studies have given some necessary terminologies for studying hate speech, for example Fortuna and Nunes [12] have listed some of the main rules for hate speech identification. In brief, hate speech is identified when disparaging stereotype about group. Together with using racial and sexist slurs with intent to harm. Add to that when indecently speak about religion or specific country.

Correspondingly, when identifying hate speech, we need to exclude some conditions. For instance, when trying to explain the meaning of some abusive words or when we use some of racial terms in another context which has no hate undertone. Add to that when writing a news article and referring to a sect which is associated with hate crime "e.g. ISIS" this referral itself won't be considered as hate speech. In like manner, Waseem and Hovy [2] have proposed 11 parameters to distinguish hate speech specifically in twitter platform, some of which are: usage of sexiest and racial terms, attacking and criticizing minority, promoting violence, distorting the truth with lies and supporting suspicious hashtags.

Given these characteristics, a reasonable list can be derived for a particular culture with certain adjustments to deal with the controversy and then from that list, hate speech can be reliably identified and recognized. Anis [7] discussed the dominant themes in Arabic hate speech particularly in the newspaper and concluded that hate speech in Arab region is generally related to religion and sectarian themes.

2.3. Text Mining and NLP for Hate Speech Detection

The problem of hate speech in social networks is technically considered as unstructured text problem. Therefore, extracting insights and pattern from such text can be a bit challenging, owing to the context-dependent interpretation of natural language. Text mining technologies have the capabilities to handle the ambiguity and variability of unstructured data [13].

Natural Language Processing or (NLP) is the main pillar of text mining, it employs a number of computational tasks in order to make human natural language tractable and understood by the machine [14]. Today, NLP researchers have moved towards the rich and controversial data available in social networks by downloading vast amount of unstructured data, these data can be mined and put into practical use. Text mining for social networks requires a number of lexical, syntactic and semantic NLP tasks aiming to give a structure to the text for further processing. These tasks include: Tokenization, which splits the text into word tokens by the spaces. Also, in this task, stop words like "in", "the" will be taken out as they make no sense to the meaning. There are a number of available tokenization tools such as Apache "OpenNLP[1]" or "Stanford Tokenizer[2]". Then, predicting Part of Speech (PoS) for each token will take a place in aim to provide lexical information. Then, parsing will take a place by representing the syntactic structure of the whole text [15]. A significant drawback of NLP nowadays is that most of the tools are

[1] https://opennlp.apache.org
[2] https://nlp.stanford.edu/software/tokenizer.shtml

exclusively designed for common languages such as English, French, Spanish [14]. Comparatively, uncommon language such as Arabic has a challenge associated with the difficulty in adapting the common languages tools. However, Arabic linguistics experts have gone through a considerably good achievements in analyzing Arabic language morphology [16]. In particular, Khoja stemmer [17] and a stemmer by Ghwanmeh et al. [18] and finally, AlKhalil Morpho system [19] which is considered as the best Arabic Morphological system [16].

2.4. Arabic Text in Social Networks

Arabic language has a very high growth rate in means of usage in social networks. Based on the Arab Social Media Report [4] the average rate of using Arabic language in social media reaches 55% in 2017. As it can be seen, the amount of Arabic content in social networks is growing substantially in recent years. Facebook stands as the most popular platform in the Arab region, followed by Twitter, LinkedIn and Instagram, with penetration rates (34%, 13%, 6.75%,1.8%) respectively.

Arabic language is known by its difficulties and challenges. In case of twitter, Salem [4] stated that it is hard to extract meaningful insights from an Arabic tweet, basically because tweets are very noisy and people don't care about spilling and grammar in their posts. Secondly, they contain great amount of variances including: writing from right to left, combining Arabic with Latin words and the usage or the neglection of diacritics [20]. Not to mention the different local informal dialect for each Arab country, this issue can be considered as the major issue for Arabic language, especially when we are considering hate speech, some Arabic terms may imply hateful meaning in one region, while it is considered an ordinary term in others. Consequently, Salem [4] claimed that many researchers tend to work with specific Arab region and try to fine-tune the used algorithms to adapt this specific region aiming to increase the accuracy of their works.

2.5. Automatic Hate Speech Detection in Social Networks

One of the main applications of social media mining is the automatic detection of events and behaviors which includes identifying people behavior in real-world events through monitoring their interactions with each other. Researchers can take an advantage of these explosive data to reach substantial insights [21]. This task depends mainly on text mining approaches such as NLP and machine learning algorithms. In twitter, researchers explored many automatic detection tasks, such as: anti-social behaviour detection, spam detection, natural disasters (e.g. earthquakes), trends and public opinion events. To achieve this task, several features and common patterns need to be identified. Then, machine learning algorithms are applied to perform the classification task to get the targeted result out of the data.

2.5.1. Features representation for hate speech detection

To perform an automatic detection task such as hate speech detection general features of the corpus need to be specified in order to enable the classification algorithms to perform the task. Some of these approaches will be presented.

Dictionaries and Lexicons. This feature usually employed in unsupervised machine learning scenarios [22]. Wiegand et al. [23] addressed the detection of profane words by taking advantage of the corpora and lexical resources. They used several features and general-purpose lexical resource to build their lexicon. Usually lexicon-based approaches are not competitive to other features used in supervised approaches since they are domain independent. Gitari et al. [24] also used a lexicon as a primary feature by aggregating opinions and giving rates to the subjective words.

Bag-of-words (BOW) and N-grams. It can be considered as word co-occurrence feature. A vectorization process is performed on tokenized words in the corpus by assigning weight for each word according to its frequency in the tweet and its frequency in between different tweets, the vectorization process is done using some statistical models (e.g. TF-IDF weight). After that, a list of words together is called BOW which will be presented as vectors of weights [25]. N-gram representation means a sequences of *N* adjacent words. Waseem and Hovy [2] analyzed the impact of using number of features in conjunction with character N-gram for detecting hate speech. They found that using character n-gram representation is a great option for detecting hate speech. BOW is limited by its need to be accompanied with other features to improve the performance, but in the other hand it is computationally expensive [26]. For *N*-grams, it needs careful selection for the value of *N* to avoid high level of distance between related words [27].

