Proceedings of the OSACT 2022 Workshop @LREC2022, pages 176–180
Marseille, 20 June 2022

iCompass at Arabic Hate Speech 2022: Detect Hate Speech Using QRNN and Transformers

Mohamed Aziz Ben Nessir, Malek Rhouma, Hatem Haddad, Chayma Fourati
iCompass, 49, rue de Marseille, Tunis, Tunisia
mohamedaziz.bennessir@etudiant-isi.utm.tn, {malek, haddad}@gmail.com,
{haddad, chayma}@icompass.digital

Abstract

This paper provides a detailed overview of the system we submitted as part of the OSACT2022 Shared Tasks on Fine-Grained Hate Speech Detection on Arabic Twitter, its outcome, and limitations. Our submission is accomplished with a hard parameter sharing Multi-Task Model that consisted of a shared layer containing state-of-the-art contextualized text representation models such as MARBERT, AraBERT, ARBERT and task specific layers that were fine-tuned with Quasi-recurrent neural networks (QRNN) for each down-stream subtask. The results show that MARBERT fine-tuned with QRNN outperforms all of the previously mentioned models.

Keywords: Multi-Task,QRNN, MARBERT, Hate Speech Detection, Arabic

1. Arabic Hate Speech Detection

Hate Speech (HS) is particularly widespread in online communication due to users' anonymity and the lack of hate speech detection tools on social media platforms. Consequently, HS detection has determined a growing interest in using Machine/Deep Learning techniques to address this issue (Schmidt and Wiegand, 2017).

We describe our submitted system to the 2022 Shared Task Fine-Grained Hate Speech Detection on Arabic Twitter. We tackled the three subtasks, namely Detect whether a tweet is offensive or not (Subtask A), Detect whether a tweet has hate speech or not (Subtask B) and Detect the fine-grained type of hate speech (Subtask C). We used state-of-the-art pretrained contextualized text representation models and fine-tuned them according to the downstream subtasks in hand. As a first approach, we used the multilingual mT5 (Xue et al., 2020) and three Arabic Language models variants: AraBERT (Antoun et al., 2020), ARBERT (Abdul-Mageed et al., 2020) and MARBERT (Abdul-Mageed et al., 2020). The achieved performances on the development dataset showed that MARBERT outperforms all of the previously mentioned models overall, either on the three subtasks. In addition, we used the Quasi-recurrent neural networks (QRNN) (Stosic et al., 2016) model combined with MARBERT to achieve the best performances.

HS Detection tasks in Arabic are challenging ones because of the lack of the labelled data and the complexity of the Arabic language (Mulki et al., 2019; Haddad et al., 2019). In addition, Hate Speech is highly dependent on the culture, political and religious background and other aspects like Arabic dialects that are different from the Modern Standard Arabic (MSA). As the provided dataset is mainly based on dialect used in the area located in the Eastern Mediterranean, we used the Levantine Hate Speech and Abusive (L-HSAB) (Mulki et al., 2019) Twitter dataset as extra resources that we added to the provided training dataset.

Examples labelled labelled as normal, offensive, and hate from the OSACT Fine-Grained Hate Speech Detection dataset are presented in Table 1.

The paper is structured as follows: Section 2 provides a description of the OSACT Fine-Grained Hate Speech Detection dataset, the used external resource and the pre-processing step. Section 3 and section 4 describe the used pre-trained models and the quasi-recurrent neural network. Section 5 presents our submitted system description. Section 6 presents our development and test results compared to the baseline results provided by OSACT2022 Shared Tasks on Fine-Grained Hate Speech Detection on Arabic Twitter. Section 7 and 8 present the discussion and the conclusion with points to possible directions for future work.

2. Data Description

The provided training dataset of the OSACT Fine-Grained Hate Speech Detection task (Mubarak et al., 2022) is about 13k tweets, labelled with the 6 Hate Speech types: race/ethnicity/nationality, religion/belief, ideology, disability/disease, social class, and gender. 35% of the tweets are offensive and 11% are hate speech as shown in Table 2.

