Transfer Learning Across Arabic Dialects for Offensive Language Detection

Fatemah Husain
Dept. of Information Science
Kuwait University
Sabah AlSalem University City, Kuwait
f.husain@ku.edu.kw

Ozlem Uzuner
Dept. of Information Sciences and Technology
George Mason University
Fairfax, USA
ouzuner@gmu.edu

***Abstract*—The Arabic language is spoken by a wide range of countries across Asia, however, it is a low-resource language that has a minimal number of linguistic resources. Moreover, the large spread of Arabic speakers spans several countries and cultures, which creates a complex variation in its dialectal form. This variation makes it very challenging to analyze online Arabic content, particularly for offensive language detection. We propose a transfer learning approach for dialectal Arabic offensives language detection based on the BERT model. The results demonstrate the effectiveness of the proposed system in improving the performance of some Arabic dialects, such as the Tunisian and the Egyptian.**

***Index Terms*—Arabic dialects, Offensive language detection, text classification, BERT model, BERT customization**

I. Introduction

The Arabic language is widely used in Asia, especially in the Middle East region. In addition, Muslims around the world learn Arabic to study religious literature. Despite its importance to a large percentage of people, there are very limited numbers of linguistic resources and studies in Arabic Natural Language Processing (NLP). Accordingly, Arabic NLP researchers are still facing lots of complexity and difficulties.

For researchers in NLP, the most significant challenge when building an automatic linguistic solution to understand Arabic text is its varieties and inconsistencies. Spoken Arabic has more than six dialects, each has some vocabularies and terms that differentiate it from the other dialects. Moreover, the variety in Arabic dialects implies the disparities in the Arabic culture, as Arabic speakers from one region share values and cultures that might consider unacceptable to other Arabic speakers from another region in the Middle East. This challenge is strongly associated with applying NLP techniques to promote online social good within the Arabic sphere by supporting the compatibility of social media with the four metaverses for social good; accessibility, diversity, humanity, and equality [1]. All of these metaverses could be met when social media manages to limit the use of offensive language and hate speech from its content.

This study investigates the problem of offensive language while considering the challenges associated with the diversities and low available resources for Arabic dialects. The main goal of this paper is to develop a solution to Arabic offensive language detection by sharing knowledge gained from one dialect to another regarding offensive language while boosting the performance of the system.

The paper starts with some background information including Arabic dialects, transfer learning, the BERT model, and offensive language. Then, we highlight some of the previous studies that apply similar approaches to our approach and demonstrate how our works complement and differ from their works. The methodology is then explained in detail. Then, we discuss the results and findings from our experiments including a thorough error analysis section. The paper ends with a conclusion to emphasize the main discoveries.

II. Background

A. Arabic Dialects

The Arabic language has three main forms: Modern Standard Arabic (MSA), classical Arabic, and dialectal Arabic. The scope of this paper focuses on the dialectal Arabic form, which has several sub-dialects depending on the countries and regions within the Middle East. For instance, the noun خوصه / Khousa in the northern area of Saudi Arabia means a knife while in the middle area of Saudi Arabia, it is usually called سكين / Sikkeen. According to [1], Arabic dialects are grouped into seven main dialects: Egyptian, Levantine, Gulf, North African, Iraqi, Yemenite, and Maltese. Dialectal Arabic is the most dominant spoken form and most of the online user-generated content is also written in dialectal Arabic.

B. Transfer Learning

Previous studies report State-Of-The-Art (SOTA) performance results by applying transfer learning for many supervised NLP tasks (e.g., classification, information extraction, etc.) [3]. Most NLP tasks share common linguistic knowledge, such as linguistic representations and structural similarities in which transfer learning adds value by allowing tasks to inform each other (e.g., syntax and semantics). In addition, labeled datasets are rare and costly; thus, transfer learning can help make the most out of using as much supervision as available. Accordingly, adopting transfer learning for NLP tasks saves time and effort by enabling language processing models to be initialized with existing prior knowledge [4].

