Offensive Language Detection in Arabizi

Imene Bensalem
ESCF de Constantine
MISC Lab, Constantine 2
University, Algeria
ibensalem@escf-constantine.dz

Meryem Ait Mout
Polytech Marseille,
Aix-Marseille Université,
France
meryem.AIT-MOUT@etu.univ-amu.fr

Paolo Rosso
Universitat Politècnica
de València,
Spain
prosso@dsic.upv.es

Abstract

Detecting offensive language in under-resourced languages presents a significant real-world challenge for social media platforms. This paper is the first work focused on the issue of offensive language detection in Arabizi, an under-explored topic in an under-resourced form of Arabic. For the first time, a comprehensive and critical overview of the existing work on the topic is presented. In addition, we carry out experiments using different BERT-like models and show the feasibility of detecting offensive language in Arabizi with high accuracy. Throughout a thorough analysis of results, we emphasize the complexities introduced by dialect variations and out-of-domain generalization. We use in our experiments a dataset that we have constructed by leveraging existing, albeit limited, resources. To facilitate further research, we make this dataset publicly accessible to the research community.

1 Introduction

Due to the unrestricted nature of online discourse, offensive language has found its way to social media platforms, which poses major challenges for maintaining a respectful and inclusive virtual environment. Processing social media texts in Arabic presents its own set of challenges, as the user-generated content in this language is often not written in standard Arabic, but instead in multiple dialects, which vary from one country to another and have no grammatical and orthographic rules. Additionally, the use of Arabizi[1] (Alghamdi and Petraki 2018; Brabetz 2022; Haghegh 2021; Yaghan 2008)–an informal system of writing Arabic using Latin alphabet and numbers, which is commonly blended with French and English–further complicates Arabic processing (Darwish 2014).

Arabizi is characterized by the numerous transliterations of a single word, which may create a new set of homonyms within Arabic and even with other languages[2]. For example, in the dataset used for this research, we were able to find 7 different Arabizi spellings of the word قلب (*heart*), which are *alb, aleb, 9alb, kalb, galb, guelb, gelb*. The 2 first ones have been found in the Lebanese dialect, whereas the rest are used in the Algerian dialect, showing different pronunciations across regions.

Due to this inconsistency in writing this vernacular digital Arabic, traditional offensive language detection methods may struggle to interpret accurately the Arabizi words and expressions unique to each dialect. To illustrate, the word *kalb*, listed above as one of the spelling forms of قلب (*heart*), is also the transliteration of كلب (*dog*), which is used, in addition to its literal meaning, as an insult in the Arab world.

Arabizi has been studied in various contexts, such as its identification and transliteration to Arabic (Darwish 2014; Shazal et al. 2020), code-switching detection (Shehadi and Wintner 2022), POS tagging (Muller et al. 2020) and sentiment analysis (Fourati et al. 2021; Guellil et al. 2021). Besides, there has been a notable increase in the number of papers focusing on Arabic offensive language detection in recent years (Husain and Uzuner 2021). Nonetheless, there is a scarcity of research dedicated to handling Arabizi specifically within the context of offensive language detection.

This paper is dedicated to addressing this gap. Our contributions are the following:

- We provide, for the first time, a

[1] Also knows as Romanized Arabic and Arabic chat alphabet.

[2] Shehadi and Wintner (2022) showed some Arabizi words that have meanings in English and Hindi.

Proceedings of the The First Arabic Natural Language Processing Conference (ArabicNLP 2023), pages 423–434
December 7, 2023 ©2023 Association for Computational Linguistics

comprehensive overview of the existing works addressing offensive language detection in the context of Arabizi;

- We assess the performance of various language models (mBERT, DziriBERT, DarijaBERTarabizi, SVM) in detecting the offensive language in the Arabizi text without transliterating it to the Arabic script[3]. Our experiments are both in-domain and out-of-domain;

- We analyze the results per each of the two dialects (Algerian, Lebanese) composing the used dataset, which allows shedding light on the behaviour of the leveraged pre-trained models;

- Finally, we make available[4] the used dataset, which we created by merging data and unifying the annotation from 4 available datasets.

The remainder of this paper is structured as follows. Section 2 offers a critical overview of the works dealing with Arabizi in the context of offensive language detection. Section 3 details the process of the dataset creation. Sections 4 and 5 are devoted to the experimentation and the presentation of their outcomes. Finally, Section 6 discusses the findings and conclusions.