2.1. External Resources and Pre-processing

Levantine Hate Speech and Abusive (L-HSAB) dataset is a publicly available hate and abusive speech dataset collected from twitter and labeled with 3 types: 468 Hate speech, 1728 Offensive speech and 3650 Normal speech. In order to increase the size of the provided datasets, we manually relabelled samples from the L-HSAB (Mulki et al., 2019) and the samples have been used as extra resource.

Label	Example
Normal	بوين الحش ؟اذا فهمته هو واشكاله ان زمن جل
Offensive	هيلق و جحلط غير كذا مافي
Hate	شغالتك هي من حرر الكويت

Table 1: Examples from the OSACT Fine-Grained Hate Speech Detection dataset.

Type	Train	Dev.	Total
Offensive	3172	404	3576
Hate - Race	260	28	288
Hate - Religion	27	4	31
Hate - Ideology	144	14	158
Hate - Disability	0	1	1
Hate - Social Class	72	10	82
Hate - Gender	456	52	508
Normal	5715	866	6581

Table 2: Provided datasets Statistics.

After adding L-HSAB, we preformed multiple re-sampling strategies. Mainly focusing on over-sampling the minority type and under-sampling the majority type to prevent the model from over-fitting. Table 3 presents statistics of the final dataset used in our three subtasks submissions.

Type	Train	Dev	Total
Offensive	4155 (+984)	404	4559
Hate - Race	1644 (+1384)	28	439
Hate - Religion	64 (+37)	4	68
Hate - Ideology	195 (+51)	14	209
Hate - Disability	5 (+5)	1	6
Hate - Social Class	78 (+6)	10	88
Hate - Gender	471 (+15)	52	523
Normal	5715 (+0)	866	6581

Table 3: Final dataset statistics used in our three subtasks submissions.

2.2. Pre-Processing

Several pre-processing pipelines from intensive strategies like translating emojis to fairly light preprocessing and removing the English tokens were experimented with. The best performances were achieved when:

1. Removing all non Arabic tokens, including ones like USER, URL, $< LF >$. Emojis were also removed.
2. Normalizing all the hashtags by simply decomposing them.
3. Removing white spaces.

Table 4 presents examples before and after the preprocessing step.

3. Pre-trained Models

Different pre-trained models were used in order to achieve the best results when fine-tuning it in a multi-task fashion.

3.1. mT5

mT5 (Xue et al., 2020) is a massive multilingual pre-trained text-to-text transformer with 57B tokens Arabic tokens gather from 53M Arabic pages. The model leverages a unified text-to-text format and scale to attain state-of-the-art results on a wide variety of NLP tasks.

3.2. AraBERT

AraBERT (V2) (Antoun et al., 2020), is a BERT based model for Modern Standard Arabic Language understanding, trained on 70M sentences from several public Arabic datasets and news websites. It was fine-tuned on 3 tasks: Sequence Classification, Named Entity Recognition and Question Answering. It was reported to achieve state-of-the-art performances even on Arabic dialects after fine-tuning by (Abu Farha and Magdy, 2020).

3.3. ARBERT

ARBERT (Abdul-Mageed et al., 2020) is also a Bert based model trained on 61GB of Modern Standard Arabic text (6.5B tokens) gathered from books, news articles, crawled data and Wikipedia.

3.4. MARBERT

MARBERT (Abdul-Mageed et al., 2020) is a large-scale pretrained language model using the BERT base's architecture. MARBERT is trained on on 128 GB of tweets from various Arabic dialects containing at least 3 Arabic words. With very light preprocessing the tweets were almost kept at their initial state to retain a faithful representation of the naturally occurring text.

4. Quasi-recurrent Neural Network

Quasi-recurrent neural network (QRNN) (Stosic et al., 2016) represents an architecture that combines the sequential manner of treating the input tokens from Recurrent Neural Networks (RNNs) and the parallel processing fashion of Convolutional Neural Networks (CNNs) to allow a longer term dependency window while also addressing several issues faced when using both architectures separately. Stacked QRNNs are reported to have a better predictive accuracy than stacked LSTMs of the same hidden size. Figure 1 represents details of the QRNN architecture.

