informatics

MDPI

Article

Arabic Offensive and Hate Speech Detection Using a Cross-Corpora Multi-Task Learning Model

Wassen Aldjanabi [1,†], Abdelghani Dahou [1,2,†], Mohammed A. A. Al-qaness [3,*], Mohamed Abd Elaziz [4], Ahmed Mohamed Helmi [5] and Robertas Damaševičius [6,*]

1 Department of Mathematics and Computer Science, Faculty of Science and Technology, University of Ahmed DRAIA, Adrar 01000, Algeria; wassen.eldjanabi@gmail.com (W.A.); dahou.abdghani@univ-adrar.edu.dz (A.D.)
2 LDDI Laboratory, Faculty of Science and Technology, University of Ahmed DRAIA, Adrar 01000, Algeria
3 State Key Laboratory for Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China
4 Department of Mathematics, Faculty of Science, Zagazig University, Zagazig 44519, Egypt; abd_el_aziz_m@yahoo.com
5 Department of Computer and Systems Engineering, Faculty of Engineering, Zagazig University, Zagazig 44519, Egypt; amhm162@gmail.com
6 Faculty of Applied Mathematics, Silesian University of Technology, 44-100 Gliwice, Poland
* Correspondence: alqaness@whu.edu.cn (M.A.A.A.-q.); robertas.damasevicius@polsl.pl (R.D.)
† These authors contributed equally to this work.

Abstract: As social media platforms offer a medium for opinion expression, social phenomena such as hatred, offensive language, racism, and all forms of verbal violence have increased spectacularly. These behaviors do not affect specific countries, groups, or communities only, extending beyond these areas into people's everyday lives. This study investigates offensive and hate speech on Arab social media to build an accurate offensive and hate speech detection system. More precisely, we develop a classification system for determining offensive and hate speech using a multi-task learning (MTL) model built on top of a pre-trained Arabic language model. We train the MTL model on the same task using cross-corpora representing a variation in the offensive and hate context to learn global and dataset-specific contextual representations. The developed MTL model showed a significant performance and outperformed existing models in the literature on three out of four datasets for Arabic offensive and hate speech detection tasks.

Keywords: multi-task learning; Arabic language model; contextual representations; offensive language; hate speech

Citation: Aldjanabi, W.; Dahou, A.; Al-qaness, M.A.A.; Abd Elaziz, M.; Helmi A.M.; Damaševičius, R. Arabic Offensive and Hate Speech Detection Using a Cross-Corpora Multi-Task Learning Model. *Informatics* **2021**, *8*, 69. https://doi.org/10.3390/informatics8040069

Academic Editor: Antony Bryant

Received: 26 August 2021
Accepted: 3 October 2021
Published: 8 October 2021

1. Introduction

In recent years, the use of the social networks has substantially increased in the Arab world. It has allowed more freedom for opinion expression, especially in the political domain. Moreover, organizations in the Arab region have embraced social media in their businesses at varying scales, assuming that it significantly affects business development. Due to the freedom of speech given to social media users, it has become relatively easy to propagate abusive or hate speech towards individuals, groups, or societies. The Cambridge Dictionary (https://dictionary.cambridge.org/fr/ (accessed on 10 May 2021)) defines hate speech as "public speech that expresses hate or encourages violence towards a person or group based on something such as race, religion, sex or sexual orientation". Online hate speech is characterized as the use of an offensive language, aimed at a specific group of people who share some common trait [1], while social networks have been recognized as a very favorable medium often used for planning and executing hate attack related activities [2]. Beyond the psychological harm, such toxic online content may be influencing

and radicalizing individuals and could lead to actual hate crimes [3]. Therefore, it is important to detect such cases of cyber-aggression and cyber-bullying in good time [4].

Natural language processing (NLP) is a field of research that deals with machine learning (ML) algorithms applied to human natural languages [5]. NLP applications aim to automatically process written and spoken human languages including sentiment analysis [6,7], sarcasm detection [8], machine translation [9], speech recognition [10], automated dialogue systems [11], urban studies [12,13], topic classification [14], similarity detection [15], text summarization [16], intent detection [17], news and social media analysis [18,19], part-of-speech (POS) tagging [20], authorship attribution [21,22], fake tweet detection [23], coreference resolution [24] and others [14,25–27]. Recently, NLP techniques have also been employed to study the sentiments and attitudes of social media users regarding the COVID-19 pandemic [28,29].

The Arabic language is the Arab world's first language, characterized by its rich and complex grammatical structure [30,31]. In addition, the Arabic morphology is perplexing because there are about 10,000 roots and more than 900 patterns that are the basis for nouns and verbs [32]. The Arabic language has different variations that are used within a community and in specific circumstances [33,34]. Figure 1 shows some typical examples of offensive and hate speech phrases found on Twitter.

(**a**) Hate speech tweet. (**b**) Offensive tweet

Figure 1. Examples of offensive and hate speech tweets in Arabic with translation to English.

Detecting hate speech (HS) is a challenging task [35] due to a lack of common understanding and agreement of what is hate speech, and the lack of high quality annotated datasets, especially for non-English languages. There have been some works on tasks related to hate speech and offensive language detection (OFF) [36,37], including in the Arabic language [38]. Most of these works assign labels to a given input; the labels vary due to the absence of a universal definition of hate and offensive speech. Arguably, all hate speech, aggressive subjects, cyberbullying, and toxic comments make different forms of offensive and hate content present or absent in different corpora. Furthermore, treating every classification task separately consumes more resources.

Previous studies often exploited content-based features (such as syntactical, lexical, and sentiment-based information) for OFF recognition [39]. Lately, however, the content-based features are fused with semantic features, word embeddings and representation learning, user activities (such as frequency of posting), follower network properties and demographic characteristics to derive more complex models [40]. Here comes the role of the multi-task learning (MTL) approach that improves the performance of multiple classification tasks by learning them jointly. In our study, we adopted an MTL model

that relies on a pre-trained Bidirectional Encoder Representation from Transformer (BERT) language model for the Arabic language [41] to perform hate speech and offensive language classification. The Arabic MTL model has experimented with two different language models to cover modern standard Arabic (MSA) and dialect Arabic (DA). Moreover, we extend the tasks of the model to improve its performance, taking advantage of different available Arabic hate speech corpora, where instead of training the MTL model on multiple tasks, we train it on multiple corpora on the same task. Thus, the MTL model can learn global and dataset-specific rich contextual representations.