2 Related work

A small number of offensive language detection works have dealt with Arabizi (Appendix A presents a summary of each of these works, along with a recap in Table 1). However, these works exhibit one or more of the following shortcomings:

- The dataset is predominantly written in Arabic script with only a minority of examples in Arabizi (Boucherit and Abainia 2022; Mohdeb et al. 2022; Röttger et al. 2022);

- The dataset is conceived to serve primarily a different task than offensive language detection (Abainia 2020; Raïdy and Harmanani 2023; Riabi et al. 2023), resulting in a small size or a low proportion of offensive examples;

- The conducted experiments or the reported results did not focus on offensive language detection in Arabizi (Abainia 2020; Boucherit and Abainia 2022; Mohdeb et al. 2022; Raïdy and Harmanani 2023);

- In the few works (Riabi et al. 2023; Röttger et al. 2022) that reported results on Arabizi, the datasets have a small number of Arabizi examples, and they did not encompass social media texts. Consequently, their results do not allow to make definitive assessments regarding the performance of models in this specific text genre.

Considering the shortcomings of the previous studies listed above, and the prevalent use of Arabizi, it became clear that there is a pressing need to pay more attention to the problem of offensive language detection on Arabizi. To address this need, there is a requirement for the creation of additional resources that would facilitate a thorough evaluation of this task.

3 Dataset

In light of the above discussion on the limitation of the available datasets, our objective is to create a single, relatively large dataset with a plausible ratio of offensive language. This dataset could be then exploited for the development of offensive language detection models. Inspired by the work of (Risch et al. 2021), we favoured leveraging the available resources instead of starting from scratch. Therefore, we decided to construct the dataset by merging the Arabizi samples from the datasets DZMP, DZOFF, DZREF and LBSA (*cf.* Table 1 in Appendix A for details on these datasets). To achieve this, we followed the subsequent steps:

Extraction of the Arabizi samples from the datasets that comprise, in addition, Arabic script. This was straightforward for the DZREF dataset, as it comprises an attribute determining whether the text is in Arabic script, Arabizi, French or English. For the DZOFF dataset, however, we made this extraction automatically by filtering out the messages that contain only the Latin alphabet and numbers.

Unification of the labels. Our goal is to obtain a dataset for binary classification where the

[3] Before the era of large language models, transliterating Arabizi to the Arabic alphabet has been a common practice in Arabic language processing tasks such as sentiment analysis (Matrane et al. 2023).

[4] https://github.com/Imene1/Arabizi-offensive-language

offensive class encompasses a wide range of abusive text including hate speech (with its subcategories such as racism and sexism), profanity and obscene content. To this end, all the labels referring to any kind of offensiveness (*cf.* Table 1), were integrated into one label (Offensive). Similarly, all the labels indicating the absence of offensive language were mapped to one label (Non-Offensive). For this task, we examined carefully the definitions of labels provided in the paper or the documentation of each dataset. For the majority of labels, it was easy to decide whether it represents offensive language or not.

Nonetheless, for a few labels, where the definition was not enough to decide, we examined, in addition, a sample of the data having this label. This was the case, of two labels: "Refusing with non-hateful words (RNH)" in the anti-refugee dataset (DZREF) and the label "sarcasm" in the Lebanese sentiment analysis dataset (LBSA). By inspecting some examples of the class RNH, we decided to consider them among the offensive class. This is because despite those messages do not contain swearing, they exhibit discrimination and xenophobia, which is in line with the wide definition of offensive language we adopted. Concerning the cases in the "sarcasm" class, we examined all the examples in this class that have the sentiment polarity "negative", assuming that the offensiveness could not be positive. This examination showed us that all the cases are non-offensive.

Merging the datasets. We have merged into one CSV file the entire examples of DZMP and LBSA datasets along with the Arabizi parts of DZOFF and DZREF datasets.

As shown in Table 2 [5], the obtained dataset comprises more than 7000 social media texts from different platforms. More than 20% of its textual examples are offensive, which is an acceptable ratio to train a detection model. Given its distinctive features, including its size, the proportion of offensive content, the two different dialects it contains, and its diverse sources from social media, we assert that this dataset is currently the most suitable choice for evaluating the performance of offensive language detection in Arabizi, which is addressed in the next section.

4 Experiments

The goal of our experiments is 3 fold:

- To estimate the performance of detecting offensive language specifically on Arabizi in two contexts: in-domain and out-of-domain.
- To analyze the performance per dialect.
- To gain insights into the misclassified cases.

