Attention-based networks for analyzing inappropriate speech in Arabic text

Mohamed BERRIMI
dept. of computer science
University of Ferhat Abbas 1
Setif, Algeria
mohamed.berrimi@univ-setif.dz

Abdelouaheb Moussaoui
dept. of computer science
University of Ferhat Abbas 1
Setif, Algeria
Abdelouaheb.Moussaoui@univ-setif.dz

Mourad Oussalah
dept. of Computer Science and Engineering
University of Oulu,
Oulu, Finland
mourad.oussalah@oulu.fi

Mohamed Saidi
dept. of Computer Science
University of Ferhat Abbas 1,
Setif, Algeria
mohamed.saidi@univ-setif.dz

***Abstract*—Analyzing social media posts and comments has become a critical task to prevent cyberbullying and hate speech. In this work we present a classification models based on the attention mechanism to analyze Arabic posts and filter out all kinds of inappropriate speech including Religious based hate speech, offensive and abusive content in different Arabic dialects. The attention-based models show promising results for four Arabic datasets. The results are presented and compared in terms of accuracy and training time**

***Index Terms*—Attention mechanism,Text classification, Hate speech detection, Social Media Mining, Arabic language, Deep learning**

I. Introduction

The Internet is becoming widely used, and social media websites play a role model in many peoples lives, especially the teens culture. Recent surveys on social media services use mentions that ninety percent of adolescents aged between 13 and 17 have used social media, where 51% of them report daily visit a social media site. On average, teens are online almost nine hours a day, not including time for homework [1]. This usage has many benefits in their daily life; life. For instance, it keeps them connected to their event of interest; it helps them in their school duties, and; it allows them to find relevant online communities and support whenever needed. Nevertheless, it could also cause serious mental health issues, harassment and exposure to online harm [2]. A survey reported that 50% of young people have been abused online [4], and 13% of them reported that they experienced cyberbullying at least once. Social media could be a conduit for accessing inappropriate content like violent images or pornography. In a multicenter study that assessed the emotional impact of different forms of bullying and cyberbullying, 68.5% of adolescents reported experiencing negative emotions such as anger, upset, and depression [21]. Cyberbullying is a serious problem in Arab countries, as a form of harassment that become increasingly common, especially among teenagers on social media. According to a survey, 20.9% of middle-school adolescents report bullying in UAE, 31.9% in Morocco, 33.6% in Lebanon, 39.1% in Oman and 44.2% in Jordan [29]. One basic existence rule of social media is to give people the ability to create, share ideas and information, and to express their opinions and beliefs without barriers. To cyberbullying impact, these rights are restricted and moderated through policies. Many social media websites have addressed solutions. A policy Rationale by Facebook mentions that they do not allow hate speech on their platform because it creates an environment of intimidation and exclusion, and in some cases, may promote real-world violence. Accordingly, many tiers and Facebook restrictions on posts have been put forward focusing on content where gender, ethnicity or religious beliefs are targeted [3]. Google on the other side have made hard restrictions and policies in a response of fighting hate speech, and expanded their use of automatic services to help detect potentially violative content and send it for human review, where they tasked over 10,000 people with detecting, reviewing, and removing content that violates Youtubes guidelines. This operation led to removing more than 17,000 channels and 100,000 videos, along with at least 500 million comments [7]. Declaring the speech to be hateful and offensive depends on the context where it is used, For example, members of a specific race can refer to each other using terms that are generally considered insults. When used consensually, the intent behind these terms is not unreasonable, but a means of retrieving terms that were historically used to demean individuals [6]. Similarly, jokes can also have either harmful meaning or amusement depending on context and discourse. This includes, for instance, expression I will kill you when two people are joking. Looking at the context of the conversation is essential to decide whether it presents a serious threat or not. As many people are exposed more to social media and online blogs, it has implied a great responsibility for data-scientists and researchers to intervein. In response to this matter, we aim to propose deep learning architectures, using soft attention to automatically detect religious conflicts and hate, violent, abusive, offensive speech in Arabic (text in MSA