This study's contributions can be summarized as follows:

- Comprehensive evaluation of single-task and MTL models built upon Transformer language models (LMs);
- We evaluated a new pre-trained model, MarBERT, to classify both DA and MSA tweets;
- We propose a model to explore the multi-corpus-based learning using Arabic LMs and MTL to improve the classification performance on Arabic offensive and hate speech detection.

The rest of the research is organized as follows: Section 2 reviews the most recent related work on offensive and hate speech detection. In Section 3, we provide details of the proposed MTL model. Section 4 presents and discusses the results of the conducted experiments. The conclusion and future works are presented in Section 5.

2. Related Works

This section presents a review of the recently proposed methods for offensive and hate speech detection from user-generated content on social media for English and Arabic languages.

2.1. Hate and Offensive Speech Detection in English

Waseem et al. [42] used a logistic regression classifier to identify hate speech (HS) tweets. They identified the appropriate features that provide the best identification performance. More so, they evaluated the proposed method with 16 K tweets achieving an F1-score of 73.93%, moreover using non-linguistic features like the gender or location can improve the performance but it is always inaccessible or unreliable on social media. In [43], a convolutional neural network (CNN) model was proposed to detect HS using four models [42] including character 4-grams, word2vec, randomly generated word vectors, and character n-grams combined with the word2vec model. The study shows that the best performance was obtained using the Word2vec model. Using Recurrent Neural Network (RNN) and Long Short Term Memory (LSTM) models with word frequency vectorization for HS classification and the previous dataset collected by [42], Pitsilis et al. [44] achieved 0.87 and 0.88 in recall and precision, respectively. Watanabe et al. [45] used text patterns and unigrams as features to train a J48graft machine learning algorithm, achieving an accuracy of 87.4% in detecting whether a tweet is hateful or not.

Basile et al. [46] studied HS towards immigrants and women for both English and Spanish tweets, using SVM and an RBF kernel trained on Universal Sentence Encoder embeddings [47]. The winning team in English achieved a 65.1% macro F1-score using an SVM classifier with an RBF kernel using the SemEval 2019 (Task5) dataset (https://github.com/cicl2018/HateEvalTeam) (accessed on 10 May 2021). Recently, BERT models have shown a good performance in HS detection. As described by Zampieri et al. [48] and Ping Liu et al. [49], the BERT outperformed SVM, CNN, and LSTM models by achieving the best performance on Semeval 2019 with a 82.9% F1 score.

Liu et al. [50] combined Multi-task Learning and the pre-training language model BERT [51] to propose a new Multi-Task Deep Neural Network (MT-DNN). MT-DNN used the General Language Understanding Evaluation (GLUE (https://gluebenchmark.com/) (accessed on 10 May 2021)) benchmark and achieved new state-of-the-art results on eight

out of nine Natural Language Understanding (NLU) tasks such as: CoLA, SST-2, STS-B, RTE, MNLI, QQP, MRPC and QNLI.

2.2. Hate and Offensive Speech Detection in Arabic

Albadi et al. [52] collected the first HS dataset with about 6.6 K Arabic HS tweets. The support vector machine (SVM) classifier and a GRU (Gated Recurrent Unit) trained on AraVec embeddings [53] were utilized for the classification task and achieved the best performance with 79% accuracy. Ousidhoum et al. [54] built a multilingual HS dataset consisting of English, French, and Arabic tweets. Amazon Mechanical Turk was used to label 13 K tweets into diverse aspects such as target attributes, target groups, directness, and hostility types. BiLSTM and Sluice networks [55] performed better than traditional bag-of-words models in most of the multi-label classification tasks.

Mulki et al. [56] created a dataset of 6 K tweets containing hate and offensive speech for the Tunisian dialect from Twitter. The authors extracted several n-gram features from each tweet using Term Frequency (TF) weighting. The extracted features were used to develop SVM and Naive Bayes (NB) classifiers, achieving an 83.6% F1-score. This work is limited to this specific dialect and does not perform well on small datasets. Djandji et al. [57] proposed a model based on AraBERT [41] with MTL during the shared task of OFF Detection in the 4th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT4 (https://edinburghnlp.inf.ed.ac.uk/workshops/OSACT4/) (accessed on 10 May 2021)) [58]. Their proposed model solved the data imbalance problem by leveraging the information from multiple tasks simultaneously and achieved the best performance with a 90% macro-F1 score.

Abu Farha et al. [59] developed a multitask learning architecture based on CNN-BiLSTM, which is trained to detect HS and offensive language. The model incorporates more data through adding sentiment information using the Mazajak Sentiment Analyser [60]. The proposed model achieved a 90.4% F1-score in OFF and 73.7% in the HS task. Hassan et al. [61] implemented various classical ML and DL approaches, such as SVM, CNN BiLSTM, and Multilingual BERT, for the HS subtask. The stacked SVMs achieved an 80.6% macro F1-score.

Otiefy et al. [62] applied several models in their proffering for SemEval2020 on identifying and categorizing offensive language. Their best model ranked 10 out of 52 participants, achieving an 88.72% F1-score using the SVM model, in which they used a combination of both character and word n-grams. Husain et al. [63] studied the impact of offensive language linguistic features on sarcastic language and sentiment content. Their system depended mainly on the pre-trained models AraBERT [41] and SalamBERT (https://huggingface.co/Fatemah/salamBERT) (accessed on 10 May 2021).

El Mahdaouy et al. [64] approached the same shared task using an end-to-end multitask learning model based on the MarBERT [65] language model. Duwairi et al. [66] investigated the ability of CNN, CNN-LSTM, and BiLSTM-CNN networks to detect hateful content on social media. These deep networks were trained and tested using the ArHS dataset, consisting of 9833 HS tweets annotated as racism, religious discrimination, abusive, and normal. The CNN model outperformed other models and achieved an accuracy of 81%. Moreover, recently suggested BERT-based Arabic named entity recognition models [67,68] could also be applied for HS recognition if trained on offensive and derogatory terms. Unlike the previous works, in this paper we propose a model trained on many DA offensive and hate speech datasets as well as MSA and evaluate a new pre-trained model to classify DA.