Before Pre-processing	After Pre-processing
#بكل_عنف_وقوة	بكل عنف و قوة
☺ ☺ URL وصارت فطاير البقالات غذاء صحي	وصارت فطاير البقالات غذاء صحي
☹ لا اثق القلوب متقلبة لا تثبت على حال RT USER	لا اثق القلوب متقلبة لا تثبت على حال

Table 4: Examples before and after pre-processing.

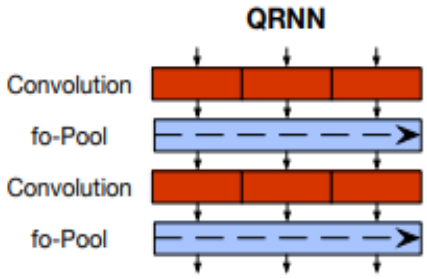

Figure 1: QRNN architecture.(Stosic et al., 2016)

5. System Description

The final submitted system is represented as in Figure 6.

5.1. Shared Layer

Preliminary results on the development dataset showed that a fine-tuned MARBERT achieved the best performances compared to the other language models. Hence, MARBERT was used as the shared part of the QRNN model and we focused our efforts on better squeezing out its performance by experimenting with different hyper-parameters values.

5.2. Subtasks Specific Layers

All of the three subtasks specific layers were essentially the same:

1. 1-dimensional convolution neural network with 128 units and a kernel size of 3.
2. 0.3 Dropout layer.
3. Bidirectional QRNN with 256 units.
4. 0.2 Dropout layers.
5. Dense layer with a Relu activation function and 64 units.

The architectures used for each subtask:

- Subask A: a dense layer with a Sigmoid activation function, 1 unit, and a threshold of 0.75.
- Subtask B and subtask C: a dense layer with a Softmax activation function and 7 units.

Given the fact that multi-task models learn better from closely related tasks, we added another output to penalize the model when it mistakes offensive comments for hate speech. This gave a 0.04 performance boost mainly for subtask C without much affecting other tasks.

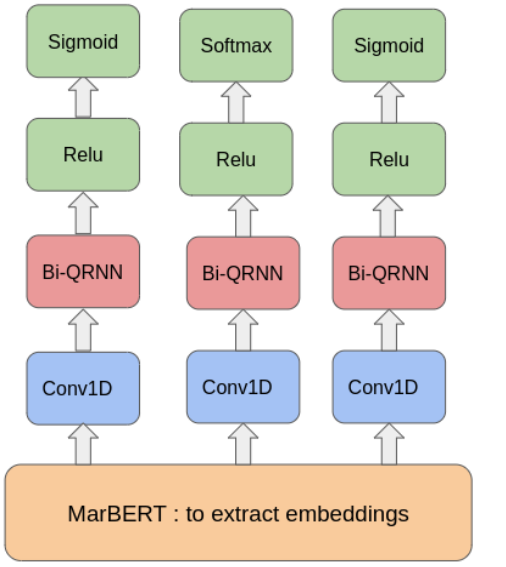

Figure 2: Submitted System Architecture.

6. Results

The submitted model was trained with a total of 16 epochs. The first 6 epochs were only used to warm up the QRNN layers, we froze MARBERT and trained them with a learning rate of 10-3, Adam optimizer, and a batch size of 350. As for the loss functions, we experimented with focal loss and categorical/binary crossentropy and submitted with categorical/binary crossentropy. In the last 10 epochs where we unfroze MARBERT, we lowered the learning rate to 10-4 keeping the same configuration.

Baseline results provided by OSACT2022 Shared Tasks on Fine-Grained Hate Speech Detection on Arabic Twitter (Mubarak et al., 2022) are presented in Table 5.

Subtask	Accuracy	Precision	Recall	F1-macro
Subtask A	0.651	0.325	0.5	0.394
Subtask B	0.893	0.447	0.5	0.472
Subtask C	0.893	0.128	0.143	0.135

Table 5: Baseline results on the development dataset.