In [5], the author proposes a taxonomy for transfer learning in NLP based on two main branches. The first branch is

called transductive transfer learning, which combines datasets for the same NLP task with labeled data only in the source domain dataset. Transductive transfer learning includes domain adaptation in which two different domain datasets are used for the same task. For example, a system for question answering on news can be applied to a corpus of customer surveys. In addition, transductive transfer learning can be applied across different languages, so that knowledge learned from one language can serve the task for the other one.

The second branch of transfer learning is applied to different NLP tasks with labeled data in the target domain dataset and is called inductive transfer learning. Inductive transfer learning includes multi-task learning, which consists of having multiple different tasks running simultaneously. It also includes sequential transfer learning, which also has multiple tasks, but they are applied in sequential order. For example, learning word representation first (called pre-training) and then applying it to the target task such as offensive language detection (called adaptation). These types of transfer learning are not discrete, they could overlap.

C. The BERT Model

The BERT model is developed based on the sequential transfer learning approach. BERT stands for Bidirectional Encoder Representations from Transformers [3]. It is an innovative language model that presents SOTA results in multiple NLP tasks, such as question answering and language inference. BERT applies pre-trained language representations to downstream tasks through a fine-tuning approach. This approach is also called transfer learning, in which the pre-trained language representations are developed using a neural network model on a known task, and then fine-tuning is performed to use the same model for a new purpose-specific task. The main feature that distinguishes BERT from the other language modeling techniques is the use of a bidirectional language model rather than a unidirectional language model during the fine-tuning process. This bidirectional learning technique consists of a Masked Language Model (MLM) with a pre-training objective that randomly masks some of the tokens from the input to predict the original masked token based only on its context [3]. The training component of BERT consists of pre-training and fine-tuning phases. A long stream of continuous text is used to help the model learn semantic and syntactic long-term dependencies in the text that is used during pre-training. Considering the quality and the size of the data is very important during training.

D. Offensive Language

Providing a discrete definition of offensive (abusive) language is a very complicated task. Several factors play crucial roles in determining what is offensive and what is not offensive. Culture and personal experience are vital elements in explaining what is considered offensive language [6]. Most researchers borrow the definition of offensive language provided by [7] in which they describe offensive language as "threats and discrimination against people, swear words or blunt insults" (p.1). There are many forms of offensive language, including hate speech, aggressive content, cyberbullying, and toxic comments [8]. The use of offensive language can cause disturbance and affect online harmony. In addition, it can reduce user trust in the online platform [9]. Fig. 1 shows an example of an offensive tweet in the Egyptian dialect.

Fig. 1: Example of an offensive tweet in Egyptian dialect

III. Related Works

The AraBERT model has been used in multiple studies for Arabic offensive language detection. In [10] a multitask AraBERT approach is developed for offensive language detection and hate speech detection simultaneously, which can also be a solution to reduce the over-fitting effect occurred by the majority not offensive and not hate comments in most datasets. One dataset from Twitter is used to evaluate their approach. The results show positive effects on the model's performance, specifically, model generalizability over both tasks.

A sequential transfer learning approach is applied using the AraBERT model by fine-tuning the model first on several offensive language datasets, and then fine-tuning it again and evaluating it on sarcasm detection and sentiment analysis datasets to study the effect of sharing features from one task to another [11]. Results demonstrate an improvement in performance for the sentiment analysis task higher than the improvement in the sarcasm detection task.

In [12], several datasets covering different platforms and dialects were used to examine the effects of fine-tuning the AraBERT model using one dataset and evaluating it on another dataset. Results highlight the limited improvements in performance achieved by fine-tuning across different datasets, particularly for the highly dialectal text.

This study differs from the previous ones by emphasizing the differences among Arabic dialects, and by applying the same language model used in pre-training the AraBERT model to continue pre-train the model. This pre-training process greatly impacts the AraBERT model's vocabulary and the weight of its content.

IV. Methodology

Across dialects transfer learning is a way of estimating the generalizability of the model in terms of performance over several Arabic dialects. Thus, a selected set of dialectal datasets are used for continuing pre-training, fine-tuning, and evaluating the model to see how one dialect informs offensive language detection in other dialects.

In order to assess the information various dialects have about each other for offensive language detection, we build on the pre-trained AraBERT model. We continue pre-training the pre-trained model on the training set of one dialect, fine-tuning it to the training data of another dialect, and applying it to the test data of the fine-tuned dialect. We do this for all pairs of dialects. We also check the ability of multi-dialects datasets to recognize offensive language in individual dialects.