Table 1 summarizes the overview of the recent related works on English and Arabic offensive and hate speech detection. As we can notice, most of the related works on Arabic offensive and hate speech detection used single-task models. In terms of MTL models, authors tend to use two tasks at most and an external task or lexicon related to the context sentiment without accounting for context diversity in each used dataset and the learned context representations.

Table 1. Summary of the related works.

Paper	Task	Model	Evaluation	Dataset
		Hate and offensive speech in English		
Wassem et al. [42]	HS	Logistic regression classifier	73.93% F1-score	16 K samples annotated for HS
Gamback et al. [43]	HS	Convolutional neural network(CNN)	78% F1-score	16 K tweets Data provided by [42]
Pitsilis et al. [44]	HS	Single and multiple LSTM classifiers	87% recall 88% precision	16 K Data provided by [42]
Basile et al. [46]	HS	Support Vector Machine (SVM)	45.1% F1-score	SemEval 2019 (Task5)
Ping et al. [49]	OFF and HS	Pre-trained BERT model	82.9 F1-score	SemEval 2019
Liu et al. [50]	Multiple tasks	Multitask Learning (MTL)	/	Glue benchmark
		Hate and offensive speech in Arabic		
Albadi et al. [52]	HS	Lexicon-based classifier and SVM classifier	79% accuracy	6.6 K of religious HS tweets
Ousidhoum et al. [54]	HS	MTL	35% F1-score	13 K trilingual HS tweets
Mulki et al. [56]	OFF and HS	SVM and Naive Bayes classifier	83.6% F1-score	6 K of Tunisian HS tweets
Djandji et al. [57]	OFF and HS	MTL	90% F1-score	OSACT4 shared task 2000 tweets
Abu Farha et al. [59]	OFF and HS	CNN-BiLSTM and MTL	90.4% F1-score	OSACT4 shared task 2000 tweets
Hassan et al. [61]	OFF and HS	CNN, CNN-BiLSTM, and multilingual BERT	80.6% F1-score	OSACT4 shared task 2000 tweets
Otiefy et al. [62]	OFF	SVM	88.72% F1-score	SemEval2020 shared task
Husain et al. [63]	Sarcasm detection and SA	AraBERT and SalamBERT	69.22% F1-score	WANLP 2021 shared task 15.5 K tweets
El mahdaouy et al. [64]	Sarcasm detection and SA	AraBERT and MTL	74.8% F1-score	WANLP 2021 shared task
Duwairi et al. [66]	HS	CNN, CNN-LSTM, and CNN-BiLSTM	81% accuracy	9.8 K HS tweets

3. Proposed Model

Even with the emergence of DL techniques in many fields and real-world applications, DL still suffers from a few limitations such as data collection, annotation, and model complexity, which urges transfer learning and multi-task algorithms. Figure 2 shows the proposed MTL model architecture for tackling Arabic offensive and hate speech classification tasks. The model consists of two components: a shared part that contains the pre-processing and the pre-trained language model and a task-specific part.

In the architecture presented in Figure 2, the MTL framework incorporates variant datasets to learn shared and specific contextual representations simultaneously. Instead of training the MTL model on different tasks, we trained it on different datasets to cover the variation in the context, and annotated classes on a shared task. Fixing the target task and exposing the model to different contexts may help to enhance the classification performance and overcome single-task model issues.

As is shown in Figure 2, the processing part is similar to BERT data preparation, where the SentencePiece [69] algorithm is used for input segmentation based on a pre-trained neural-based tokenization vocabulary. The segmentation algorithm is a multi-layered RNN (recurrent neural netwrok) used to map each token to an embedding vector represented in word, segment, and positional (ID) embeddings. Furthermore, during the input segmentation [CLS] and [SEP] tokens will be padded to the word sequence at the beginning and end of the sentence, respectively. At this stage, the model receives an input representing the word sequence (tweet) from a dataset, which will be converted to a set of embedding vectors. Each word in the inputted sequence will be mapped to an embedding vector generated from summing up its corresponding word, segment, and positional embeddings. To learn shared global contextual representations across all corpora, the shared part is employed to fine-tune the weights of a pre-trained multilayer bidirectional transformer encoder such as AraBERT [41], and MarBERT [65]. The AraBERT and MarBERT are transformer encoders consisting of a self-attention mechanism to learn the

contextual representations for each inputted word. The transformer encoder architecture of BERT is shown in Figure 3, which consists of 12 transformer layers.

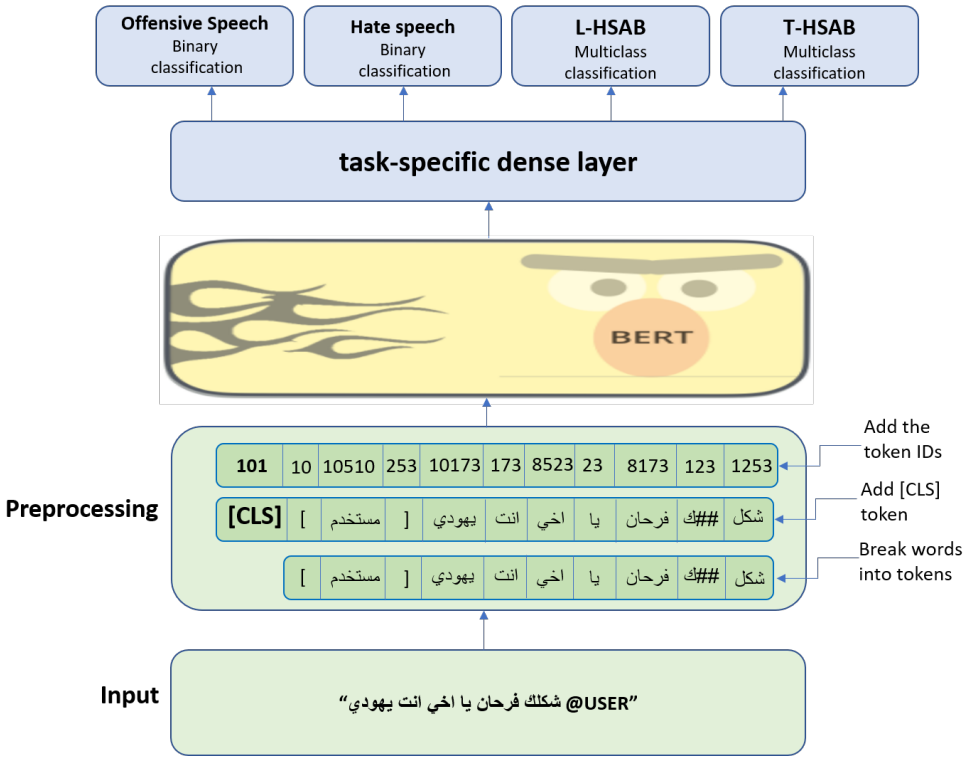


Figure 2. The architecture of the proposed MTL model.