and different dialects). In summary, we discuss the challenges for automatic detection of inappropriate speech in Arabic texts, including competing definitions, availability and processing of data sets, and existing approaches. We also propose a new approach based on the attention mechanism, which aims to give better performance in terms of precision, model complexity, training time, and interpretability of decisions being made. We summarize our contribution as follow: Section 2 provides an analysis of the most recent works dealing with hate speech on social media. Section 3 describes the four Arabic datasets used in our study, and some analysis is also presented. Section 4 discusses our proposed Deep attention model for faster and accurate text classification. Section 5 we present our experiments and results, also comparisons on SOTA approaches and recent proposed works on the used data. Furthermore, concluding on the importance of fighting inappropriate speech on social media and the effectiveness of attention-based models in Natural language processing, and their added value on the explainability of deep learning.

II. Related works

This section presents different aspects of hate speech, as well as researches associated with the Arabic language. Due to the sensitivity of this matter, many studies were conducted to protect personals on social media, especially adolescents.

There are more than 168.1M Arabic speakers on the Internet, and Arabs are nowadays onlines more than ever, therefore many studies including datasets proposals have been published to limit the phenomena of offensive speech.

Elmadany et al. [9] proposed a deep learning system based on Bidirectional Transformers BERT for offensive language detection. In their experiments, they used two types of data: data distributed by the Offensive Language Detection shared task and an automatically collected dataset.

The first data contains 10,000 tweets manually annotated for two sub-task: offensive speech, and hate speech. Their systems performance came up with 89.60% accuracy (82.31% macro F1) for hate speech and 95.20% accuracy and 70.51% macro F1 on official TEST data.

Alshehri et al. [10] worked on Understanding and Detecting Dangerous Speech in Social Media on Arabic texts. They manually curate a multi-dialectal lexicon of physical harm threats. They used to collect a large dataset of threatening speech from Arabic Twitter, and manually annotate a subset of the data for dangerous speech, then trained BERT model for detecting hate speech. Their system yield a F1 score between 53.42% and 59.60% on detecting hate speech.

Haddad et al. [18] proposed an Attention-based Deep Neural Networks to detect the offensive speech in the Arabic language. They worked on the OffensEval 2020 dataset where they conducted many experiments using the Bidirectional GRU model augmented with an attention layer, that achieved 0.859 F1 score for the task of offensive language detection, and 0.75 F1 score for the task of hate speech detection.

Haidar et al. [32] collected a dataset for detecting Arabic Cyberbullying on social media, then trained machine learning classifiers such as Nave Bayes and SVM. They obtained a precision of 90.1 and 93.4% and published another paper using Deep learning where trained simple feed-forward layers architecture, achieving 94.56%.

Albadi N. [30] addressed the religious-based hate content problem on social media, in their study they presented a large annotated Arabic dataset, along with collection of lexicon consisting of terms commonly found in religious discussions, for hate speech detection. In their study [30] the authors trained a Gated Recurrent nets with GRU cells, using pretrained word embeddings to detect religious hate speech with 0.84 (AUROC).

Using the same data, Chowdhury et al. [11] proposed ARHNet (Arabic Religious Hate Speech Net) model incorporates both Arabic Word Embeddings and Social Network Graphs for the detection of religious hate speech, their system obtained f1-score of 0.78. Many recent studies have addressed the problem of hate speech in English, Chatzakou et al. [12] proposed on a concrete study to understand the characteristics of abusive behavior in Twitter to detect Cyberbullying and cyberaggression in English text. They analyzed 1.2 million users and 2.1 million tweets, comparing users participating in discussions around seemingly normal topics, to those more likely to be haterelated, and also explored specific manifestations of abusive behavior, i.e., cyberbullying and cyberaggression, in one of the hate-related communities. Using various state-of-the-art machine learning algorithms, they classify these accounts with over 90% accuracy and AUC. In the next section, We discuss some of the Arabic datasets available to train and measure the performance of inappropriate speech detection models.