A. Datasets

Four dialectal datasets were used to evaluate the experiments. The corpora all contain tweets, as a way of controlling any effects of the platform on the transfer of models. The datasets that we utilize are L-HSAB [13] which contains Levantine tweets, Egyptian Tweets [14] dataset which contains Egyptian tweets, T-HSAB [15] which contains Tunisian, and OSACT [16] which contains tweets from a mix of dialects. We utilize the datasets which contain individual dialects to see how individual dialects inform each other. While utilizing the multi-dialects dataset to see if a wealth of dialects can support individual, resource-constrained dialects. Only binary classes are applied; offensive or not offensive. Thus, we convert different types of offensive language to offensive class. For example, the L-HSAB and T-HSAB datasets differentiate between hate and abusive language classes; which were both converted to offensive classes. We randomly apply an 80%-20% split to all datasets to create the train-test portions. Table I shows the distribution of each dataset. As can be seen, the Egyptian Tweets dataset is significantly smaller than the others.

TABLE I: Datasets distribution

Dataset (Dialect)	Train	Test
Egyptian Tweets (Egyptian)	880 tweets (not offensive = 353, offensive = 527)	220 tweets (not offensive = 100, offensive = 120)
T-HSAB (Tunisian)	4,819 samples (not offensive = 3,068, offensive = 1,751)	1,205 samples (not offensive = 752, offensive = 453)
L-HSAB (Levantine)	4,676 tweets (not offensive = 2,919, offensive = 1,757)	1,170 tweets (not offensive = 731, offensive = 439)
OSACT (Multi-dialect)	8,000 tweets (not offensive = 6,403, offensive = 1,597)	2,000 tweets (not offensive = 1,606, offensive = 394)

B. Preprocessing

Based on findings from previous research [17], which demonstrate that pre-processing does not improve the performance of the BERT models. In all experiments, all datasets were used without pre-processing.

C. Classification Model

1) The AraBERT Models: The experiments utilize the AraBERT (the first version of AraBERT, also called AraBERTv1-base and bert-base-arabert) model from the HuggingFace library [1]. The experiment is developed in Python using the PyTorch-Transformers library, and evaluation metrics were developed using the Scikit-Learn Python library. Google Colab Pro was used to conduct all experiments. The full experiment's code is available on GitHub [2].

2) Continue Pre-training: To continue pre-training AraBERT, the BertForMaskedLM class from Huggingface library is used. First, we create a LineByLineTextDataset from Huggingface library interface for reading and tokenizing the continued pre-training corpus with a block size of 128 words per line and AraBERT wordPiece tokenizer. Second, we use the DataCollatorForLanguageModeling class to create batches out of the corpus and specify 15% MLM probability for training. Third, we apply the TrainingArguments class to define the hyperparameters as the following: 5e-05 learning_rate, 1 epoch, and 32 batch size. Fourth, we call the trainer class to train the model. All hyperparameters were selected based on the recommended values by Huggingface tutorial[3].

3) Fine-Tuning: In all experiments, we apply the same experiment settings during fine-tuning to the continued pre-trained AraBERT models developed during the previous step in addition to the original AraBERT model: maximum length = 128, patch size = 16, epoch = 5, epsilon = 1e-8, and learning rate = 2e-5. We apply the BertForSequenceClassification BERT model, which has a linear layer for sentence classification (or regression) on top of the pooled output pooled output from the encoder to be used with a simple Feed Forward Neural Network (FFNN) layer to build the classifier.

D. Performance Evaluation

We use macro-averaged precision, recall, F1, and accuracy score to evaluate the classifiers' performance. Manual inspection and in-depth error analysis are also conducted to further evaluate the behavior of the models and their points of failure.

V. Results and Discussion

Table II provides a baseline performance of the original AraBERT model without further training, applied off the shelf to the three dialects. It reports the results for each dialect using its training set to fine-tune the model and its testing set to evaluate the model.