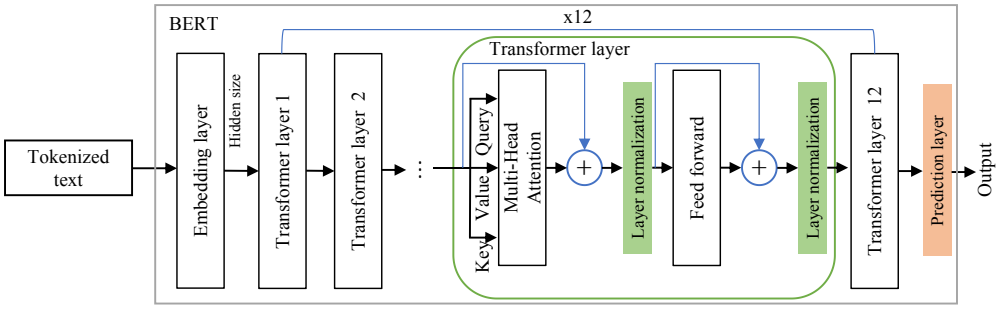


Figure 3. The architecture of the BERT model and its components.

Unlike BERT, which trains on a single task, the shared part containing the Arabic LM is trained using the combined loss from all datasets. The learned contextual embeddings for each word will be concatenated and stored in a single vector named [CLS]. Later, the [CLS] representing the semantic representation of the inputted sentence is fed to the task-specific part. The task-specific part acts as a single-sentence classifier on each dataset to classify their samples. The classifier is a fully connected layer followed by a softmax to estimate the probability of a sentence's contextual representation vector, which is labeled as class c as shown in Equation (1). The cross-entropy loss over the softmax output is used in our experiment to train the MTL model on both binary and multi-class classifications.

$$P_r(c|X) = softmax(W^T_{OSACT-OFF}.x), \quad (1)$$

where $W_{OSACT-OFF}$ is the dataset specific parameter matrix.

4. Experimental Setup

This section presents the experimental setup used in our study. In our experiments, we investigated the usage of pre-trained Arabic LMs on single task and MTL models. We experimented with different MTL models with variant datasets and pre-trained LMs during the training process.

4.1. Description of Datasets

We used three different Arabic offensive and hate speech datasets. OSACT is the dataset provided in the shared task, named the open-source Arabic corpora and corpora processing tools (OSACT) [70]. OSACT contains two datasets with the same context, including the OSACT-OFF for offensiveness with labels (OFF or NOT OFF) and OSACT-HS for hate speech with labels (HS or NOT HS). L-HSAB is the Levantine hate speech and abusive dataset collected by Mulki et al. [56]. T-HSAB is the first Tunisian hate speech and abusive dataset provided by Haddad et al. [71]. In the following sections, we will detail the statistics for each dataset. Table 2 summarizes the statistics of each dataset and the number of samples used during training and testing.

Table 2. Summary of the samples distribution in each dataset.

Dataset	**Language**	**Total Samples**	**Label**	**Training Set**	**Development Set**	**Test Set**
OSACT-HS	Arabic MSA Arabic DA	10 K	HS NOT HS	361 6639	44 956	101 1899
OSACT-OFF	Arabic MSA Arabic DA	10 K	OFF NOT OFF	1410 5590	179 821	402 1598
L-HSAB	Syrian DA Lebanese DA	6024	Abusive Hate Normal	1226 325 2539	258 74 544	243 67 567
T-HSAB	Tunisian DA	5846	Abusive Hate Normal	791 757 2668	166 160 578	169 161 574

4.2. Experiment Settings

The Adam optimizer [72], with a learning rate of 2×10^{-5}, was used in all experiments. We used HuggingFace's Transformers library [73] to utilize the pre-trained BERT models. Single task models and MTL models were trained for ten and five epochs, respectively. This is because training an MTL model on several tasks can take a longer time than single-task models.

4.3. Performance Measures and Models Training

Our experiments used the macro-average, which calculates metrics for each label, and finds their unweighted mean when the task is binary classification. In the case of multi-class classification tasks, the metrics are calculated for each class. Later, the average of the resulted scores is weighted by the number of true instances for each class.

Equation (2) calculates the accuracy of the TP (true-positive) and TN (true-negative) output when the model correctly predicts the positive and negative classes, respectively. On the other hand, the FP (false-positive) and FN (false negative) are the output, when the model incorrectly predicts the positive and the negative classes into negative and positive, respectively. The following formulas are used to determine other metrics such as F1-score, precision, and recall.

$$Accuracy(Acc) = \frac{TP + TN}{TP + TN + FP + FN} \times 100\%, \quad (2)$$

$$F1 = \frac{2 * Recall * Precision}{Recall + Precision} \times 100\%, \quad (3)$$

where

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

and

$$Recall = \frac{TP}{TP + FN}. \tag{5}$$

5. Results and Discussions

This section presents the conducted experiments and their results alongside discussion of the reported results. In the end, we compared the best-trained MTL models with recently developed models in the literature.

5.1. Experimental Series 1

This section will experiment with a range of models, including a single-task model and MTL models with three and four datasets, respectively. The model's setup is as follows:

- AraBERT v02: is a single-task model fine-tuned on both OSACT-OFF and OSACT-HS datasets separately;
- MTL-A-L and MTL-A-T: are MTL models with AraBERT used in the shared part, and OSACT-OFF, OSACT-HS, and T-HSAB are used in the specific task part;
- MTL-M-L and MTL-M-T: are MTL models with MarBERT covering Maghreb region dialect and MSA used in the shared part, and OSACT-OFF, OSACT-HS, and T-HSAB are used in the specific task part;
- MTL-AraBERT and MTL-MarBERT: are MTL models with AraBERT and MarBERT used in the shared part, respectively. In both models, all four datasets are used in the task part.