Latent Dirichlet Allocation (LDA). It is a probabilistic topic modeling method. It is mainly used to give an estimation of the latent topics in data set and these latent topics will be used as features instead of words. However, LDA is suitable for unsupervised and semi-supervised machine learning settings. Xiang et al. [28] claimed that BOW did not work well for abusive text detection in twitter. Instead, they include highly expressive topical feature and other lexicon features by using LDA model. this approach can be an alternative for that supervised methods.

Word embedding and Word2Vec. The emergence of word embedding mitigated the data sparsity problem by bringing up an extra semantic feature by generating distributed representations that introduces dependence between words. Word2Vec is one of the techniques to construct word embedding. According to Lilleberg et al. [29] word2vec has given a lot of interest by researchers in text mining field and it is compatible with both supervised and unsupervised machine learning models.

2.5.2. Machine learning for hate speech detection

After preparing the text to work with machine, classification algorithms can take a place to perform the detection task. In terms of classifiers, machine learning approaches can be categorized into: supervised, semi-supervised and unsupervised approaches.

Supervised learning. This approach is domain dependent since it relies on a manual labeling of a large volume of text. Labeling task is time and effort consuming but it is more efficient for domain-dependent events. Most of the approaches used for hate speech detection tasks are supervised methods. For instance, Burnap and Williams [30] have used several supervised classifiers to detect hate speech in twitter, their results showed that all classifiers have performed the same but the different settings of features changed the accuracy of the model. Consequently, the choice of the classifier depends on the features that can be extracted from the corpus.

Semi-supervised learning. In this paradigm, algorithms are trained using both of labeled and unlabeled data. Using labeled data in conjunction with unlabeled data can effectively enhance the performance, this can be seen in Hua et al. [31] model. They argued that unsupervised learning has limited ability to handle small scale events. On the contrary, supervised learning has the capability to effectively capture small scale events but the need to manually label the data set decreases the scalability of the model. To achieve the right balance between these two situations authors suggested a semi-supervised approach. Moreover, Xiang et al. [28] replaced the costly manual annotation with an automatically generated feature, They claimed that their approach can be a good alternative to the costly supervised approaches to detect hate speech.

Unsupervised learning. It is a domain-independent approach and is capable to handle a diversity of content while maintaining scalability [32]. It does not rely on human labor to label a large volume training set, instead, it dynamically extracts domain-related key terms. Gitari et al. [24] utilized a bootstrapping approach to build their lexicon by starting with small seed of hate verb and then expanded it iteratively. The best results from their model were obtained when they incorporated semantic hate and them-based features.

2.5.3. Deep learning

Deep learning models show promising future in text mining tasks. It depends entirely on the artificial neural networks but with extra depth. It tries to mimic the event in layers of neurons and attempt to learn in a real sense to identify patterns in the provided text. However, deep learning approaches are not always better than the traditional supervised approaches. The performance of deep learning is subject to the right choice of algorithm and number of hidden layers as well as the feature representation technique. Al-Smadi et al. [33] proved the previous assumption by comparing the performance of both Recurrent Neural Network (RNN) and Support Vector Machine (SVM). Their comparison showed that SVM outperformed RNN for specific set of features. So, they suggested to use (LSTM) and different algorithm for the embedding for their future work. For hate speech detection, Pitsilis et al. [34] used RNN model with word frequency vectorization to implement the features instead of the word embedding to break the barrier of language dependency in word embedding approach. Their results outperformed the current state of art deep learning approaches for hate speech detection.

3. RELATED WORK

This section presents a comprehensive review on the key works and existing studies related to the area of automatic detection and hate speech in particular.

3.1. Current state in Hate Speech Detection and Related Concepts

There are some researches that have discussed different related terminologies which serves similar related concept to the phenomena of hate speech (e.g. cyber-bullying, abusive language, radicalization detection). The analysis of these different terminologies will definitely help to reach insights from different perspectives in current situation and will also contribute in spotting and recognizing the interrelationship among these terminologies.

3.1.1. Abusive language detection

It is the general concept that covers all the hurtful language. Hate speech is considered under the umbrella of abusive language. This terminology also covers profanity (the use of inappropriate words). However, many researches refer to abusive language as offensive language. Chen et al. [35] used YouTube comments as a dataset to detect offensive language. They used a combination of lexical and syntactic features and they incorporated user's writing style to predict user's behaviour in the future. Also, Wiegand et al. assumed that they can filter abusive words from the negative polar expressions. They took advantage of a base lexicon by taking a small subset of negative polar expressions and then via crowdsourcing, the abusive words were labelled. Similar approach was proposed by Xiang et al. [28] to detect offensive content in twitter. Their features were mainly based on the linguistic regularities of the profane terms also based on statistical topic modelling on a relatively big dataset. For a deep learning scenario, Park and Fung [36] compared the performance of one-step and two-step classifiers by using the dataset provided by Waseem and Hovy [2]. Based on their results, they believe that combining 2 classifiers (e.g. CNN and logistic regression) can boost up the performance. Moreover, Chen et al. [37] used FastText as

their neural network classifier to detect abusive text from various social networks platforms. They found that FastText performance is lower than using SVM as a classifier.

3.1.2. Cyberbullying detection

The electronic form of traditional bullying is called cyberbullying. which is the aggression and harassment that is targeted to an individual who is unable to defend himself [38]. Bullying is known with its repetitive act to the same individual, unlike hate speech which is more general and not necessarily intended to hurt a specific individual. Dinakar et al. [39] research is one of the pioneers and most cited researches for the textual cyberbullying detection. Their experiment was based on a corpus of 4500 YouTube comments. Their result showed that showing the polarities of the dataset outperformed categorizing the dataset into a multiclass. Both of Nahar et al. [40] and Capua et al. [41] presented unsupervised approach for cyberbullying detection. Özel et al. [42] work is unique to this area because they have investigated Turkish language in order to detect cyberbullying from twitter and Instagram text. Their results showed that Naïve Bayes Multinational showed the best results in both accuracy and total training and testing time. Finally, Pawar et al.[43] utilized distributed computing for cyberbullying detection. Their work focuses mostly on the robust performance rather than the accuracy alone.