Table III shows the performances of dialectal AraBERT models, each built by continuing pre-training the AraBERT models on the training portion of the training dialect, fine-tuned to the training portion of the held-out dialect, and applied to the testing portion of the held-out dialect.

In general, Egyptian and Tunisian gain from continuing pre-training AraBERT on other dialects. Levantine consistently

[1]https://huggingface.co/aubmindlab/bert-base-arabertv01

[2]https://github.com/Fatemah-Husain

[3]https://huggingface.co/blog/how-to-train

TABLE II: Performance results of the individual dialectal models

Fine-tuning and Evaluating Dataset (Dialect)	Precision	Recall	F1	Accuracy
Egyptian Tweets (Egyptian)	0.86	0.66	0.66	0.68
T-HSAB (Tunisian)	0.79	0.77	0.78	0.80
L-HSAB (Levantine)	**0.86**	**0.86**	**0.86**	**0.87**

TABLE III: Performance results of dialectal continued pre-trained models

Pre-training Dialect	Fine-Tuning and Evaluating Dialect	Precision	Recall	F1	Accuracy
Egyptian Tweets (Egyptian)	T-HSAB (Tunisian)	0.79	0.77	0.78	0.80
	L-HSAB (Levantine)	0.83	0.83	0.83	0.84
T-HSAB (Tunisian)	Egyptian Tweets (Egyptian)	0.68	0.64	0.64	0.66
	L-HSAB (Levantine)	0.82	0.83	0.83	0.84
L-HSAB (Levantine)	Egyptian Tweets (Egyptian)	0.76	0.74	0.74	0.75
	T-HSAB (Tunisian)	0.77	0.77	0.77	0.79
Egyptian Tweets (Egyptian) and T-HSAB (Tunisian)	L-HSAB (Levantine)	0.85	0.85	0.85	0.86
Egyptian Tweets (Egyptian) and L-HSAB (Levantine)	T-HSAB (Tunisian)	0.79	0.78	0.79	0.80
T-HSAB (Tunisian) and L-HSAB (Levantine)	Egyptian Tweets (Egyptian)	0.72	0.70	0.70	0.72
Concatenation of all three dialectal datasets	Egyptian Tweets (Egyptian)	0.69	0.68	0.68	0.69
	T-HSAB (Tunisian)	0.80	0.77	0.78	0.80
	L-HSAB (Levantine)	0.84	0.84	0.84	.85

TABLE IV: Performance results from multi-dialects continued pre-trained models

Pre-training Dialect	Fine-Tuning and Evaluating Dialect	Precision	Recall	F1	Accuracy
OSACT (Multi-dialects)	Egyptian Tweets (Egyptian)	0.56	0.51	0.36	0.47
	T-HSAB (Tunisian)	0.66	0.57	0.54	0.66
	L-HSAB (Levantine)	0.73	0.56	0.50	0.66
OSACT (Multi-dialects) and all three dialectal datasets	Egyptian Tweets (Egyptian)	0.78	0.78	0.78	0.78
	T-HSAB (Tunisian)	0.79	0.80	0.80	0.81
	L-HSAB (Levantine)	0.86	0.87	0.87	0.87

benefits the other dialects but does not gain in performance by training on other dialects. This finding is in line with the findings from [18], which highlight the reduction in test's effectively for small test sets because of the test's assumption that the test set distribution does not differ significantly from the population distribution. Overall, the results highlight the positive impact of transferring knowledge across Arabic dialects.

Table IV below shows the results of continued pre-training AraBERT using the OSACT dataset to evaluate whether the use of multi-dialects can support individual, resource-constrained dialects.

As can be noticed from Table IV, the results report significant improvement for all three dialects' evaluation performance after adding the training sets from each dialectal dataset to the multi-dialects corpus to continue pre-train AraBERT.

A. Error Analysis

Table V, table VI, and table VII in the appendix show samples of tweets from each dataset with their actual and predicted labels by each model.

To better understand the variations among each model's result, table VIII and table IX in the appendix depict the behavior of each model and how they are similar and different in understanding the data. Each cell in the table contains samples that are misclassified by its intersecting column and row's models and correctly classified by the other models. This analysis supports identifying points of failure for each model.