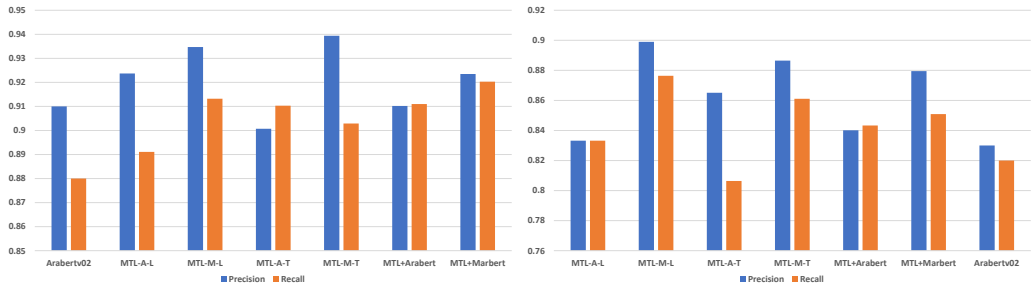
Table 3 reports the performance of different models using the macro F1-score and weighted F1-score for binary and multi-class classification, respectively. As can be noticed, MTL models outperform the single-task model in terms of accuracy and F1-score as listed for OSACT-OFF and OSACT-HS datasets. It can also be noticed that the MTL model trained using MarBERT boots the performance on OSACT-OFF, OSACT-HS, and L-HSAB datasets compared to the usage of AraBERT. This is due to the presence of dialectical Arabic in these datasets. The MTL models trained on three datasets in the specific-task part perform better than the MTL models trained on four tasks. This can be attributed to the data imbalance, the variation between L-HSAB and T-HSAB dialects, and the limited vocabulary presented in Arabic LMs.

Table 3. Performance of single-task and MTL models. Best results are shown in bold.

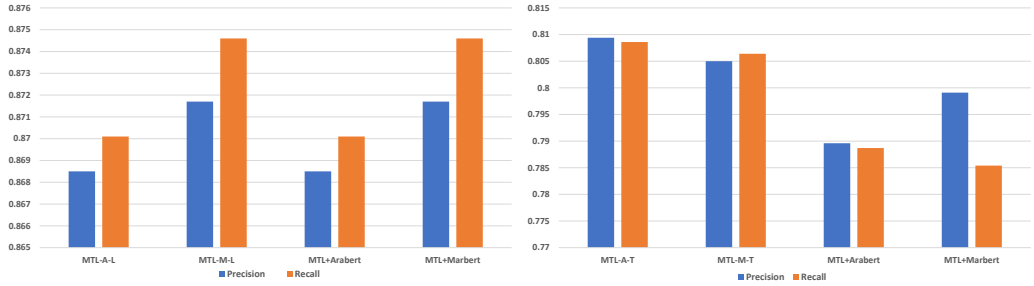
Model	OSACT-OFF *		OSACT-HS *		L-HSAB **		T-HSAB **	
	Acc (%)	F1 (%)	Acc (%)	F1 (%)	Acc (%)	F1 (%)	Acc (%)	F1 (%)
AraBERT v02	94	90.0	97	83.0	-	-	-	-
MTL-A-L	94.20	90.62	96.80	83.32	94.20	86.90	-	-
MTL-M-L	95.20	**92.34**	97.90	**88.73**	87.46	**87.18**	-	-
MTL-A-T	93.85	90.54	97.03	83.28	-	-	80.86	**80.50**
MTL-M-T	95.05	91.97	97.65	87.33	-	-	80.64	80.25
MTL-AraBERT	94.25	91.06	96.95	84.17	86.09	86.02	78.87	78.89
MTL-MarBERT	95.00	92.19	97.50	86.46	82.55	83.46	78.54	79.02

* Binary classification; * Multi-class classification (3 classes).

As shown in Figure 4, MTL-M-L has reported the highest precision and recall on three datasets, namely OSACT-OFF, OSACT-HS, and L-HSAB. On the other hand, MTL-A-T has the highest results on T-HSAB with 0.83 and 0.82 for precision and recall, respectively. Thus, MTL models trained on multiple datasets will give them the ability to share global and specific contextual representations from shared and specific-task parts, respectively.



(**a**) Models' performance on OSACT-OFF dataset. (**b**) Models' performance on OSACT-HS dataset.



(**c**) Models' performance on L-HSAB dataset. (**d**) Models' performance on T-HSAB dataset.

Figure 4. Precision and recall of the experimental series 1.

5.2. Experimental Series 2

This experiment compares our best-trained model, which reports the highest F1-score on all datasets, to the existing recent works in the literature. Table 4 shows that MTL-M-L and MTL-A-T outperformed existing single-task and MTL models on three out of four datasets, including OSACT-OFF, OSACT-HS, and L-HSAB. It can be noticed that our models did not achieve better results on T-HSAB compared to the results reported by Haddad et al. [71] due to the lack of sufficient samples for the Tunisian dialect presented in OSACT-OFF, OSACT-HS. The reported results show that training an MTL model on the same task using different datasets presenting different contexts, dialects, and classes can improve the contextual representation of the pre-trained LMs and lead to better results compared to single-task training.

Table 4. Comparison with the state-of-the-art models. Best results are shown in bold.

Model	OSACT-OFF *		OSACT-HS *		L-HSAB **		T-HSAB **	
	Acc (%)	F1 (%)	Acc (%)	F1 (%)	Acc (%)	F1 (%)	Acc (%)	F1 (%)
Abu farha et al. [59]	-	87.7	-	76	-	-	-	-
Djandji et al. [57]	-	90	-	82.28	-	-	-	-
Mulki et al. [56]	-	-	-	-	88.4	74.4	-	-
Haddad et al. [71]	-	-	-	-	-	-	87.9	**83.6**
MTL-M-L	95.20	**92.34**	97.90	**88.73**	87.46	**87.18**	-	-
MTL-A-T	93.85	90.54	97.03	83.28	-	-	80.86	80.50

* Binary classification; ** Multi-class classification (3 classes).