We carried out our experiments with 3 variants of BERT. Below is a succinct overview of them.

Multilingual BERT (a.k.a. mBERT) [6] (Devlin et al. 2019): BERT Language model pre-trained on 104 languages including Arabic. Previous experiments using this model in the context of POS tagging and dependency parsing (Muller et al. 2020), as well as sentiment analysis (Fourati et al., 2021), proved it can generalize to handle Arabizi by fine-tuning it using datasets in this form of Arabic.

DziriBERT [7] (Abdaoui et al. 2021): BERT Language model pre-trained on more than one million tweets in the Algerian dialect including Arabizi.

DarijaBERT-arabizi[8] (Gaanoun et al. 2023): a variant of BERT pre-trained on more than 4 Million texts on the Moroccan dialect (a.k.a. Darija) written in Arabizi.

As baselines, we used SVM with TF-IDF as features and the majority class heuristic.

To provide a robust estimate of those model's performance, we applied the following evaluation setup:

For the **in-domain** context, we fine-tuned the three BERT-like models and trained SVM through 5-fold cross-validation using the created dataset.

The obtained models were then tested on two **out-of-domain** datasets, which are the Arabizi part of EGMHC (Röttger et al. 2022) and DZTRB (Riabi et al. 2023). The former comprises synthetic Arabizi texts in Egyptian and the second comprises Algerian Arabizi collected from a news website and a song lyrics corpus (*cf.* Appendix A for further details on these datasets).

The hyper-parameters used to fine-tune BERT models are displayed in Table 3. Adam optimizer was used in all the models.

[5] Due to space limitations, Tables 2-8 are included in Appendix C.

[6] `https://github.com/google-research/bert.` The cased version is used.

[7] `https://github.com/alger-ia/dziribert`

[8] `https://huggingface.co/SI2M-Lab/DarijaBERT-arabizi`

5 Results and discussion

5.1 In-domain results

Table 4 displays the performance scores of the models in terms of F1 measured on the 2 classes offensive and non-offensive as well as the macro-averaged F1 and the accuracy. It should be noted that those measures are computed on the predictions file wherein the results obtained from the cross-validation folds are appended. We also reported the average of the F1 scores computed on the 5 folds (they are displayed between parenthesis on the table).

The results show that all the models outperformed the majority class baseline. Even SVM, which is not context-aware and does not have any prior knowledge on Arabizi was able to classify correctly 17% of the offensive texts with a precision of 97%[9], resulting in an F1 of 0.29.

The mBERT model reached an F1 score of 0.92, showing an improvement of +0.32 in comparison with SVM's result. This means that the contextual embeddings that this model learned from multiple languages allowed it to capture some of the patterns in the Arabizi text even though it was not pre-trained with this form of Arabic, which confirms the findings of previous studies (Fourati et al. 2021; Muller et al. 2020).

DziriBERT and DarijaBERT-arabizi perform almost equally and surpass mBERT, most notably in the offensive class. This shows the advantageous impact of pre-training BERT with Arabizi. Additionally, those results suggest that knowledge can be effectively transferred across Arabic dialects, even when expressed using Latin script. This is illustrated by the good performance of DarijaBERT-arabizi on our dataset, which comprises Algerian and Lebanese Arabizi, despite being pre-rained on Moroccan Arabizi.

In the following section, we will delve deeper into the analysis of performance per dialect to gain further insights.

5.2 Performance per Dialect

Table 5 shows the performance scores computed on the examples of each dialect separately. With regard to the Lebanese dialect, the performance of all the models in the non-offensive class is outstanding and superior to their performance in the offensive class. This result is indeed expected since the Lebanese sub-dataset is extremely imbalanced (the ratio of the offensive texts is only 6%), which makes it easy to reach a high-performance score on the majority class. This is evidenced by the F1 score of 0.97, which was reached on the non-offensive class just by a random guess using the majority class heuristic.

Interestingly, although the Lebanese dialect was unseen in the pre-training phase of the three used BERT models, DziriBERT and DarijaBERT-arabizi generated good results on the offensive class, with a raise of +0.14 and +0.13 respectively in F1 score in comparison with mBERT. This supports our previous remark concerning the transferability of knowledge across dialects, meaning that knowledge on the Algerian dialect (and also the Moroccan dialect) was useful in improving the offensive language detection performance on the Lebanese dialect despite the fact that those dialects are very different from each other.