3.1.3. Radicalization detection

This concept is usually referred to as a motive towards violent extremism. Usually radical groups have an ideology that considers violence as a legitimate action when it serves to address their concerns [44]. Radicalization and hate speech are closely related and usually mentioned as if they have the same meaning but actually radicalization comes under hate speech as it has specific tendencies towards religious believes. Wadhwa and Bhatia [45] referred to radical groups as "cyber-extremists". They investigated the possibility of the detection of such act in Twitter using unsupervised approach. They came with the conclusion that fully unsupervised approach will not be able to detect the right topics for this issue, manual intervention is necessary to reach better results because tweets have a dynamic nature. Agarwal and Ashish [46] introduced a semi-supervised approach to detect radicalization in twitter. They had a mixture of labeled and unlabeled data. They mainly counted on the hashtags with radical tendencies (e.g. #Terrorism) to identify extremism promoting tweets. Fernandez and Alani [47] believe that the major reason behind the inaccuracy of previous approaches that detect radicalization is because they mainly rely on the appearance of terminologies and expressions regardless of their context.

3.1.4. Hate speech detection

Starting from the early stages, Warner and Hirschberg [48] were one of the first initiatives to automate the detection of hate speech in the World Wide Web. Their research was specifically to detect anti-Semitic and their work included a number of challenges that should be overcome by now. The first racial oriented research was by Kwok and Wang [49] who decided to follow Warner and Hirschberg path and implemented a supervised model to detect racist tweets. Then, Burnap and Williams [50] were motivated to investigate the spread of hate speech immediately after Lee Rigby murder in UK. They trained a supervised classifier to find hateful and non-hateful tweets related to this particular event. Waseem and Hovy's work [2] was a baseline research for many following researchers as they have investigated the predictive features for hate speech detection. They allowed the access to their huge corpus of 16K tweets that is dedicated for hate speech researches in English language.

Hate speech in other languages was also investigated. Del Vigna et al. [51] designing a model to detect hate speech Italian language in Facebook. Their result showed that the classifiers

were not able to discriminate between three levels of hate. In addition, Alfina et al. [52] investigated the ability to detect hate in Indonesian language. They took advantage of a political event to collect their sample "Jakarta Governor election 2017". For the German hate speech Jaki [53] initiated a quick response to the recent German NetDG law by proposing a model to detect hate speech in German language. He also established a comprehensive qualitative and quantitative analysis in what constitute hate speech from the political communication perspective.

Recent works have shifted to the employment of deep learning for such task, Gambäck and Sikdar [54] applied deep learning approach on Waseem and Hovy's dataset [2]. Their results outperformed by means of precision and recall. The same corpus was also used by Badjatiya et al. [55] for comparing different combinations of deep learning models. In addition, Zhang et al. [56] explored combining convolutional and gated recurrent unit networks, they also compared their model performance with all previous deep learning models. They stated that their work sets a new benchmark for future researches in this area. Finally, Pitsilis et al. [34] believe that deep neural networks have a high potential to solve the issue of hate speech detection. Their deep learning approach outperformed all the state-of-art approaches.

3.2. Arabic Hate Speech Detection

A limited number of Arabic researches have contributed to that particular area. In the other hand, many Arabic researches were investigated in similar areas which we can call "Anti-social behaviors" such as, Abusive or offensive language and cyberbullying.

3.2.1. Arabic anti-social behaviour detection

Starting with Abusive language detection. Abozinadah paved the way in this area and contributed in three researches tailored for this area. First, Abozinadah et al. [57] proposed a model in response to Arab governments needs of blocking such abusive contents. They created their own test set and made it publicly available. Then in [58] Abozinadah and H. Jones, Jr. enhanced the previous work by proposing a lexicon that is fed by an Arabic word correction method to enhance the detection of such abusive words. A third work by Abozinadah is [59] which used statistical learning approach for the detection process to overcome the limitation in the BOW approach presented in other previous works. Mubarak et al. [60] work aimed to build a a large scale corpus of Arabic tweets that are classified to (Obscene, offensive and clean) and made it available for next researchers.

Another two contributions by Alakrot et al. [61][62]. In the first work, they have constructed a corpus of Arabic comments from YouTube and made it publicly available for abusive detection purposes. In their second work, an empirical examination of the dataset has been performed. They concluded that a combining N-gram and stemming may results in lower performance. Alshehri et al. [63] followed the same path of Abozinadah and Azalden but the concept is slightly different. They created a large scale of adult content in Arabic Twitter. Consequently, a large lexicon was built based on that corpus.

In addition to the previous behaviors, cyberbullying is another serious issue that has been addressed by many researchers in English language. For the Arabic language, Haidar et al. [64] made the first attempt to detect cyberbullying in Arabic language. Their work was the first step into this area, since it needs a lot of enhancements such as considering more features related to cyberbullying and choosing better feature representation. Alduailej and Khan [65] discussed the main challenges of detecting cyberbullying in Arabic language. The fundamental challenge was that we need to discover the context before deciding whether it is considered cyberbullying or not.

Finally, radicalization and extremism are another two anti-social behaviors that have been studied in Arabic language scenarios. Magdy et al. [66] classified twitter users whether they are supporting or opposing ISIS by discriminating the language that shows support for ISIS. Kaati et al. [67] proposed a model that detects whether a user is more likely to support Jihadist groups. Their experiment showed that AdaBoost classifier worked well for English tweets but it did not give the expected performance in Arabic tweets.