6. Conclusions

On social media, conflicts and violent behaviors become more explicit with every posted hate tweet or abusive content, affecting people's lives, especially in the Arab world. The automatic detection of these posts is challenging due to dialectal Arabic, which does not comply with any grammatical rules. We have developed a multitask learning model that incorporates two different pre-trained Arabic LMs based on Transformers, namely AraBERT and MarBERT. The choice of these LMs is to improve the learning of our model on MSA and DA and learn global contextual representation in the MTL shared part. In the task-specific part of the developed MTL model, the training is performed based on the variation

of the datasets as tasks to build our offensive and hate speech detection system. Compared to single-task learning models, the developed MTL model showed better classification performance and has outperformed existing models in the literature on three out of four evaluated datasets. The study opens the door for future research directions, where the MTL models can be extended to other applications and can exploit content diversity expressed in different datasets and lexicons. Meanwhile, experimenting with different dataset sources to improve the vocabulary coverage and learning better contextual representations is still an open problem, and is worth more investigation.

Author Contributions: Conceptualization, A.D. and M.A.A.A.-q.; data curation, W.A. and A.D.; formal analysis, A.D. and M.A.E.; funding acquisition, R.D.; investigation, R.D.; methodology, W.A., and M.A.E.; project administration, M.A.A.A.-q.; resources, W.A. and A.D.; software, W.A. and A.D.; supervision, M.A.A.A.-q.; writing—original draft preparation, W.A., and A.D.; writing—review and editing, M.A.A.A.-q., M.A.A., A.M.H. and R.D.; validation, M.A.E. and R.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by by LIESMARS Special Research Funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data used are open source as mentioned in the paper.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Castaño-Pulgarín, S.A.; Suárez-Betancur, N.; Vega, L.M.T.; López, H.M.H. Internet, social media and online hate speech. Systematic review. *Aggress. Violent Behav.* **2021**, *58*, 101608.
2. Chetty, N.; Alathur, S. Hate speech review in the context of online social networks. *Aggress. Violent Behav.* **2018**, *40*, 108–118.
3. Ul Rehman, Z.; Abbas, S.; Khan, M.A.; Mustafa, G.; Fayyaz, H.; Hanif, M.; Saeed, M.A. Understanding the language of ISIS: An empirical approach to detect radical content on twitter using machine learning. *Comput. Mater. Contin.* **2020**, *66*, 1075–1090.
4. Mladenovic, M.; Ošmjanski, V.; Stankovic, S.V. Cyber-Aggression, Cyberbullying, and Cyber-grooming. *ACM Comput. Surv.* **2021**, *54*, 1–42.
5. Cambria, E.; White, B. Jumping NLP curves: A review of natural language processing research. *IEEE Comput. Intell. Mag.* **2014**, *9*, 48–57.
6. Mäntylä, M.V.; Graziotin, D.; Kuutila, M. The evolution of sentiment analysis—A review of research topics, venues, and top cited papers. *Comput. Sci. Rev.* **2018**, *27*, 16–32.
7. Kapočiūtė-Dzikienė, J.; Damaševičius, R.; Woźniak, M. Sentiment analysis of Lithuanian texts using traditional and deep learning approaches. *Computers* **2019**, *8*, 4, doi:10.3390/computers8010004.
8. Kumar, A.; Dikshit, S.; Albuquerque, V.H.C. Explainable Artificial Intelligence for Sarcasm Detection in Dialogues. *Wirel. Commun. Mob. Comput.* **2021**, *2021*, 2939334.
9. Rivera-Trigueros, I. Machine translation systems and quality assessment: A systematic review. *Lang. Resour. Eval.* **2021**, doi:10.1007/s10579-021-09537-5.
10. Wani, T.M.; Gunawan, T.S.; Qadri, S.A.A.; Kartiwi, M.; Ambikairajah, E. A Comprehensive Review of Speech Emotion Recognition Systems. *IEEE Access* **2021**, *9*, 47795–47814.
11. Kapočiute-Dzikiene, J. A domain-specific generative chatbot trained from little data. *Appl. Sci.* **2020**, *10*, 2221.
12. Yang, J.A.; Tsou, M.H.; Janowicz, K.; Clarke, K.C.; Jankowski, P. Reshaping the urban hierarchy: Patterns of information diffusion on social media. *Geo-Spat. Inf. Sci.* **2019**, *22*, 149–165.
13. Lock, O.; Pettit, C. Social media as passive geo-participation in transportation planning–how effective are topic modeling & sentiment analysis in comparison with citizen surveys? *Geo-Spat. Inf. Sci.* **2020**, *23*, 275–292.
14. Kapočiūtė-Dzikienė, J.; Krilavičius, T. Topic classification problem solving for morphologically complex languages. In *International Conference on Information and Software Technologies*; Springer: Cham, Switzerland, 2016; Volume 639, pp. 511–524.
15. Mansoor, M.; Ur Rehman, Z.; Shaheen, M.; Khan, M.A.; Habib, M. Deep learning based semantic similarity detection using text data. *Inf. Technol. Control* **2020**, *49*, 495–510.
16. El-Kassas, W.S.; Salama, C.R.; Rafea, A.A.; Mohamed, H.K. Automatic text summarization: A comprehensive survey. *Expert Syst. Appl.* **2021**, *165*, 113679.
17. Kapočiūtė-Dzikienė, J.; Salimbajevs, A.; Skadiņš, R. Monolingual and cross-lingual intent detection without training data in target languages. *Electronics (Switzerland)* **2021**, *10*, 1412.
18. Islam, M.R.; Liu, S.; Wang, X.; Xu, G. Deep learning for misinformation detection on online social networks: A survey and new perspectives. *Soc. Netw. Anal. Min.* **2020**, *10*, 1–20.