In the context of the Algerian dialect, DarijaBERT-arabizi achieved the highest performance, showing only a marginal distinction from DziriBERT. This outcome is quite predictable because it is expected that knowledge transfer from the Moroccan dialect to the Algerian one would be effective given the substantial similarities between these dialects. Appendix B provides an analysis of the misclassified cases with examples from the dataset.

5.3 Out-of-domain results

Table 7 reports the results of testing the models that were fine-tuned through the experiments described in Section 5.1 on two unseen datasets. It should be noted that the EGMHC dataset comprises also texts in the Arabic script, but we reported, in Table 7, the results computed only on the Arabizi examples, which all belong to the positive class[10]. Therefore, using accuracy would be enough to measure the performance on this dataset. Additionally, the results obtained on DZTRB were computed only on its test set (DZTRB$_{test}$), with the aim of allowing their comparison with the results reported in (Riabi et al. 2023) (displayed in the last three lines). Note that, the models in Riabi's et al. paper have been trained on the training set of DZTRB and tested on

[9] Precision and Recall are not displayed in the tables of results. We mentioned them for illustration reasons.

[10] The positive class in this dataset is *hateful*, which we mapped to *offensive* to be compatible with the label of the dataset used previously to fine-tune the models.

its test set. Our main observations on the obtained results are below.

The SVM model failed to identify any offensive instances in either dataset, suggesting that the content in EGMHC and DZTRB deviates significantly from the training data, thereby impeding the model's ability to generalize.

Overall, the performance is very poor on EGMHC, indicating, again, a high dissimilarity of this dataset with the ones used to pre-train and fine-tune the models. This dissimilarity could be related to the fact that this dataset is in the Egyptian dialect, which was unseen in the fine-tuning and pre-training phases of all the used models.

On the other hand, the performance on the $\text{DZTRB}_{\text{test}}$ dataset was not as low as the one on EGMHC, and fairly close to the results achieved by the models fine-tuned on the training subset of this dataset, obtained from Riabi et al. paper[11]. This could be explained by the fact that the texts in DZTRB are in the Algerian dialect, which is a seen dialect in the fine-tuning phase. This allows the models to generalise to some extent on this dataset.

Unlike the in-domain results, mBERT generated the highest accuracy score on EGMHC and the highest F1 on the offensive class on $\text{DZTRB}_{\text{test}}$. Nonetheless, its score remains too poor to be significant.

Those results show different difficulty degrees for the models to generalise across datasets, illustrated by the very low results on EGMHC and the moderate results on DZTRB. In both cases, this implies the necessity of domain adaptation to improve performance. In this context, we were able to find only a couple of works addressing the topic of domain adaptation across Arabic dialects in the context of offensive language detection (Husain and Uzuner 2022) and sentiment analysis (El Mekki et al. 2021). Consequently, further investigation in this area is warranted.

6 Conclusion

In this research paper, we have explored the fascinating topic of offensive language detection within Arabizi. Regarding this topic is still underexplored, our study aimed to shed light on the performance of models in this specific linguistic context. Throughout our investigation, the following key findings have emerged:

Feasibility of detecting offensive language in Arabizi: despite the complexity of Arabizi, our experiments demonstrated that offensive language in this form of Arabic could be detected with high-performance scores without transliterating it to the Arabic script. This was evidenced by an F1 score that researched 0.96 using cross-validation on a dataset comprised of texts collected from different sources and in two distinct dialects, Algerian and Lebanese. However, the generalizability of the models across datasets is a challenge, especially if the dialect is different, as shown through our out-of-domain experiments.

The role of pre-trained language models: we showed that while a plausible performance could be reached by multilingual BERT, the best results are obtained by the models pre-trained partially or totally with Arabizi, which are DziriBERT and DarijaBERT-arabizi, respectively. On the other hand, SVM, a traditional machine learning model, generated poor results. This highlights the importance of transfer learning and context-aware models in dealing with the complexity of Arabizi.

Challenges in dialectal variation: our dialect-specific analysis of results along with our inspection of the misclassified cases revealed that despite the transferability of knowledge between dialects, it remains essential to tailor approaches to each dialect for better performance. This is particularly important because the vocabulary of offensive language may vary among the various dialects. This finding was underscored by our out-of-domain experiments, showing the difficulty of models to generalize on an unseen dialect (Egyptian).