Our results on the development datset with and without extra resources are presented in Table 6 where * refers to results after adding the extra resources.

Subtask	Accuracy	Precision	Recall	F1-macro
Subtask A*	0.825	0.855	0.812	0.845
Subtask A	0.848	0.808	0.834	0.819
Subtask B*	0.943	0.824	0.823	0.820
Subtask B	0.930	0.792	0.778	0.784
Subtask C*	0.931	0.558	0.559	0.555
Subtask C	0.918	0.382	0.480	0.734

Table 6: Results on the development dataset without and with using extra resources.

the confusion matrix for Subtask A on the validation is presented in figure 3

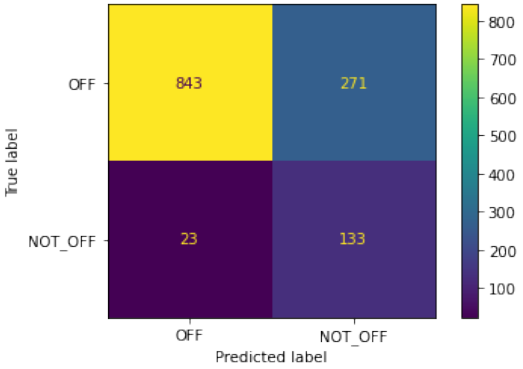

Figure 3: SubtaskA confusion matrix.

the confusion matrix for Subtask B on the validation is presented in figure 4

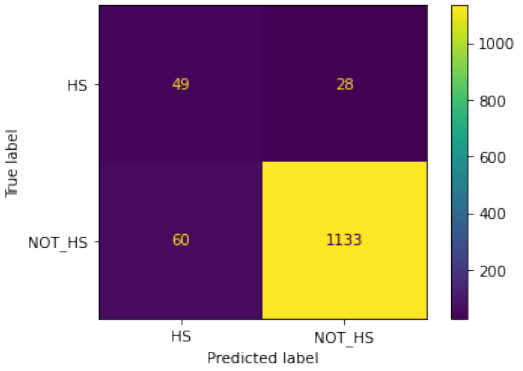

Figure 4: SubtaskB confusion matrix.

the confusion matrix for Subtask C on the validation is presented in figure 5
Our results on the test dataset are presented in Table 7.

7. Discussion

Different language models were used in this work. However, MARBERT achieved the best results. This

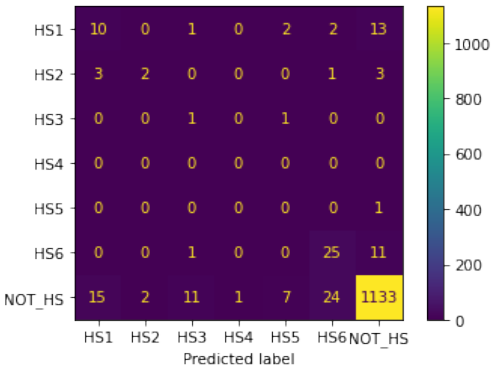

Figure 5: SubtaskC confusion matrix.

Subtask	Accuracy	Precision	Recall	F1-macro
Subtask A	0.854	0.841	0.837	0.839
Subtask B	0.941	0.869	0.801	0.831
Subtask C	0.919	0.548	0.531	0.528

Table 7: Results on the test dataset.

was the case because it was pre-trained on various Arabic dialects and therefore works better with dialectal data.

In addition, the data imbalance decreased the model performance. In fact, the training data set presents skewed class proportions. Relating to offensiveness, "Not Offensive" is the most frequent value with a count of 5,715 labels over a total of 8887. As for hate speech, the majority of samples fall under "Not Hate Speech" with a count of 7928 over 8887.

Category	Top Label	Label Frequency
OFF Label	NOT OFF	5715
HS label	NOT HS	7928
Vulgar label	NOT VLG	8753
Violence label	NOT VIO	8826

Table 8: Top labels and their frequencies in provided the training dataset.