3.2.2. Arabic hate speech detection

In English language, hate speech detection has been intensively investigated by more than 14 contributors who investigated all the categories of hate speech (racial, sexism, religious and general hate). In contrast, Arabic language has limited available resources for detecting various categories of hate speech, actually, only one contribution has been found in this area which is specifically targeted to the detection of "Religious" Arabic hate speech. Albadi et al. [10] were the first to tackle the problem of religious hatred in Arabic twitter, but they didn't encounter the other categorizations of hate speech. They built and scored a lexicon of the most common religious terms. They tested various classifiers for this task including GRU RNN which outperformed the rest of classifiers. They stated their reasons behind choosing GRU rather than LSTM, they claimed that GRU works better with smaller datasets and it is faster with respect to training time, also it has lower probability to overfit small datasets.

3.3. Summery and Analysis

The next tables present a summary of all the discussed papers and they are organized according to their respective time series. These tables cover the following topics respectively: English Anti-social behaviours, English hate speech and finally, Arabic Anti-social behaviours. These tables can serve as a quick reference for all the key works done in the automatic detection in social media. All the approaches and their respective experiments results are listed in a concise manner.
Table 1 consolidates all the terminologies related to hate speech and their corresponding contributions. Table 2 summarizes all the multilingual contributions and papers which are directly related to hate speech. Finally, table 3 which gives an emphasize on the Arabic language by summarizing all the works that deals with the detection of anti-social behaviour in social media platforms. For the results column, the best results in each paper is pointed.

Table 1. Summary of the current state of anti-social behaviour detection, and their respective results, in the metric: Precision (P), Recall (R), F1-Score (F).

Author	Year	Platform	ML approach	Features Representation	Algorithm	P	R	F
Abusive Language (English)								
Chen et al. [35]	2012	YouTube	Un-Supervised	Lexical and syntactic	Match Rules	0.98	0.94	-
Xiang et al.[28]	2012	Twitter	Semi-Supervised	Topic modelling	Logistic Regression	-	-	0.84
Park, Fung [36]	2017	Twitter	Supervised	Character and Word2vec	Hybrid CNN	0.71	0.75	0.73
Chen et al. [37]	2017	Youtube, Myspace, SlashDot	Supervised	Word embeddings	FastText	-	0.76	-
Wiegand et al.[23]	2018	Twitter, Wikipedia, UseNet	Supervised	Lexical, linguistics and word embedding	SVM	0.82	0.80	0.81

Cyberbullying (English)								
Dinakar et al.[39]	2011	YouTube	Supervised	Tf-idf, lexicon, PoS tag, bigram	SVM	0.66	-	-
Nahar et al. [40]	2014	Myspace, Slashdot	Semi-Supervised	Linguistic features	Fuzzy SVM	0.69	0.82	0.44
Capua et al.[41]	2016	YouTube, Form-Spring, Twitter	Un-Supervised	Semantic and syntactic features	GHSOM network and K-mean	0.60	.094	0.74
Pawar et al.[43]	2018	Form-spring	Supervised	Bag of words	M-NB and Stochastic Gradient Descent	-	-	0.90
Cyberbullying (Turkish)								
Özel et al.[42]	2017	Twitter, Instagram	Supervised	Bag of words	M-Naïve Bayes	-	-	0.79
Radicalization (English)								
Wadhwa , Bhatia [45]	2013	Twitter	Un-Supervised	Topic identification, N-grams	Topic-entity mapping	-	-	-
Agarwal , Sureka [46]	2015	Twitter	Semi-Supervised	Linguistic, Term Frequency	LibSVM	-	-	0.83
Fernande z and Alani [47]	2018	Twitter	Supervised	Semantic Context	SVM	0.85	0.84	0.85

Table 2. Summary of the current state of hate speech and their respective results, in metrics: Precision (P), Recall (R), F1-Score (F).

Author	**Year-Platform**	**Classes**	**ML Approach**	**Features Representation**	**Algorithm**	**P**	**R**	**F**
Religious hate speech (English)								
Warner and Hirschberg [48]	2013-Yahoo news-group	Anti-Semitic, not anti-Semitic.	Supervised	Template-based, PoS tagging	SVM	0.59	0.68	0.63
Racial hate speech (English)								
Kwok and Wang[49]	2013-Twitter	Racist, Non-racist	Supervised	Unigram	Naïve Bayes	-	-	-
General hate speech (English)								
Burnap and Williams [50]	2014-Twitter	Yes, No	Supervised	BOW, Dependencies, Hateful Terms	Bayesian Logistic Regression	0.89	0.69	0.77
Gitari et al. [24]	2015-Blog	No hate, Weakly hate, Strongly hate	Semi-Supervised	Lexicon, Semantic, theme-based features	Rule based	0.73	0.68	0.70

Djuric et al. [68]	2015-Yahoo Finance	Hateful, Clean	Supervised	Paragraph2vec, CBOW	Logistic regression	-	-	-
Waseem and Hovy[2]	2016-Twitter	Hate, not hate	Supervised	Character n-grams	Logistic regression	0.72	0.77	0.73
Watanabe et al.[3]	2018-Twitter	Hateful, Offensive, Clean	Supervised	Sentiment-Based, Semantic, Unigram,	J48graft	0.79	0.78	0.78
Malmasi and Zampieri [69]	2018-Twitter	Hate, offensive, Ok	Supervised	N-grams, Skip-grams, hierarchical word clusters	RBF kernel SVM	0.78	0.80	0.79
Gambäck and Sikdar [54]	2017-Twitter	Non-hate, Racism, Sexism, Both	Supervised	Character N-grams, word2vec	CNN	0.85	0.72	0.78
Badjatiya et al. [55]	2017-Twitter	Sexist, Racist, Neither sexist nor racist	Supervised	Random Embedding,	LSTM and GBDT	0.93	0.93	0.93
Pitsilis et al. [34]	2018-Twitter	Neutral, Racism or Sexism	Supervised	Word-based frequency vectorization	RNN and LSTM	0.90	0.87	0.88
Zhang et al. [70]	2018-Twitter	Racism, Sexism, Both, Non-hate	Supervised	Word embeddings	CNN+GRU	-	-	0.94
General hate speech (Italian)								
Del Vigna et al. [51]	2017-Face-book	Hate, Not hate	Supervised	Morpho-syntactical, sentiment polarity, word embedding lexicons.	SVM	0.75	0.68	0.71
					RNN and LSTM	0.70	0.75	0.72
General hate speech (Indonesian)								
Alfina et al. [52]	2017-Twitter	Hate speech, Non-hate speech	Supervised	BOW and n-gram	Random Forest Decision Tree	-	-	0.93
General hate speech (German)								
Jaki. [53]	2018-Twitter	Muslim, Terrorist, Islamo-fascistoid	Un-Supervised	Skip grams and Character trigrams	K-means, single-layer averaged Perceptron	0.84	0.83	0.84