Generally, misclassified samples contain several types of nouns; names of figures, animals, countries, organizations, etc. Moreover, incorrectly labeled samples from the Tunisian dataset (T-HSAB) were detected, such as "المغرب الكبير الدول العربية والأعجمية يجب فتحها للأسف / The Grand Morocco, Arab and foreign countries, must unfortunately be conquered", which is labeled as offensive and was misclassified by concatenated model and the Levantine model. Similarly, the Levantine dataset (L-HSAB) includes some incorrectly annotated samples, for instance, "انت واشكالك عباد المال تبحثون عن الازمات لجل مصالح شخصيه / You and your fellows are money slaves are looking for crises for personal interests", labeled as not offensive. Previous studies also highlighted the same issues with the datasets [19].

The Tunisian model fails in classifying simple offensive Levantine tweets, containing explicit offensive words, which were correctly classified by all other models, such as "طبعا الكلاب معروفي مين مقصودة / Of course dogs know who is meant".

The findings from the analysis demonstrate that the Egyptian model and the Tunisian model demonstrate limited performance in classifying the long Levantine samples. Similarly, the Levantine model misclassifies short Tunisian samples, containing one or two words only, such as "برا شيط / go shit yourself", while the other models were better in classifying

similar short samples.

Conclusion

Developing technical solutions to process the Arabic language automatically is a very challenging task due to the variability and complexity of the Arabic content. Particularly for the Arabic text on social media, which consists of different dialects and cultures. This paper proposes a transfer learning approach across Arabic dialects for offensive language detection. The experiments report variations in the results. For the Tunisian and Egyptian dialects, they gain better performance by applying our proposed transfer learning approach than the Levantine dialect. The findings demonstrate the potential of our approach in improving the performance of Arabic offensive language detection. We recommend further investigating our approach on larger offensive language datasets covering more Arabic dialects to better understand its benefits and effects on the model's performance.

References

[1] H. Duan, J. Li, S. Fan, Z. Lin, X. Wu, and W. Cai, "Metaverse for Social Good: A University Campus Prototype," Proceedings of the 29th ACM International Conference on Multimedia, October 2021, https://doi.org/10.1145/3474085.3479238

[2] N. Habash, "Introduction to Arabic Natural Language Processing," Synthesis Lectures on Human Language Technologies, 3(1), pp. 1–187. Morgan and Claypool Publishers, 2010.

[3] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova,"BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," 2018. arXiv preprint arXiv:1810.04805.

[4] P. Azure, "Transfer Learning for Natural Language Processing," ISBN 9781617297267, 2021.

[5] S. Ruder, "Neural Transfer Learning for Natural Language Processing," (Doctoral dissertation, NUI Galway), 2019.

[6] B. Vandersmissen, "Automated detection of offensive language behavior on social networking sites," Ghent University, 2012.

[7] G. Wiedemann, E. Ruppert, R. Jindal, and C., "Biemann Transfer Learning from LDA to BiLSTM-CNN for Offensive Language Detection in Twitter," Proceedings of GermEval 2018, 14th Conference on Natural Language Processing (KONVENS 2018), 2018.

[8] A. Schmidt, and M. Wiegand, "Survey on Hate Speech Detection Using Natural Language Processing," Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media, pp.1–10, Association for Computational Linguistics (ACL), Valencia, Spain, 2017.

[9] A. Founta, D. Chatzakou, N. Kourtellis, J. Blackburn, A. Vakali, and I. Leontiadis, " A Unified Deep Learning Architecture for Abuse Detection," Proceedings of the 10th ACM Conference on Web Science (WebSci '19), pp.105–114, Association for Computing Machinery, New York, NY, USA, 2018.

[10] M. Djandji, F. Baly, and H. Hajj, "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, May 2020.

[11] F. Husain and O. Uzuner, "Leveraging Offensive Language for Sarcasm and Sentiment Detection in Arabic," Proceedings of the Sixth Arabic Natural Language Processing Workshop, pp. 364–369, Association for Computational Linguistics, April 2021.

[12] F. Husain and O. Uzuner, "Fine-Tuning Approach for Arabic Offensive Language Detection System: BERT-Based Model," Future Technologies and Innovations (FTI) Proceedings: 4th International Conference on Computer Applications and Information Security (ICCAIS'2021), pp. 1–5, March 2021.