Those findings suggest that future research efforts, in the context of offensive language detection in Arabizi, have to focus on the development of more datasets and pre-trained language models for the various Arabic dialects, as the majority of the existing resources concern the Algerian dialect. They also highlight the necessity of domain adaption research, most notably across dialects.

Finally, we anticipate that the findings of this study and the dataset we have made publicly accessible will pave the way for further research on this topic.

[11] Even the results of Riabi et al. are low. The best macro F1 is 0.61 as indicated in Table 7. See Appendix D for further details on DZTRB dataset.

Limitation

The limitation of our research is twofold. First, we used in our experiments a dataset comprising only Arabizi texts. However, this does not reflect the distribution in the real world, wherein Arabic script and Arabizi coexist together, sometimes in a single message. Moreover, code-switching to French and English occurs frequently in Arabizi, an aspect that we did not investigate in our experimentations. Therefore, it would be important, in future studies, to consider these two real-world aspects.

Second, our research did not address the transliteration of Arabizi to the Arabic script. We used instead models compatible with Arabizi. Thus, it is still unknown whether the transliteration improves the performance.

Acknowledgements

This publication was made possible by NPRP grant 13S-0206-200281 from the Qatar National Research Fund (a member of Qatar Foundation). The findings achieved herein are solely the responsibility of the authors.

References

Kheireddine Abainia. 2020. DZDC12: a new multipurpose parallel Algerian Arabizi–French code-switched corpus. *Language Resources and Evaluation*, 54(2):419–455.

Amine Abdaoui, Mohamed Berrimi, Mourad Oussalah, and Abdelouahab Moussaoui. 2021. DziriBERT: a Pre-trained Language Model for the Algerian Dialect. *Preprint arXiv 2109.12346*.

Hamdah Alghamdi and Eleni Petraki. 2018. Arabizi in Saudi Arabia: A deviant form of language or simply a form of expression? *Social Sciences*, 7(9).

Oussama Boucherit and Kheireddine Abainia. 2022. Offensive Language Detection in Under-resourced Algerian Dialectal Arabic Language. *arXiv preprint arXiv:2203.10024*:1–9.

Giulia Brabetz. 2022. Arabizi: A Linguistic Manifestation of Glocalization in the Arabic Language Area? *Maydan: rivista sui mondi arabi, semitici e islamici*, 2:103–129.

Kareem Darwish. 2014. Arabizi Detection and Conversion to Arabic. *ANLP 2014 - EMNLP 2014 Workshop on Arabic Natural Language Processing, Proceedings*:217–224.

Mithun Das, Saurabh Kumar Pandey, and Animesh Mukherjee. 2023. Evaluating ChatGPT's Performance for Multilingual and Emoji-based Hate Speech Detection.

Jacob Devlin, Ming Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, 1(Mlm):4171–4186.

Abdellah El Mekki, Abdelkader El Mahdaouy, Ismail Berrada, and Ahmed Khoumsi. 2021. Domain Adaptation for Arabic Cross-Domain and Cross-Dialect Sentiment Analysis from Contextualized Word Embedding. *NAACL-HLT 2021 - 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference*:2824–2837.

Chayma Fourati, Hatem Haddad, Abir Messaoudi, Moez Ben Haj Hmida, Aymen Ben Elhaj Mabrouk, and Malek Naski. 2021. Introducing A large Tunisian Arabizi Dialectal Dataset for Sentiment Analysis. *WANLP 2021 - 6th Arabic Natural Language Processing Workshop, Proceedings of the Workshop*:226–230.

Kamel Gaanoun, Abdou Mohamed Naira, Anass Allak, and Imade Benelallam. 2023. DarijaBERT : A Step Forward in NLP for the Written Moroccan Dialect. Technical report.

Imane Guellil, Ahsan Adeel, Faical Azouaou, Fodil Benali, Ala Eddine Hachani, Kia Dashtipour, Mandar Gogate, Cosimo Ieracitano, Reza Kashani, and Amir Hussain. 2021. A Semi-supervised Approach for Sentiment Analysis of Arab(ic+izi) Messages: Application to the Algerian Dialect. *SN Computer Science*, 2(2):1–18.

Mariam Haghegh. 2021. Arabizi across Three Different Generations of Arab Users Living Abroad: A Case Study. *Arab World English Journal For Translation and Literary Studies*, 5(2):156–173.

Fatemah Husain and Ozlem Uzuner. 2021. A Survey of Offensive Language Detection for the Arabic Language. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 20(1):1–44.