Table 3. Summary of the Arabic contributions in anti-social behaviour detection and their respective results, in the metrics: Precision (P), Recall (R), F1-Score (F).

Author	Year-Platform	Classes	ML Approach	Features Representation	Algorithm	P	R	F
Abusive language (Arabic)								
Abo-zinadah et al. [57]	2015-Twitter	Abuser, Normal	Supervised	Profile and tweet-based features, bag of words, N-gram, TF-IDF	Naïve Bayes	0.85	0.85	0.85
Abo-zinadah and H. Jones, Jr. [58]	2016-Twitter	Abusive, Legitimate Accounts	Un-Supervised	Lexicon, bag of words (BOW), N-gram	SVM	0.96	0.96	0.96
Abo-zinadah and H. Jones, Jr.[59]	2017-Twitter	Non-Abusive, Abusive	Supervised	PageRank (PR) algorithm, Semantic Orientation (SO) algorithm, statistical measures.	SVM	0.96	0.96	0.96
Mubarak et al.[60]	2017-Twitter, Arabic News Site	Obscene, Offensive and Clean	Un-supervised	unigram and bigram, Log Odds Ratio (LOR), Seed Words lists	None. Just performed extrinsic evaluation	0.98	0.45	0.60
Alakrot et al. [62][61]	2018-YouTube	Offensive, In-offensive	Supervised	N-gram	SVM	0.88	0.80	0.82
Violent content (Arabic)								
Abdelfatah et al. [71]	2017-Twitter	Violent, Non-violent	Un-supervised	Sparse Gaussian process latent variable model, morphological features, Vector Space Model	K-means clustering	0.56	0.60	0.58
Adult content (Arabic)								
Alshehri et al.[63]	2018-Twitter	Adult, Regular user	Supervised	Lexicon, N-grams, bag-of-means (BOM)	SVM	0.70	0.93	0.78
Cyberbullying (Arabic)								
Haidar et al.[64]	2017-Facebook, Twitter	Yes, No	Supervised	Tweet to SentiStrength Feature Vector	SVM	0.93	0.94	0.92
Terrorism (Arabic)								
Magdy et al. [66]	2016-Twitter	Pro-ISIS and Anti-ISIS	Supervised	Temporal patterns, Hashtags	SVM	0.87	0.87	0.87
Kaati et al. [67]	2016-Twitter	Support or Oppose Jihadism	Semi-Supervised	Data dependent features and data independent features.	AdaBoost	0.56	0.86	0.86

Religious hate speech (Arabic)								
Albadi et al.[10]	2018-Twitter	Hate, Not hate	Supervised	Word embeddings (AraVec)	GRU-based RNN	0.76	0.78	0.77

4. DISCUSSION AND FUTURE WORK

After exploring the literature, Arabic hate speech detection challenges can be pointed out based on what have been discussed in previous works.

4.1. Arabic Hate Speech Detection Challenges

Hate speech detection is not a simple keyword spotting, it is a complex task with many challenges. Based on the review conducted in the previous section, we can spot several research challenges in the automated detection of Arabic hate in social media.

First barrier is that there are a few of researches in hate speech detection that can result in high precision and recall rates. These few researches are mostly dedicated for languages with Latin characters, on the other hand, there is a gap in the Arabic language researches in this area. Many researchers confessed that Arabic language is the major challenge due to its complexity and richness in both of its derivations and inflections, add to that, the varieties of dialects used by Arab users in twitter. Secondly, Arabic hate speech detection is a multidisciplinary problem and it needs to be investigated from different dimensions, one of the challenges is related to the social and political perspective, will we be able to discriminate different hate speech contexts for different Arab cultures? As there is no unified definition for what constitutes hate speech. Coupled with the issue of legitimacy, hate speech contains a broad and loose range of expressions and sometimes, trivial issues can be included and considered as part of it which makes it hard to discriminate which case is more critical.

Moving to the technical perspective, choosing the most appropriate machine learning approach is another challenging decision. Previous works employed mostly all the varieties of techniques. According to tables 1,2,3, majority of researchers relied on supervised machine learning approaches in their automatic detection task. Un-supervised approaches come to the next place of popularity and semi-supervised approaches are the least used techniques. We need to consider all the factors that can affect our decision in the right choice of the approach. For instance, one major factor is the size of the corpus, as some ML algorithms works pretty well with small datasets. Others such as Neural Networks needs more intensive and complex training.

Resent researches are oriented towards deep learning to solve complex learning tasks. Researchers claimed that deep learning is powerful when it comes to finding data representation for classification and obviously it has a promising future in the field of the automatic detection. Choosing to adopt deep learning needs commitment in both of preparing and training the model with large amount of data. Generally, there are two main architectures for deep neural networks that are usually utilized for NLP tasks, these models are: RNN and CNN. In the previous tables, there were 4 hate speech researches that adopted deep learning, two of them were RNN and the two others were CNN. These researches concluded with the effectiveness of both approaches. for that reason, more investigation needs to be done to make the appropriate choice of deep learning architecture.