III. Datasets

To evaluate the approaches, we used four available datasets related to offensive, abusive, hate speech in the Arabic language.

- Mubarak et al. [5] collected an Arabic dataset for the task of detecting offensive speech, obtained from comments deleted from Aljazira.com which is popular Arabic news. news. It was then manually moderated so that as pointed out in the comment guidelines, the posts are removed if it contains a personal attack, racist, sexist, any form of offensiveness. The authors initially obtained up to 400K comments on approximately 10K articles that cover many domains then selected randomly 32K deleted comments whose lengths are between 3 and 200 characters to ease subsequent annotation. The selected comments were annotated using CrowdFlower, where three annotators were asked to classify comments as obscene, offensive, or clean. The comments are written in Modern Standard Arabic (MSA) and different dialects.
- Albadi N. [30] proposed the first Arabic dataset related to religious-based hate content; this dataset focuses on the four most common religious beliefs in the Middle East region (Islam, christianity, Judaism and Atheism) [30]. Since Islam is the most practiced religion in this region,

the dataset included the two main sects of Islam, namely Sunni and Shia, which comprises 87-90%, and 10-13% of all Muslims, respectively.
- We also used the Subtask A dataset shared within The 4th Workshop on Open-Source Arabic Corpora and Processing tools. The dataset contains 10,000 tweets that were manually annotated (labels are: OFF or NOT OFF), we could retrieve only 7500 tweets.
- The L-HSAB [35] and T-HSAB [23] are two different datasets that contain Levantine Hate Speech and ABuse and Tunisian Hate speech and ABuse texts, collected from Twitter. L-HSAB dataset combines 5,846 Syrian/Lebanese political tweets, T-HSAB combines 6,024 Tunisian comments, both labeled with Abuse, Hate, and Normal. In this study, we combined the two datasets, since they present the same labels to obtain a larger dataset of Tunisian and Levantine Arabic dialects of Abusive and Hate speech.

A. Preprocessing

Social media posts generally lack uniformity in writing styles and do not respect the grammar standards; this makes building reliable language models much difficult. Therefore we normalized the datasets as follows:

- Normalizing links, user mentions, and numbers.
- Removing the hashtags by deleting underscores and the # symbol.
- Removing punctuations, emojis, and words with a length of 1.
- Normalized the words by reducing the repeated characters if the repetition count three or more.
- Stopwords are irrelevant tokens that add no meaning to the sentence, but in this task, altering them could completely change the meaning, therefore we kept all stopwords. adds much value to the context in sentiment analysis.
- Many of the tweets are not in MSA, finding a good stemmer could be challenging, hence we decided to not lemmatize nor stemm the words.
- Removed all non-arabic scripts from the texts, except for T-HSAB dataset that contains texts in Arabizi.

IV. Appraoche

This section presents the main corp of our research, where we illustrate the different models and approaches that we applied during the experimental study. Attention Mechanism is becoming widely used and is one of the most popular mechanisms in the recent Natural language processing research field. Attention was first introduced for machine translation tasks by Bahdanau et al. [25]

The intuition behind attention was inspired by the human biological systems; our visual processing system tends to focus selectively on parts of the frame (scene being looked at) while ignoring other irrelevant information in a manner that can assist in the perception. Attention makes use of this notion by allowing the model to dynamically pay attention to only certain parts of the input that helps in performing the task, and neglect other parts, making the processing faster.

The rapid advancement in modeling attention in neural networks is primarily due to three reasons:

- Attention is the corp mechanism in many state-of-the-art SOTA models like BERT [36], Transformer [37] and used in different tasks such as image captioning [31].
- Beside the remarkable performance on the main task, The mechanism also brings an important feature to the world of neural networks which is the interpretability of the results.
- The mechanism was mainly proposed to overcome the limitation of by RNNs when dealing with long input sequences in Machine translation.