19. Krilavičius, T.; Medelis, Ž.; Kapočiūtė-Dzikienė, J.; Žalandauskas, T. News media analysis using focused crawl and natural language processing: Case of Lithuanian news websites. In *International Conference on Information and Software Technologies*; Springer: Berlin/Heidelberg, Germany, 2012; pp. 48–61.
20. Tesfagergish, S.G.; Kapočiūtė-Dzikienė, J. Part-of-speech tagging via deep neural networks for northern-Ethiopic languages. *Inf. Technol. Control* **2020**, *49*, 482–494.
21. Neal, T.; Sundararajan, K.; Fatima, A.; Yan, Y.; Xiang, Y.; Woodard, D. Surveying stylometry techniques and applications. *ACM Comput. Surv.* **2017**, *50*, 1–36.
22. Venckauskas, A.; Karpavicius, A.; Damaševičius, R.; Marcinkevičius, R.; Kapočiūte-Dzikiené, J.; Napoli, C. Open class authorship attribution of lithuanian internet comments using one-class classifier. In Proceedings of the 2017 Federated Conference on Computer Science and Information Systems (FedCSIS), Prague, Czech Republic, 3–6 September 2017; pp. 373–382.
23. Tesfagergish, S.G.; Damaševičius, R.; Kapočiūtė-Dzikienė, J. Deep Fake Recognition in Tweets Using Text Augmentation, Word Embeddings and Deep Learning. In *International Conference on Computational Science and Its Applications*; Springer: Cham, Switzerland, 2021; Volume 12954 LNCS, pp. 523–538.
24. Žitkus, V.; Butkiene, R.; Butleris, R.; Maskeliunas, R.; Damaševičius, R.; Woźniak, M. Minimalistic Approach to Coreference Resolution in Lithuanian Medical Records. *Comput. Math. Methods Med.* **2019**, *2019*, 9079840 .
25. Behera, R.K.; Das, S.; Rath, S.K.; Misra, S.; Damasevicius, R. Comparative study of real time machine learning models for stock prediction through streaming data. *J. Univers. Comput. Sci.* **2020**, *26*, 1128–1147.
26. Shao, Z.; Sumari, N.S.; Portnov, A.; Ujoh, F.; Musakwa, W.; Mandela, P.J. Urban sprawl and its impact on sustainable urban development: A combination of remote sensing and social media data. *Geo-Spat. Inf. Sci.* **2021**, *24*, 241–255.
27. Xu, L.; Ma, A. Coarse-to-fine waterlogging probability assessment based on remote sensing image and social media data. *Geo-Spat. Inf. Sci.* **2021**, *24*, 279–301.
28. Amin, S.; Uddin, M.I.; Al-Baity, H.H.; Zeb, M.A.; Khan, M.A. Machine learning approach for COVID-19 detection on twitter. *Comput. Mater. Contin.* **2021**, *68*, 2231–2247.
29. Babić, K.; Petrović, M.; Beliga, S.; Martinčić-Ipšić, S.; Jarynowski, A.; Meštrović, A. COVID-19-Related Communication on Twitter: Analysis of the Croatian and Polish Attitudes. In *Proceedings of Sixth International Congress on Information and Communication Technology*; Springer: Singapore, 2022; Volume 216, pp. 379–390.
30. Habash, N.Y. Introduction to Arabic natural language processing. *Synth. Lect. Hum. Lang. Technol.* **2010**, *3*, 1–187.
31. Shaalan, K.; Siddiqui, S.; Alkhatib, M.; Abdel Monem, A. Challenges in Arabic natural language processing. In *Computational Linguistics, Speech and Image Processing for Arabic Language*; World Scientific: Singapore, 2019; pp. 59–83.
32. Darwish, K. Building a shallow Arabic morphological analyser in one day. In Proceedings of the ACL-02 Workshop on Computational Approaches to Semitic Languages, Philadelphia, PA, USA, 11 July 2002.
33. Ray, S.K.; Shaalan, K. A Review and Future Perspectives of Arabic Question Answering Systems. *IEEE Trans. Knowl. Data Eng.* **2016**, *28*, 3169–3190.
34. Guellil, I.; Saâdane, H.; Azouaou, F.; Gueni, B.; Nouvel, D. Arabic natural language processing: An overview. *J. King Saud Univ. Comput. Inf. Sci.* **2021**, *33*, 497–507.
35. MacAvaney, S.; Yao, H..; Yang, E.; Russell, K.; Goharian, N.; Frieder, O. Hate speech detection: Challenges and solutions. *PLoS ONE* **2019**, *14*, e0221152.
36. Fortuna, P.; Nunes, S. A survey on automatic detection of hate speech in text. *ACM Comput. Surv.* **2018**, *51*, 1–30.
37. Ayo, F.E.; Folorunso, O.; Ibharalu, F.T.; Osinuga, I.A. Machine learning techniques for hate speech classification of twitter data: State-of-The-Art, future challenges and research directions. *Comput. Sci. Rev.* **2020**, *38*, 100311.
38. Khairy, M.; Mahmoud, T.M.; Abd-El-Hafeez, T. Automatic Detection of Cyberbullying and Abusive Language in Arabic Content on Social Networks: A Survey. *Procedia CIRP* **2021**, *189*, 156–166.
39. Rosa, H.; Pereira, N.; Ribeiro, R.; Ferreira, P.C.; Carvalho, J.P.; Oliveira, S.; Coheur, L.; Paulino, P.; Veiga Simão, A.M.; Trancoso, I. Automatic cyberbullying detection: A systematic review. *Comput. Hum. Behav.* **2019**, *93*, 333–345.
40. Van Hee, C.; Jacobs, G.; Emmery, C.; DeSmet, B.; Lefever, E.; Verhoeven, B.; De Pauw, G.; Daelemans, W.; Hoste, V. Automatic detection of cyberbullying in social media text. *PLoS ONE* **2018**, *13*, e0203794.
41. Antoun, W.; Baly, F.; Hajj, H. Arabert: Transformer-based model for arabic language understanding. *arXiv* **2020**, arXiv:2003.00104.
42. Waseem, Z.; Hovy, D. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In Proceedings of the NAACL Student Research Workshop, San Diego, CA, USA, 12–17 June 2016; pp. 88–93.
43. Gambäck, B.; Sikdar, U.K. Using convolutional neural networks to classify hate-speech. In Proceedings of the First Workshop on Abusive Language Online, Vancouver, BC, Canada, 4 August 2017; pp. 85–90.
44. Pitsilis, G.K.; Ramampiaro, H.; Langseth, H. Effective hate-speech detection in Twitter data using recurrent neural networks. *Appl. Intell.* **2018**, *48*, 4730–4742.
45. Watanabe, H.; Bouazizi, M.; Ohtsuki, T. Hate Speech on Twitter: A Pragmatic Approach to Collect Hateful and Offensive Expressions and Perform Hate Speech Detection. *IEEE Access* **2018**, *6*, 13825–13835.
46. Basile, V.; Bosco, C.; Fersini, E.; Nozza, D.; Patti, V.; Rangel Pardo, F.M.; Rosso, P.; Sanguinetti, M. SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*; Association for Computational Linguistics: Minneapolis, MN, USA, 2019; pp. 54–63.