4.2 Machine Learning Model

When working with a specific language (e.g. Arabic) and particular region, this task can be considered as domain-dependent task. Consequently, supervised approaches are the best candidates for this task. However, Since the revolution of deep learning and deep neural networks for NLP tasks, we will consider narrowing our choice to these robust models. Yin et al. [72] conducted a comparative study between the two deep neural networks "RNN and CNN" as they are the most commonly used deep learning models for NLP tasks. Basically, RNN has two types: GRU and LSTM and it supports sequential architectures. CNN in the other hand has a hierarchical architecture. Yin experiments showed that RNN are well suited for the long-ranged context dependencies. While CNN is better in extracting local features. GRU and CNN results can be compared with respect to text size, GRU is better when the sentences are bit longer. Finally, they concluded that deep neural network performance is highly dependable on tuning the hyperparameters.

4.3 Conclusion and Future Work

Arab regions and worldwide are now more aware of the problem of spreading hate through the social networks. Many countries are working hard in regulating and countering such speech. This attention raised the need for automating the detection of hate speech. In this paper we analyzed the concept of hate speech and specifically "cyber hate" which is conducted in the means of social media and the internet sphere. Moreover, we differentiated between the different anti-social behaviors which include (Cyberbullying, Abusive and offensive language, Radicalization and hate speech). After that we presented a comprehensive study on how text mining can be used in social networks. we investigated some challenges which can be a guide for the implementation of Arabic hate speech detection model. In addition, these recommendations will help in drawing a road map and a blueprint for the future model. The future work will include incorporating the latest deep learning architectures to build a model that is capable to detect and classify Arabic hate speech in twitter into distinct classes. A data set will be collected from twitter, and for intensifying the training of our neural network we will including data from additional platform "e.g. Facebook" as it is the most used platform in the Arab region.

ACKNOWLEDGEMENT

The authors would like to thank Deanship of scientific research in King Saud University, for funding and supporting this research through the initiative of DSR Graduate Students Research Support (GSR).

REFERENCES

[1] C. Blaya, "Cyberhate: A review and content analysis of intervention strategies," Aggress. Violent Behav., no. May, pp. 0–1, 2018.

[2] Z. Waseem and D. Hovy, "Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter," Proc. NAACL Student Res. Work., pp. 88–93, 2016.

[3] H. Watanabe, M. Bouazizi, and T. Ohtsuki, "Hate Speech on Twitter: A Pragmatic Approach to Collect Hateful and Offensive Expressions and Perform Hate Speech Detection," IEEE Access, vol. 6, pp. 13825–13835, 2018.

[4] F. Salem, "Arab Social Media Report 2017: Social Media and the Internet of Things: Towards Data-Driven Policymaking in the Arab World," 2017.

[5] F. Miro-Llinares and J. J. Rodriguez-Sala, "Cyber hate speech on twitter: Analyzing disruptive events from social media to build a violent communication and hate speech taxonomy," Int. J. Des. Nat. Ecodynamics, vol. 11, no. 3, pp. 406–415, 2016.

[6] A. Brown, "What is hate speech? Part 1: The Myth of Hate," Law Philos., vol. 36, no. 4, pp. 419–468, 2017.

[7] M. Y. Anis and U. S. Maret, "Hatespeech in Arabic Language," in International Conference on Media Studies, 2017, no. September.

[8] N. Chetty and S. Alathur, "Hate speech review in the context of online social networks," Aggress. Violent Behav., vol. 40, no. May, pp. 108–118, 2018.

[9] A. Jha and R. Mamidi, "When does a compliment become sexist? Analysis and classification of ambivalent sexism using twitter data," Proc. Second Work. NLP Comput. Soc. Sci., pp. 7–16, 2017.

[10] N. Albadi, M. Kurdi, and S. Mishra, "Are they Our Brothers? Analysis and Detection of Religious Hate Speech in the Arabic Twittersphere," 2018 IEEE/ACM Int. Conf. Adv. Soc. Networks Anal. Min., pp. 69–76, 2018.

[11] T. Gelashvili and K. A. Nowak, "Hate Speech on Social Media," Lund University, 2018.

[12] P. Fortuna and S. Nunes, "A Survey on Automatic Detection of Hate Speech in Text," ACM Comput. Surv., vol. 51, no. 4, pp. 1–30, 2018.

[13] R. Irfan et al., "A survey on text mining in social networks," Knowl. Eng. Rev., vol. 30, no. 2, pp. 157–170, 2015.

[14] J. Hirschberg and C. D. Manning, "Advances in natural language processing," Science (80-.)., vol. 349, no. 6245, p. 261 LP-266, Jul. 2015.

[15] S. Sun, C. Luo, and J. Chen, A review of natural language processing techniques for opinion mining systems, vol. 36, no. November 2016. 2017.

[16] M. M. Najeeb, A. A. Abdelkader, and M. B. Al-Zghoul, "Arabic Natural Language Processing Laboratory serving Islamic Sciences," Int. J. Adv. Comput. Sci. Appl., vol. 5, no. 3, pp. 114–117, 2014.

[17] S. Khoja and R. Garside, "Stemming arabic text," Lancaster, UK, Comput. Dep. Lancaster Univ., 1999.

[18] S. H. Ghwanmeh, G. Kanaan, R. Al-Shalabi, and S. Rabab'ah, "Enhanced Algorithm for Extracting the Root of Arabic Words," 2009 Sixth Int. Conf. Comput. Graph. Imaging Vis., pp. 388–391, 2009.

[19] M. Boudchiche, A. Mazroui, M. Ould Abdallahi Ould Bebah, A. Lakhouaja, and A. Boudlal, "AlKhalil Morpho Sys 2: A robust Arabic morpho-syntactic analyzer," J. King Saud Univ. - Comput. Inf. Sci., vol. 29, no. 2, pp. 141–146, 2017.