In this paper, we apply Soft attention mechanism [25] to detect different types of inappropriate speech in the Arabic texts. Therefore, we propose three architectures, all of them contains an embedding layer as a first layer.

- **EL** LSTM layer and two dense layers: In this study, we did not use pre-trained embeddings for semantic extraction, we set the Embedding layer to be trainable, the output of this layer is an embedding vector of size (50, 150), was fed into an LSTM layer with 250 hidden units was used to capture long-distance contextual information, the LSTM layer will return the full sequence instead of just returning the last output hidden state. Recurrent neural nets expect to receive sequential data with the same vector lengths, and Collected texts from social media, do not the same text length, for this reason, short texts are padded with special word paddings that do not add any meaning (zeros 0 for instance), for each dataset we looked for the longest sentence and padded all short sentences with 0 value to have an equal length of the longest sentence.as a result, all words will be represented with same vectors length.
- **ESoA**: Soft attention layer [25] and final output layer: The output of the embedding layer will be fed to a soft attention layer that uses a weighted average of all masked states in the input sequence to create the context vector. The use of the soft weighing method makes the neural network suitable for effective learning by backpropagation in a quadratic calculation. [28] The core idea behind applying soft attention is to compute a weight distribution on the input sequence and assign higher weights to the relevant parts of the input sequence.

- **ELSoA** Part of this architecture is illustrated in Figure 1. We add to the previous architecure an LSTM layer, to see the impact of the attention mechanism with and without recurrent networks.

V. EXPERIMENTS AND RESULTS

This section summarizes the major findings of our work. First, we compare the three models, on each dataset, based on the training time and number of parameters in table 1, then

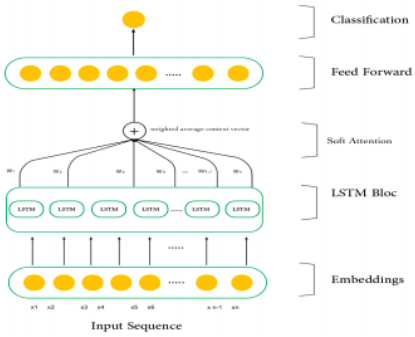

Fig. 1. Figure Proposed LSTM + soft Attention architecture ELSoA

we move to compare the results obtained by each model on different datasets and also with previous works.

To conduct the experiments, we used a TESLA T4 GPU card and applied the three proposed models to perform experiments with the different datasets described earlier, in order to analyze the performance of those models we used word embeddings as a feature learning techniques.

For training the dataset, we set Adam as our Optimizer for all the experiments, and trained all the models for ten 10 epochs. We report the training process in seconds per each epoch and binary cross-entropy as a loss function for the first three models, except the last one that contains three output classes, we set categorical cross-entropy.

The performance of the models are reported in three metrics namely Accurac, Precision and F1 mesure.

$$Precision = \frac{\text{TP}}{\text{TP + FP}}$$

$$F1 = \frac{2 \cdot precision \cdot recall}{precision + recall}$$

$$Accuracy = \frac{TP+TN}{TP+FN+TN+FP}$$

For each data, these are the hyperparameters of all the models.

- For the religious hate speech dataset [30] we trained the embedding layer on a total of 20000 tokens from the original dataset, the batch size is set to 62 with a single neuron in the output layer and a binary cross-entropy as a loss function.
- Dataset 2 [5] This dataset initially contained three classes, in this study, we merged the offensive and hate speech classes to be one class, we ended with a binary classification dataset, same as reported in the original paper. The batch size is set to 80 and binary cross-entropy as a loss function.
- OffensEval 2020 dataset: The retrieved dataset contained only 7500 tweets (instead of 10000 reported in the competition).
- LHSAB+THSAB dataset [35] [23]: The datasets are presented in section 3, in this study we combined the two datasets in order to obtain a large Arabic dialects for hate speech. The resulting data contains 11869 tweets labeled as Abusive, Hate speech and normal texts.