Fatemah Husain and Ozlem Uzuner. 2022. Transfer Learning Across Arabic Dialects for Offensive Language Detection. *2022 International Conference on Asian Language Processing, IALP 2022*:196–205.

Yassir Matrane, Faouzia Benabbou, and Nawal Sael. 2023. A systematic literature review of Arabic dialect sentiment analysis. *Journal of King Saud*

University - Computer and Information Sciences, 35(6):101570.

Djamila Mohdeb, Meriem Laifa, Fayssal Zerargui, and Omar Benzaoui. 2022. Evaluating transfer learning approach for detecting Arabic anti-refugee/migrant speech on social media. *Aslib Journal of Information Management*, 74(6):1070–1088.

Benjamin Muller, Benoit Sagot, and Djamé Seddah. 2020. Can Multilingual Language Models Transfer to an Unseen Dialect? A Case Study on North African Arabizi. *arXiv preprint arXiv:2005.00318*.

Maria Raïdy and Haidar Harmanani. 2023. A Deep Learning Approach for Sentiment and Emotional Analysis of Lebanese Arabizi Twitter Data. In *ITNG 2023 20th International Conference on Information Technology-New Generations, Advances in Intelligent Systems and Computing 1445*, pages 27–35.

Arij Riabi, Menel Mahamdi, and Djamé Seddah. 2023. Enriching the NArabizi Treebank : A Multifaceted Approach to Supporting an Under-Resourced Language. In *Proceedings of the 17th Linguistic Annotation Workshop (LAW-XVII)*, pages 266–278. Association for Computational Linguistics (ACL).

Julian Risch, Philipp Schmidt, and Ralf Krestel. 2021. Data Integration for Toxic Comment Classification: Making More Than 40 Datasets Easily Accessible in One Unified Format. *WOAH 2021 - 5th Workshop on Online Abuse and Harms, Proceedings of the Workshop*:157–163.

Paul Röttger, Haitham Seelawi, Debora Nozza, Zeerak Talat, and Bertie Vidgen. 2022. MULTILINGUAL HATE CHECK : Functional Tests for Multilingual Hate Speech Detection Models. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 154–169. Association for Computational Linguistics (ACL).

Ali Shazal, Aiza Usman, and Nizar Habash. 2020. A Unified Model for Arabizi Detection and Transliteration using Sequence-to-Sequence Models. *Proceedings of the Fifth Arabic Natural Language Processing Workshop*:167–177.

Safaa Shehadi and Shuly Wintner. 2022. Identifying Code-switching in Arabizi. *WANLP 2022 - 7th Arabic Natural Language Processing - Proceedings of the Workshop*:194–204.

Mohammad Ali Yaghan. 2008. “Arabizi”: A Contemporary Style of Arabic Slang. *Design Issues*, 24(2):39–52.

A Summaries of the Related Works

Typically, non-Arabic characters and numbers are considered noise and hence deleted during the processing of Arabic text or the creation of datasets. This practice results in the omission of the Arabizi script. Indeed, only a few offensive language detection datasets involve this form of Arabic. In this section, we provide an overview of those few works.

Abainia (2020) created a multipurpose dataset (DZMP) [12] in Arabizi comprising 12 Algerian sub-dialects. The dataset is collected from Facebook and annotated for several tasks, namely code-switching, sub-dialect identification, emotion detection, gender identification, and abusive language detection. The abusive comments constitute only 12% of the dataset. To the best of our knowledge, this dataset has not been yet exploited in abusive language detection experiments.

Mohdeb et al. (2022) addressed the problem of detecting anti-refugee and anti-migrant speech. They created a dataset (DZREF) composed of more than 4500 YouTube comments in the Algerian dialect including 434 comments in Arabizi with code-switching to French and English. Their experiments, using different variants of BERT, showed that the performance of the hate speech detection models is impacted negatively when including the Arabizi comments. However, further investigation is needed in this regard since the percentage of Arabizi in the used dataset is too small (only 9%).