[20] A. Alshutayri and E. Atwell, "Creating an Arabic Dialect Text Corpus by Exploring Twitter, Facebook, and Online Newspapers," no. May, 2018.

[21] A. Goswami and A. Kumar, "A survey of event detection techniques in online social networks," Soc. Netw. Anal. Min., vol. 6, no. 1, pp. 1–25, 2016.

[22] A. Assiri, A. Emam, and H. Al-Dossari, "Towards enhancement of a lexicon-based approach for Saudi dialect sentiment analysis," J. Inf. Sci., vol. 44, no. 2, pp. 184–202, 2018.

[23] M. Wiegand, J. Ruppenhofer, A. Schmidt, and C. Greenberg, "Inducing a Lexicon of Abusive Words – a Feature-Based Approach," in Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), 2018, pp. 1046–1056.

[24] N. D. Gitari, Z. Zuping, H. Damien, and J. Long, "A lexicon-based approach for hate speech detection," Int. J. Multimed. Ubiquitous Eng., vol. 10, no. 4, pp. 215–230, 2015.

[25] S. George K and S. Joseph, "Text Classification by Augmenting Bag of Words (BOW) Representation with Co-occurrence Feature," IOSR J. Comput. Eng., vol. 16, no. 1, pp. 34–38, 2014.

[26] C.-F. Tsai, "Bag-of-Words Representation in Image Annotation: A Review," ISRN Artif. Intell., vol. 2012, pp. 1–19, 2012.

[27] P. Burnap and M. L. Williams, "Us and them: identifying cyber hate on Twitter across multiple protected characteristics," EPJ Data Sci., vol. 5, no. 1, 2016.

[28] G. Xiang, B. Fan, L. Wang, J. Hong, and C. Rose, "Detecting offensive tweets via topical feature discovery over a large scale twitter corpus," Proc. 21st ACM Int. Conf. Inf. Knowl. Manag. - CIKM '12, p. 1980, 2012.

[29] J. Lilleberg, Y. Zhu, and Y. Zhang, "Support vector machines and Word2vec for text classification with semantic features," Proc. 2015 IEEE 14th Int. Conf. Cogn. Informatics Cogn. Comput. ICCI*CC 2015, pp. 136–140, 2015.

[30] P. Burnap and M. L. Williams, "Cyber Hate Speech on Twitter: An Application of Machine Classification and Statistical Modeling for Policy and Decision Making," Policy & Internet, vol. 7, no. 2, pp. 223–242, 2015.

[31] T. Hua, F. Chen, L. Zhao, C.-T. Lu, and N. Ramakrishnan, "STED: semi-supervised targeted-interest event detectionin in twitter," in Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, 2013, pp. 1466–1469.

[32] R. Pandarachalil, S. Sendhilkumar, and G. S. Mahalakshmi, "Twitter Sentiment Analysis for Large-Scale Data: An Unsupervised Approach," Cognit. Comput., vol. 7, no. 2, pp. 254–262, 2015.

[33] M. Al-Smadi, O. Qawasmeh, M. Al-Ayyoub, Y. Jararweh, and B. Gupta, "Deep Recurrent neural network vs. support vector machine for aspect-based sentiment analysis of Arabic hotels' reviews," J. Comput. Sci., vol. 27, pp. 386–393, 2018.

[34] 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, Dec. 2018.

[35] Y. Chen, Y. Zhou, S. Zhu, and H. Xu, "Detecting offensive language in social media to protect adolescent online safety," in Proceedings - 2012 ASE/IEEE International Conference on Privacy, Security, Risk and Trust and 2012 ASE/IEEE International Conference on Social Computing, SocialCom/PASSAT 2012, 2012, no. December, pp. 71–80.

[36] J. H. Park and P. Fung, "One-step and Two-step Classification for Abusive Language Detection on Twitter," in AICS Conference, 2017.

[37] H. Chen, S. McKeever, and S. J. Delany, "Abusive text detection using neural networks," in CEUR Workshop Proceedings, 2017, vol. 2086, pp. 258–260.

[38] S. Salawu, Y. He, and J. Lumsden, "Approaches to Automated Detection of Cyberbullying: A Survey," IEEE Trans. Affect. Comput., vol. 3045, no. c, pp. 1–20, 2017.

[39] K. Dinakar, R. Reichart, and H. Lieberman, "Modeling the detection of Textual Cyberbullying.," Soc. Mob. Web, vol. 11, no. 02, pp. 11–17, 2011.

[40] V. Nahar, S. Al-Maskari, X. Li, and C. Pang, "Semi-supervised Learning for Cyberbullying Detection in Social Networks," in Databases Theory and Applications, 2014, pp. 160–171.

[41] M. Di Capua, E. Di Nardo, and A. Petrosino, "Unsupervised cyber bullying detection in social networks," in 2016 23rd International Conference on Pattern Recognition (ICPR), 2016, pp. 432–437.

[42] S. A. Özel, E. Saraç, S. Akdemir, and H. Aksu, "Detection of cyberbullying on social media messages in Turkish," in 2017 International Conference on Computer Science and Engineering (UBMK), 2017, pp. 366–370.

[43] R. Pawar, Y. Agrawal, A. Joshi, R. Gorrepati, and R. R. Raje, "Cyberbullying Detection System with Multiple Server Configurations," 2018 IEEE Int. Conf. Electro/Information Technol., pp. 90–95, 2018.

[44] B. Doosje, F. M. Moghaddam, A. W. Kruglanski, A. De Wolf, L. Mann, and A. R. Feddes, "Terrorism , radicalization and de-radicalization," Curr. Opin. Psychol., vol. 11, pp. 79–84, 2016.

[45] P. Wadhwa and M. P. S. Bhatia, "Tracking on-line radicalization using investigative data mining," in 2013 National Conference on Communications (NCC), 2013, pp. 1–5.

[46] S. Agarwal and A. Sureka, "Using KNN and SVM Based One-Class Classifier for Detecting Online Radicalization on Twitter," in International Conference on Distributed Computing and Internet Technology, 2015, pp. 431–442.