The data imbalance in provided the training dataset further illustrated in Figure 7 .
Indeed, class proportions vary substantially, especially among hate speech classes. As illustrated in Table 2, there are 27 instances of "HS2" (i.e. hate speech based on religion) versus 456 instances of "HS6" (i.e. hate speech based on gender). In particular, there are no instances of "HS4" (i.e. hate speech based on disability). The data imbalance problem has a substantial effect on subtask C (i.e. multi-class classification) and explains the resulting relatively-low macro-averaged F1 score. Indeed, it stems from limited data relating to (from least to most available) hate speech based on disability/disease, hate speech based on religion/belief, and

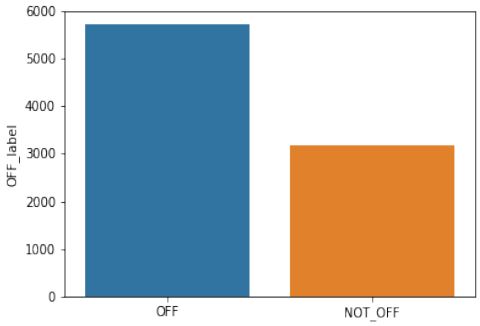

Figure 6: Offensive speech statistics.

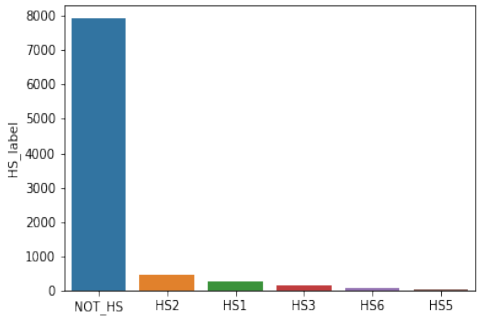

Figure 7: Hate speech statistics.

hate speech based on social class.

8. Conclusion

In this paper, we demonstrated how promising Multi-Tasking is for Hate & Abusive speech detection by fine-tuning the pre-trained model MARBERT with quasi-recurrent neural networks. Despite the small sized annotated data and the presence of different Arabic dialects, the model achieved satisfactory results.

With respect to the model, further work should explore meta-learning, Focal loss, semi-supervised learning, and ways to make use of violent and vulgar labels in the multi-task architecture.

As for the data, further work should focus on the need for disability data collection, disaggregation, and analysis. Indeed, persons with disabilities and their representative organizations must be at the core of data collection. The same goes for religious minorities in order to address the data gap.

9. Bibliographical References

Abdul-Mageed, M., Elmadany, A., and Nagoudi, E. M. B. (2020). Arbert & marbert: deep bidirectional transformers for arabic. *arXiv preprint arXiv:2101.01785*.

Abu Farha, I. and Magdy, W. (2020). 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*, pages 86–90, Marseille, France, May. European Language Resource Association.

Antoun, W., Baly, F., and Hajj, H. (2020). Arabert: Transformer-based model for arabic language understanding. *arXiv preprint arXiv:2003.00104*.

Haddad, H., Mulki, H., and Oueslati, A. (2019). T-hsab: A tunisian hate speech and abusive dataset. In *Proceedings of 7th International Conference on Arabic Language*, pages 1251–263, Nancy, France. Springer Nature.

Mubarak, H., Hassan, S., and Chowdhury, S. A. (2022). Emojis as anchors to detect arabic offensive language and hate speech. *arXiv preprint arXiv:2201.06723*.

Mulki, H., Haddad, H., Ali, C. B., and Alshabani, H. (2019). L-hsab: A levantine twitter dataset for hate speech and abusive language. In *Proceedings of the third workshop on abusive language online*, pages 111–118.

Schmidt, A. and Wiegand, M. (2017). A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10, Valencia, Spain, April. Association for Computational Linguistics.

Stosic, D., Stosic, D., Zanchettin, C., Ludermir, T., and Stosic, B. (2016). Qrnn: q-generalized random neural network. *IEEE transactions on neural networks and learning systems*, 28(2):383–390.

Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., and Raffel, C. (2020). mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.