[13] H. Mulki, H. Haddad, C. Bechikh Ali, and H. Alshabani, "L-HSAB: A Levantine Twitter Dataset for Hate Speech and Abusive Language," Proceedings of the Third Workshop on Abusive Language Online, pp. 111–118, Association for Computational Linguistics (ACL), 2019, https://doi.org/10.18653/v1/W19-3512.

[14] H. Mubarak, K. Darwish, and W. Magdy, "Abusive Language Detection on Arabic Social Media," Proceedings of the First Workshop on Abusive Language Online, pp. 52–56, Association for Computational Linguistics (ACL), Vancouver, BC, Canada, August 2017, https://doi.org/10.18653/v1/w17-3008.

[15] H. Haddad, H. Mulki, and A. Oueslati, "T-HSAB: A Tunisian Hate Speech and Abusive Dataset, In International Conference on Arabic Language Processing," pp. 251–263, Springer, Cham, October 2019.

[16] H. Mubarak, K. Darwish, W. Magdy, T. Elsayed, and H. Al-Khalifa, "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, pp. 48–52, May 2020.

[17] F. Husain and O. Uzuner, "Investigating the Effect of Preprocessing Arabic Text on Offensive Language and Hate Speech Detection," ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP), vol. 21, pp. 1–20, July 2022.

[18] R. Dror, G. Baumer, S. Shlomov, and R. Reichart, "The Hitchhiker's Guide to Testing Statistical Significance in Natural Language Processing," In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Vol.1, pp. 1383–1392, Association for Computational Linguistics (ACL), 2018.

[19] F. Husain and O. Uzuner, "A survey of offensive language detection for the Arabic language," ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP), vol. 20, pp. 1–44, April 2021.

Appendix

This section includes tables discussed in the Error Analysis section. It illustrates examples of output from each model we applied and compares all model's outputs to provide an in-depth analysis of the model's behavior.

TABLE V: Sample tweets with their actual and predicted labels by models for the Egyptian Tweets dataset (W: wrong prediction, C: correct prediction)

Tweet	Actual Label	Baseline AraBERT	Levantine Model	Tunisian Model	Levantine and Tunisian Model	Concatenated Model	Multi-Dialect and Three Dialects
التهنئة تكون للمسلمين اللذين لايعادون الاسلام ويشوهون صورته وأهله، أما أنت فقد بعت دينك بغرض من الدنيا،إلى جهنم وبئس المصير Congratulations are given to Muslims who do not hate Islam and distort its image and its people. As for you, you sold your religion for the purpose of this world, go to hell and misery of fate	Offensive	W	C	C	C	C	C
اهدى بس عشان أول لما ها يدعوا لتعديل الدستور أنت أول واحد هاتقول نعم "Be quiet, first of all, when they call for amending the constitution, you will be the first one who will say yes."	Offensive	C	W	C	C	C	C
محمد البرادعي رمز افتخر به ولكن الاغلب ولكن الاغلب يجهلون مواقف هذا الرجل العظيم Mohamed El- Baradei is a symbol I am proud of, but most are ignorant of the positions of this great man.	Not Offensive	C	C	W	C	C	W
ما هو لو كام موجود علي الساحه المصريه دلوقت كنت مش هصدقه Had he been on the Egyptian scene, I would have not believed him	Not Offensive	W	C	C	W	C	C
ومتى تخلعون الكذب والنفاق عن وجوهكم When you put off lying and hypocrisy from your faces	Offensive	C	C	C	C	W	W
الله ينصرك ربي ويعزك ويرفعك انت وكل مخلص قابع بالمعتقلات ظلما وزورا May God grant you victory, may my Lord comfort you and exalts you and every sincere person lying in the detention camps unjustly and unfairly	Not Offensive	W	W	W	W	W	W
انزل يا حمدين واوعدك حتكون شهيد وتقابل الشهداء وحياه امي لو شفتك لعملها "Come down, Hamdeen, and I promise you that you will be a martyr and meet the martyrs. I swear by my mother, I will do that."	Offensive	C	C	C	C	C	C