We evaluated the models on an unseen part of the data. Accuracy, Precision, and F1-score were the metrics used to evaluate the performance of the models through all experiments. It is also important to consider both performance and time, especially in critical systems requiring a fast response [20] The attention models were the least time consuming over all the models, and also the less parametrized.

TABLE I
MODELS COMPARAISON ACCROSS ALL DATASETS

	Model	# Parameters	Training/epoch (seconds)
Data 1	EL	9.851.000	20
	ESoA	5.006.866	3
	ELSoA	1352000	19
Data 2	EL	13,851,401	112
	ESoA	3,010,451	3
	ELSoA	3,351,651	107
Data 3	EL	16,301,801	17
	ESoA	1,020,851	1
	ELSoA	1,402,051	17
Data 4	EL	10,851,803	40
	ESoA	1,402,051	1
	ELSoA	1,852,053	41

TABLE II
PERFORMANCE REPORT ON DATASET [5]

D1	Accuracy	P	F1
ESoA	90.5	90.39	90.36
ELSoA	**90.79**	90.78	90.49
EL	89.83	88.93	89.95

The tables show that the use of LSTM, which produces less reliable models, requires longer computational time and much weights than both attention models, the concatenation of the LSTM model and the attention models enhances performance in the majority of datasets.We did not find the full dataset, but our models have performed better than all reported results on the OSACT 2020 competition in all metrics. The first observation is the remarkable performance of Soft Attention compared to LSTM based models.The AM improved performance accuracy by + 1.2% on majority of data, This thus demonstrates the ability to produce and use contextual representations and focus only on parts on the relevant parts of the input sequences.

TABLE III
PERFORMANCE REPORT ON RELIGIOUS HATE SPEECH DATASET [30]

Data 2	Accuracy	Precision	F1 score
[26]	/	/	78%
[30]	77%	76%	77%
EL	95.5%	95.31%	95.96%
ESoA	96.47%	97.77	97.86%
ELSoA	**96.59%**	97.64%	97.92%

The good performance could be related to both word embeddings and the attention mechanism that both helps in

TABLE IV
PERFORMANCE REPORT ON OFFENSEVAL 2020 DATASET

Data 3	Accuracy	Precision	F1 score
Djandjil et al. [33]	90%	90.67%	93.7%
Haddad et al [18]	93.85%	90.17%	90.5%
ESoA	97.4%	97.14%	97%
ESoA	**97.47%**	97.77%	97.17%
EL	96.33%	97.12%	97.1%

TABLE V
PERFORMANCE REPORT ON T-HSAB AND L-HSAB COMBINED DATASETS

Data 4	Accuracy	Precision	F1 score
EL	95.90%	95.95%	95.89%
ESoA	95.67%	95.77%	95.67%
ELSoA	96.18%	96.14%	96.09%

prioritizing contextual information representation and selecting most relevant words contributing to the task.

Comparing our results with previous works in three first datasets, its observable that our models outperformed the latter with a remarkable margin.

The

VI. Conclusion

In this work, we show the critical issue with hate speech and offensive content on Arabic social media communities, and we propose a novel deep learning architecture based on the attention mechanism for smooth and accurate learning and classification. Experimental results over four datasets show that the Attention-based models outperform other architectures by a significant margin in terms of performance and processing time. For future work, we aim to investigate the use of pretrained Arabic embeddings as well as TF-IDF feature extraction and test our models on multilingual datasets.

References

[1] Social Media and Teens, March 2018, https://www.aacap.org/AACAP/Families and Youth/Facts for Families/FFFGuide/Social-Media-and-Teens-100.aspx. accessed on June 28,2020.

[2] Costello, C. R. (2017, October). Social Media, Youth, and the Law: Legal Risks and Protection for Young People Interacting Online. In 64th Annual Meeting. AACAP.