[47] M. Fernandez and H. Alani, "Contextual semantics for radicalisation detection on Twitter," CEUR Workshop Proc., vol. 2182, 2018.

[48] W. Warner and J. Hirschberg, "Detecting Hate Speech on the World Wide Web," no. Lsm, pp. 19–26, 2012.

[49] I. Kwok and Y. Wang, "Locate the Hate: Detecting Tweets against Blacks," Twenty-Seventh AAAI Conf. Artif. Intell., pp. 1621–1622, 2013.

[50] P. Burnap and M. L. Williams, "Hate Speech, Machine Classification and Statistical Modelling of Information Flows on Twitter: Interpretation and Communication for Policy Decision Making," in Proceedings of the Conference on the Internet, Policy & Politics, 2014, pp. 1–18.

[51] F. Del Vigna, A. Cimino, F. Dell'Orletta, M. Petrocchi, and M. Tesconi, "Hate me, hate me not: Hate speech detection on Facebook," CEUR Workshop Proc., vol. 1816, pp. 86–95, 2017.

[52] I. Alfina, R. Mulia, M. I. Fanany, and Y. Ekanata, "Hate speech detection in the Indonesian language: A dataset and preliminary study," 2017 Int. Conf. Adv. Comput. Sci. Inf. Syst. ICACSIS 2017, vol. 2018–Janua, no. October, pp. 233–237, 2018.

[53] S. Jaki and T. De Smedt, "Right-wing German Hate Speech on Twitter : Analysis and Automatic Detection," p. 29, 2018.

[54] B. Gambäck and U. K. Sikdar, "Using Convolutional Neural Networks to Classify Hate-Speech," Assoc. Comput. Linguist., no. 7491, pp. 85–90, 2017.

[55] P. Badjatiya, S. Gupta, M. Gupta, and V. Varma, "Deep Learning for Hate Speech Detection in Tweets," in Proceedings of the 26th International Conference on World Wide Web Companion, 2017, pp. 759–760.

[56] Z. Zhang, D. Robinson, and J. Tepper, "Detecting Hate Speech on Twitter Using a Convolution-GRU Based Deep Neural Network," in ESWC 2018: The Semantic Web, 2018, pp. 745–760.

[57] E. A. Abozinadah, A. V Mbaziira, and J. H. Jones Jr., "Detection of abusive accounts with Arabic tweets," Int. J. Knowl. Eng., vol. 1, no. 2, pp. 113–119, 2015.

[58] E. A. Abozinadah and J. H. Jones, Jr, "Improved Micro-Blog Classification for Detecting Abusive Arabic Twitter Accounts," Int. J. Data Min. Knowl. Manag. Process, vol. 6, no. 6, pp. 17–28, 2016.

[59] E. A. Abozinadah and J. H. Jones, "A Statistical Learning Approach to Detect Abusive Twitter Accounts," Proc. Int. Conf. Comput. Data Anal. - ICCDA '17, pp. 6–13, 2017.

[60] H. Mubarak, K. Darwish, and W. Magdy, "Abusive Language Detection on Arabic Social Media," Proc. First Work. Abus. Lang. Online, pp. 52–56, 2017.

[61] A. Alakrot, L. Murray, and N. S. Nikolov, "Dataset Construction for the Detection of Anti-Social Behaviour in Online Communication in Arabic," Procedia Comput. Sci., vol. 142, pp. 174–181, 2018.

[62] A. Alakrot, L. Murray, and N. S. Nikolov, "Towards Accurate Detection of Offensive Language in Online Communication in Arabic," Procedia Comput. Sci., vol. 142, pp. 315–320, 2018.

[63] A. A. E. M. B. N. H. Alhuzali and M. Abdul-Mageed, "Think Before Your Click: Data and Models for Adult Content in Arabic Twitter," in Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), 2018.

[64] B. Haidar, M. Chamoun, and A. Serhrouchni, "A Multilingual System for Cyberbullying Detection : Arabic Content Detection using Machine Learning," Adv. Sci. Technol. Eng. Syst. J., vol. 2, no. 6, pp. 275–284, 2017.

[65] A. H. Alduailej and M. B. Khan, "The challenge of cyberbullying and its automatic detection in Arabic text," in 2017 International Conference on Computer and Applications (ICCA), 2017, pp. 389–394.

[66] W. Magdy, K. Darwish, and I. Weber, "#FailedRevolutions: Using Twitter to Study the Antecedents of ISIS Support," in AAAI Spring Symposium Series, 2016.

[67] L. Kaati, E. Omer, N. Prucha, and A. Shrestha, "Detecting Multipliers of Jihadism on Twitter," Proc. - 15th IEEE Int. Conf. Data Min. Work. ICDMW 2015, pp. 954–960, 2016.

[68] N. Djuric, J. Zhou, R. Morris, M. Grbovic, V. Radosavljevic, and N. Bhamidipati, "Hate Speech Detection with Comment Embeddings," in Proceedings of the 24th International Conference on World Wide Web, 2015, pp. 29–30.

[69] S. Malmasi and M. Zampieri, "Challenges in Discriminating Profanity from Hate Speech," J. Exp. Theor. Artif. Intell., vol. 30, pp. 187–202, 2018.

[70] Z. Zhang and L. Luo, "Hate Speech Detection: A Solved Problem? The Challenging Case of Long Tail on Twitter," vol. 1, no. 0, pp. 1–5, 2018.

[71] K. E. Abdelfatah, G. Terejanu, and A. A. Alhelbawy, "UNSUPERVISED DETECTION OF VIOLENT CONTENT IN ARABIC SOCIAL MEDIA," Comput. Sci. Inf. Technol. (CS IT), pp. 1–7, 2017.

[72] W. Yin, K. Kann, and M. Yu, "Comparative Study of CNN and RNN for Natural Language Processing," arXiv Prepr. arXiv 1702.01923, 2017.