TABLE VI: Sample tweets with their actual and predicted labels by models for the L-HSAB (Levantine) dataset (W: wrong prediction, C: correct prediction)

Tweet	Actual Label	Baseline AraBERT	Egyptian Model	Tunisian Model	Egyptian and Tunisian Model	Concatenated Model	Multi-Dialect and Three Dialects
ما حدا باعصك الا اللواء No one other than the Major General is hurting you	Offensive	W	C	C	C	C	C
خازوق فيك ولي بشد عمشدك عم بيحكي عن جعجع شو خص سعد الحريري ولا عملكن عقدة Screw you and people who think like you. He was talking about Geagea and not Saad Hariri, but you guys want any excuse to blame Hariri.	Offensive	C	W	C	C	C	C
لنك رخيص Because you are worthless	Offensive	C	C	W	C	C	C
قل كلمتك وامشي ولايهمك كلام الحاقدين والمتزلفين كن مع الحق ولاتبالي Say your word and walk, and do not care about haters' talk. Be with the truth and do not care	Not Offensive	C	C	C	W	C	C
هلق كلهم ضرب عالفلسطينيين فليش التدقيق All hit are on the Palestinians, why caring	Not Offensive	C	C	C	C	W	W
كملة الاخوة كتّاب العدل بيطلعلهم وزير بالحكومة؟ عقدة بيت نعيم May God grant you victory, may my Lord comfort you and exalts you and every sincere person lying in the detention camps unjustly and unfairly	Not Offensive	W	W	W	W	W	W
انضب يا حقير منعرفكن كلاب مين يلعنك واطي "Come down, Hamdeen, and I promise you that you will be a martyr and meet the martyrs. I swear by my mother, I will do that."	Offensive	C	C	C	C	C	C

TABLE VII: Sample tweets with their actual and predicted labels by models for the for the T-HSAB (Tunisian) dataset (W: wrong prediction, C: correct prediction)

Tweet	Actual Label	Baseline AraBERT	Levantine Model	Egyptian Model	Levantine and Egyptian Model	Concatenated Model	Multi-Dialect and Three Dialects
تعمل وشعب تونسي بهيم هههههههه Working and Tunisian people are bad Hahaha	Offensive	W	C	C	C	C	C
خرجتو زين العبدين وجبتو عميل الماسون The Masonic agent went out and got Zain Al-Abideen	Not Offensive	C	W	C	C	C	C
يعيشكم خاوتي تونسة ممكن تشرفوني قناتي ومردودين Long live my Tunisian brothers, can you honor my channel and return	Not Offensive	C	C	W	C	C	C
لوكان جات بلحق نضرة الإسلام ماكنتيش تحكي ياغبيا عفوا مقصودي بيا If he had come with the freshness of Islam, you wouldn't say I am sorry," excuse me	Not Offensive	C	C	C	W	C	C
اكبر كذابة عرفتها البشرية حاسبة روحها حاجة ركي متسوايش بصلة يل جسوسة منافقة The biggest liar known to mankind, You are nothing	Offensive	C	C	C	C	W	W
برافو بشير حق المرة فاتت هوما فتو المرأة يداويها راجل عطونا فتوة جديدة تنو صوت المرأة عورة ربي يهديهم Bravo Bashir, the truth of the time. The woman's fatwa, was healed by a man who gave us a new fatwa	Offensive	W	W	W	W	W	W
تكفيري فخ العياذ با وجود للسلفية الجهادية العلمية غيرهما السلفية هدي النبي صلى وسلم خربتم ديننا بارك Radical Takfiri, God forbid	Offensive	C	C	C	C	C	C

TABLE VIII: Misclassified samples across the models (Part 1)