[3] Facebook, Community Standards, Hate speech, https://www.facebook.com/communitystandards/hate speech, accessed on June 28,2020

[4] The Common Sense Census: Media Use by Tweens and Teens, 2019, https://www.commonsensemedia.org/research/the-common-sensecensus-media-use-by-tweens-and-teens-2019, accessed on June 28,2020

[5] Mubarak, H., Darwish, K., & Magdy, W. (2017, August). Abusive language detection on Arabic social media. In Proceedings of the first workshop on abusive language online (pp. 52-56).

[6] Help Center,Twitter Rules and policies, Hateful conduct policy, https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy, accessed on June 28,2020.

[7] Official YouTube Blog:The Four Rs of Responsibility, Part 1: Removing harmful content, https://youtube.googleblog.com/2019/09/the-four-rs-ofresponsibility-remove.html?m=1, Tuesday,September 3,2019, accessed on June 28,2020.

[8] Fox News, YouTube removed 17,000 channels, 500 million comments under new hateful conduct policy, https://www.foxnews.com/tech/google-removed-hateful-17000-channels-500-million-comments, September 3, accessed on June 28,2020.

[9] Elmadany, A., Zhang, C., Abdul-Mageed, M., & Hashemi, A. (2020). Leveraging Affective Bidirectional Transformers for Offensive Language Detection. arXiv preprint arXiv:2006.01266.

[10] Alshehri, A., Nagoudi, E. M. B., & Abdul-Mageed, M. (2020). Understanding and Detecting Dangerous Speech in Social Media. arXiv preprint arXiv:2005.06608.

[11] Chowdhury, A. G., Didolkar, A., Sawhney, R., & Shah, R. (2019, July). ARHNet-Leveraging Community Interaction for Detection of Religious Hate Speech in Arabic. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop (pp. 273-280).

[12] Chatzakou, D., Leontiadis, I., Blackburn, J., Cristofaro, E. D., Stringhini, G., Vakali, A., & Kourtellis, N. (2019). Detecting cyberbullying and cyberaggression in social media. ACM Transactions on the Web (TWEB), 13(3), 1-51.

[13]] Van Hee, C., Lefever, E., Verhoeven, B., Mennes, J., Desmet, B., De Pauw, G., & Hoste, V. (2015). Automatic detection and prevention of cyberbullying. In International Conference on Human and Social Analytics (HUSO 2015) (pp. 13-18). IARIA.

[14] D. Yin, B. D. Davison, Z. Xue, L. Hong, A. Kontostathis, and L. Edwards, Detection of Harassment on Web 2.0, in Proceedings of the Content Analysis in the Web 2.0 (CAW2.0) April 21, 2009, Madrid, Spain. CAW 2.0, Apr. 2009, pp. 12311238, CAW 2.0, URL: http://wbox0.cse.lehigh.edu/brian/pubs/2009/CAW2/ harassment.pdf [accessed: 2020-06-25].

[15] Ousidhoum, N., Lin, Z., Zhang, H., Song, Y., & Yeung, D. Y. (2019). Multilingual and multi-aspect hate speech analysis. arXiv preprint arXiv:1908.11049.

[16] Ibrohim, M. O., & Budi, I. (2019, August). Multi-label hate speech and abusive language detection in Indonesian twitter. In Proceedings of the Third Workshop on Abusive Language Online (pp. 46-57).

[17] Chaudhari, S., Polatkan, G., Ramanath, R., & Mithal, V. (2019). An attentive survey of attention models. arXiv preprint arXiv:1904.02874.

[18] Haddad, B., Orabe, Z., Al-Abood, A., & Ghneim, N. (2020, May). Arabic Offensive Language Detection with Attention-based Deep Neural Networks. In Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection (pp. 76-81).