47. Cer, D.; Yang, Y.; Kong, S.Y.; Hua, N.; Limtiaco, N.; John, R.S.; Constant, N.; Guajardo-Céspedes, M.; Yuan, S.; Tar, C.; et al. Universal sentence encoder. *arXiv* **2018**, arXiv:1803.11175.
48. Zampieri, M.; Malmasi, S.; Nakov, P.; Rosenthal, S.; Farra, N.; Kumar, R. Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval). *arXiv* **2019**, arXiv:1903.08983.
49. Liu, P.; Li, W.; Zou, L. NULI at SemEval-2019 Task 6: Transfer Learning for Offensive Language Detection using Bidirectional Transformers. In *Proceedings of the 13th International Workshop on Semantic Evaluation*; Association for Computational Linguistics: Minneapolis, MN, USA, 2019; pp. 87–91.
50. Liu, X.; He, P.; Chen, W.; Gao, J. Multi-task deep neural networks for natural language understanding. *arXiv* **2019**, arXiv:1901.11504.
51. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
52. Albadi, N.; Kurdi, M.; Mishra, S. Are they Our Brothers? Analysis and Detection of Religious Hate Speech in the Arabic Twittersphere. In Proceedings of the 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), Barcelona, Spain, 28–31 August 2018; pp. 69–76.
53. Ashi, M.; Siddiqui, M.; Nadeem, F. Pre-trained Word Embeddings for Arabic Aspect-Based Sentiment Analysis of Airline Tweets. In *Advances in Intelligent Systems and Computing*; Springer International Publishing: Cham, Switzerland, 2019; pp. 241–251, doi:10.1007/978-3-319-99010-1_22.
54. Ousidhoum, N.; Lin, Z.; Zhang, H.; Song, Y.; Yeung, D.Y. Multilingual and Multi-Aspect Hate Speech Analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*; Association for Computational Linguistics: Hong Kong, China, 2019; pp. 4675–4684.
55. Ruder, S.; Bingel, J.; Augenstein, I.; Søgaard, A. Latent multi-task architecture learning. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp. 4822–4829.
56. Mulki, H.; Haddad, H.; Ali, C.B.; Alshabani, H. L-hsab: A levantine twitter dataset for hate speech and abusive language. In Proceedings of the Third Workshop on Abusive Language Online, Florence, Italy, 1 August 2019; pp. 111–118.
57. Djandji, M.; Baly, F.; Antoun, W.; Hajj, H. 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*; European Language Resource Association: Marseille, France, 2020; pp. 97–101.
58. Mubarak, H.; Darwish, K.; Magdy, W.; Elsayed, T.; Al-Khalifa, H. Overview of OSACT4 Arabic Offensive Language Detection Shared Task. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*; European Language Resource Association: Marseille, France, 2020; pp. 48–52.
59. Abu Farha, I.; Magdy, W. Multitask Learning for Arabic Offensive Language and Hate-Speech Detection. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*; European Language Resource Association: Marseille, France, 2020; pp. 86–90.
60. Abu Farha, I.; Magdy, W. Mazajak: An Online Arabic Sentiment Analyser. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*; Association for Computational Linguistics: Florence, Italy, 2019; pp. 192–198, doi:10.18653/v1/W19-4621.
61. Hassan, S.; Samih, Y.; Mubarak, H.; Abdelali, A.; Rashed, A.; Chowdhury, S.A. ALT Submission for OSACT Shared Task on 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*; European Language Resource Association: Marseille, France, 2020; pp. 61–65.
62. Otiefy, Y.; Abdelmalek, A.; El Hosary, I. WOLI at SemEval-2020 Task 12: Arabic Offensive Language Identification on Different Twitter Datasets. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation (Online)*; International Committee for Computational Linguistics: Barcelona, Spain, 2020; pp. 2237–2243.
63. Husain, F.; Uzuner, O. Leveraging Offensive Language for Sarcasm and Sentiment Detection in Arabic. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop (Virtual)*; Association for Computational Linguistics: Kyiv, Ukraine, 2021; pp. 364–369.
64. El Mahdaouy, A.; El Mekki, A.; Essefar, K.; El Mamoun, N.; Berrada, I.; Khoumsi, A. Deep Multi-Task Model for Sarcasm Detection and Sentiment Analysis in Arabic Language. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop (Virtual)*; Association for Computational Linguistics: Kyiv, Ukraine, 2021; pp. 334–339.
65. Abdul-Mageed, M.; Elmadany, A.; Nagoudi, E.M.B. ARBERT & MARBERT: Deep Bidirectional Transformers for Arabic. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2021; pp. 7088–7105, doi:10.18653/v1/2021.acl-long.551.
66. Duwairi, R.; Hayajneh, A.; Quwaider, M. A Deep Learning Framework for Automatic Detection of Hate Speech Embedded in Arabic Tweets. *Arab. J. Sci. Eng.* **2021**, *46*, 4001–4014.
67. Alsaaran, N.; Alrabiah, M. Arabic Named Entity Recognition: A BERT-BGRU Approach. *Comput. Mater. Contin.* **2021**, *68*, 471–485.
68. Boudjellal, N.; Zhang, H.; Khan, A.; Ahmad, A.; Naseem, R.; Shang, J.; Dai, L. ABioNER: A BERT-Based Model for Arabic Biomedical Named-Entity Recognition. *Complexity* **2021**, *2021*, 1–6.
69. Kudo, T.; Richardson, J. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv* **2018**, arXiv:1808.06226.
70. Husain, F. OSACT4 Shared Task on Offensive Language Detection: Intensive Preprocessing-Based Approach. *arXiv* **2020**, arXiv:2005.07297.

71. Haddad, H.; Mulki, H.; Oueslati, A. T-HSAB: A Tunisian Hate Speech and Abusive Dataset. In *Communications in Computer and Information Science*; Springer International Publishing: Cham, Switzerland, 2019; pp. 251–263, doi:10.1007/978-3-030-32959-4_18.
72. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
73. Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; et al. HuggingFace's Transformers: State-of-the-art natural language processing. *arXiv* **2019**, arXiv:1910.03771.