Models	Egyptian Model	Tunisian Model	Levantine Model	Concatenated Model
Egyptian Model	**Levantine dataset:** لو في دولة محترمة حالا ما كنا صوتك منسمعوع قولة زياد الرحباني مجرص الحزب فيك If we were living in a respectable country, people like you wouldn't have a voice. As Ziad Rahbani said: you are embarrassing the political party you adhere to. *(Offensive predicted as not offensive)* **Tunisian:** حرية لعرا توانسا احرار متربين تفوو Freedom for nacked, Tunisian, free, educated people *(Offensive predicted as not offensive)*	**Levantine dataset:** ليدي ديما صادق الله يسعدك حبيبتي وين ما كنت وبحب قلك خليها حلقه يا جبل ما يهزك ريح تحياتي القلبية الك Lady Dima Sadiq, may God be pleased with you, my love, where were you, and with the love of your heart. *(Not offensive*	**Tunisian dataset:** عنصرية كانهم ليسو أفارقة إفريقي حني تكن أسمر هكدا براك الاخر Racist, as if they were not African Africans when you were showing brown *(Offensive predicted as not offensive)*	**Tunisian dataset:** يامريم نكره برا اتلهى بصغارك خير تافها Maryam, we hate righteousness,its better to play with your children *(Offensive predicted as (Offensive predicted as not offensive)* **Levantine:** Not available
Tunisian Model	**Levantine dataset:** خدو هالتغريده وروحوا نامو انتزعة السهرة الهارب من العدالة يحي وزير العدل خبصها مبارح هاجم العدل والعدالة والداخلية I'll tweet this and sleep as my evening has been ruined. An outlaw just became the minister of justice. He messed up yesterday by attacking judges and the ministry of interior. *(Not offensive predicted as offensive)*	**Egyptian dataset:** من انحرافات صلاح نصر الى مخابرات الشتم و السب و تربية الشراميط. From the deviations of Salah Nasr to the intelligence of insulting and raising sluts. *(Offensive predicted as not offensive)* **Levantine:** طبعا الكلاب معروفي مين مقصودة Of course dogs know who is meant *(Offensive predicted as not offensive)*	**Egyptian dataset:** ابثو قابلوني see me *(Offensive predicted as not offensive)*	**Levantine dataset:** جاسوس برتبة وزير A spy with the rank of Minister *(Offensive classified as not offensive)* **Egyptian:** علماء السوء Bad scholars *(Not offensive predicted as offensive)*
Levantine Model	**Tunisian dataset:** لطفي العبدلي يمثلني Lotfi Al-Abdali represents me *(Not offensive predicted as offensive)*	**Egyptian dataset:** مصر بتراثها و تاريخها اختزلت في السيسي Egypt with its heritage and history reduced to Sisi *(Not offensive predicted as offensive)*	**Egyptian dataset:** تحيا مصر امالدنيا يارب مصر دايما من نصر لنصر احفظ بلادنا يارب واملاها من خيرك Long live Egypt # Mother of the World, Lord of Egypt, always from victory to victory. *(Not offensive predicted as offensive)* **Tunisian dataset:** برا شيط go shit yourself *(Offensive predicted as not offensive)*	**Tunisian dataset:** المغرب الكبير الدول العربية والأعجمية يجب فتحها للأسف The Grand Morocco, Arab and foreign countries, must unfortunately be conquered *(Offensive predicted as not offensive)* **Egyptian dataset:** نكبة ليه بس هو إحنا بنستمد الشعارات والطهارة غير منها وهى ورقة التوت اللى بنغطى بيها مساؤنا وأفعالنا دى نعمة لكتير مش نكبة Why a catastrophe, but it is that we derive slogans and purity other than from it, and it is the berries leaf with which we cover our misfortunes and our actions, this is a blessing for many, not a calamity *(Offensive predicted as not offensive)*

TABLE IX: Misclassified samples across the models (Part 2)

Models	Egyptian Model	Tunisian Model	Levantine Model	Concatenated Model
Concatenated Model				**Tunisian dataset:** يسترعلى الامه الاسلاميه ربي ويطرح مجلس النواب شرك شرك شرك كبيروخروج ديننا الاسلامي The Islamic Ummah is covered up, my Lord, and the House of Representatives proposes a major polytheism, and the exit of our Islamic religion *(Not offensive predicted as offensive)* **Levantine dataset:** الاسد لا حضور له مع الخراف The lion (Al-Assad) does not attend with the sheep *(Offensive predicted as not offensive)* **Egyptian dataset:** للاسف الميليشيات العفنه ضيعت القضية.الفلسطينية !! Unfortunately, the rotten militias lost the Palestinian conflict. *(Offensive predicted as not offensive)*