[19] Lin, Z., Feng, M., Santos, C. N. D., Yu, M., Xiang, B., Zhou, B., & Bengio, Y. (2017). A structured self-attentive sentence embedding. arXiv preprint arXiv:1703.03130.

[20] Neppalli, V.K.; Caragea, C.; Squicciarini, A.; Tapia, A.; Stehle, S.J. Sentiment analysis during Hurricane Sandy in emergency response. Int. J. Disaster Risk Reduct. 2017, 21, 213222

[21] Bottino, S. M. B., Bottino, C. M. C., Regina, C. G., Correia, A. V. L., & Ribeiro, W. S. (2015). Cyberbullying and adolescent mental health: systematic review. Cadernos de Sade Pblica, 31(3), 463475. doi:10.1590/0102-311x00036114

[22] Ybarra ML. Linkages between depressive symptomatology and internet harassment among young regular internet users. Cyberpsychol Behav 2004; 7:247-57. 14.

[23] Haddad, H., Mulki, H.,& Oueslati, A. (2019, October). THSAB: A Tunisian Hate Speech and Abusive Dataset. In International Conference on Arabic Language Processing (pp. 251-263). Springer, Cham

[24] Chang FC, Lee CM, Chiu CH, Hsi WY, Huang TF, Pan YC. Relationships among cyberbullying, school bullying, and mental health in Taiwanese adolescents. J Sch Health 2013; 83:454-62. 15.

[25] Bahdanau, Dzmitry & Cho, Kyunghyun & Bengio, Y.. (2014). Neural Machine Translation by Jointly Learning to Align and Translate. ArXiv. 1409.

[26] Chowdhury, A. G., Didolkar, A., Sawhney, R., & Shah, R. (2019, July). ARHNet-Leveraging Community Interaction for Detection of Religious Hate Speech in Arabic. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop (pp. 273-280).

[27] Juvonen J, Gross EF. Extending the school grounds? Bullying experiences in cyberspace. J Sch Health 2008, 78:496-505

[28] Galassi, A., Lippi, M., & Torroni, P. Attention in Natural Language Processing.

[29] Kazarian and J. Ammar, School Bullying in the Arab World: A Review, The Arab Journal of Psychiatry , vol. 24, no. 1, pp. 37 - 45, 2013

[30] Albadi, N., Kurdi, M., & Mishra, S. (2018, August). Are they our brothers? Analysis and detection of religious hate speech in the Arabic Twittersphere. In 2018 IEEE/ACM International Conference on Ad-

vances in Social Networks Analysis and Mining (ASONAM) (pp. 69-76). IEEE

[31] Huang, Lun & Wang, Wenmin & Chen, Jie & Wei, Xiao-Yong. (2019). Attention on Attention for Image Captioning

[32] Haidar, Batoul & Maroun, Chamoun & Serhrouchni, Ahmed. (2018). Arabic Cyberbullying Detection: Using Deep Learning. 284-289. 10.1109/ICCCE.2018.8539303.

[33] Djandji, M., Baly, F., & Hajj, H. (2020, May). Multi-Task Learning using AraBert for Offensive Language Detection. In Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection (pp. 97-101).

[34] R. M. Duwairi, R. Marji, N. Shaban, and S. Rushaidat, Sentiment analysis in Arabic tweets, in information and communication systems (icics), 2014 5th international conference on. IEEE, 2014, pp. 16

[35] Mulki, Hala & Haddad, Hatem & Bechikh Ali, Chedi & Alshabani, Halima. (2019). L-HSAB: A Levantine Twitter Dataset for Hate Speech and Abusive Language. 10.18653/v1/W19-3512.

[36] Devlin, Jacob & Chang, Ming-Wei & Lee, Kenton & Toutanova, Kristina. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.

[37] Vaswani, Ashish & Shazeer, Noam & Parmar, Niki & Uszkoreit, Jakob & Jones, Llion & Gomez, Aidan & Kaiser, Lukasz & Polosukhin, Illia. (2017). Attention Is All You Need