Röttger et al. (2022) constructed a particular dataset known as Multilingual Hatecheck. It is a functional test, which encompasses synthetic texts in 10 languages. The Arabic subset (EGMHC), which is mostly in the Egyptian dialect, contains 3570 cases of both hateful and non-hateful content. These cases were carefully crafted by language experts using numerous templates, where the hate speech target and the slur word vary across the cases. The purpose of this dataset is to allow a controlled evaluation of hate speech detection models based on 25 fine-grained functionalities. Each functionality reflects the ability of the model to correctly classify specific kinds of hate or non-hate speech (e.g., implicit derogation, counter

[12] Throughout the paper, we use acronyms to denote each dataset. We constructed these acronyms by combining an abbreviation of the dialect (based on the ISO 3166-1 alpha-2 code of the respective country) with additional letters that indicate the dataset's primary purpose.

Authors	Source	Main Task (dataset acronym)	Dialect	Overall Size (size and proportion of Arabizi)	Annotation related to off. language	% off. examples in the Arabizi part
(Abainia 2020)	f	**M**ultipur**p**ose (DZMP)	DZ	2400 (2400, 100%)	**Abusive** , Not Abusive	12%
(Boucherit and Abainia 2022)	f	**Off**ensive language detection (DZOFF)	DZ	8749 (1415, 16%)	**Abusive**, **offensive**, none	61%
(Mohdeb et al. 2022)	▶	Anti-**ref**ugees and anti-migrant hate speech detection (DZREF)	DZ	4586 (434, 9.5%)	**Hate**, **Incitement**, Sympathetic, **Refusing with non-hateful words**, Comment (not hateful, nor sympathetic)	44%
(Raïdy and Harmanani 2023)	🐦	**S**entiment **a**nalysis (LBSA)	LB	3134 (100%)	**Bullying**, Courtesy words, **Foul language**, Joke, Known fact, **Racism**, **Sarcasm**, Saying, **Sectarianism**, **Sexism**, None	6%
(Röttger et al. 2022)	Synthetic text	Functional test for hate speech detection (EGMHC)	EG	3570 (133, 4%)	Hateful, Non-hateful	100%
(Riabi et al. 2023)	News website and song lyrics	Treebank (DZTRB)	DZ	1287 (1287, 100%)	Offensive , Non-offensive	22%

Table 1: Datasets comprising annotated offensive content in Arabizi. In Dialect column, the acronyms DZ, LB, EG refer to Algerian, Lebanese and Egyptian, respectively. In Size column, the first figure refers to the total number of examples in the dataset, while the figures inside parentheses are the number of Arabizi examples and their proportions to the overall size. To construct our dataset, the Arabizi data of the first 4 datasets were merged, wherein the annotation labels appearing in bold were mapped to the label offensive, and the rest were mapped to the label non-offensive. The two last datasets are not collected from social media, and we used them for the out-of-domain experiments.

speech), or cases exhibiting lexical or syntactic phenomena (e.g., negation, spelling variation including Arabizi). This dataset was used to evaluate an XLM-T model (a multilanguage model) that was fine-tuned using known offensive language datasets in 3 Latin languages. The model achieved an accuracy rate of 60.9% on the 133 Arabizi examples. In a recent work (Das et al. 2023), this dataset was also employed to evaluate ChatGPT, which achieved an accuracy of 75.9% on the Arabizi examples. However, it was not able to classify 20.3% of the cases.

Boucherit and Abainia (2022) proposed a dataset of offensive language detection (DZOFF) in the Algerian dialect crawled from Facebook. The dataset contains more than 8500 texts, in Arabic and Arabizi scripts, sampled from public pages and groups of controversial topics. Each text has been labelled as abusive, offensive or normal following the definitions provided by the authors. According to these definitions, offensive language includes any offence targeting individuals, groups or entities, whereas abusive speech corresponds to swearing or obscene content. The examples in Arabizi represent 16% of the dataset. The dataset was used in binary (offensive and abusive language classes are merged) and multiclass classification experiments employing traditional and deep learning models. Although a significant portion of the dataset is in Arabizi, the paper did not report the performance of the models specifically on this script.

The work of (Raïdy and Harmanani 2023) concerns sentiment analysis in the Lebanese dialect. In addition to the polarity labels (positive and negative), the created dataset (LBSA), which is entirely in Arabizi, contains labels providing hints on the tweets' content e.g., sexism, sectarianism, jokes, and sarcasm, among others. Only a small proportion of texts (6%) have labels referring to offensive content. Moreover, those labels have not been considered in the conducted experiments.

Contemporaneously with our work[13], Riabi et al. (2023) enriched the North African Arabizi Treeback with offensive language annotation. The dataset is composed of 1287 sentences in the Algerian dialect sampled from two sources: a corpus of user comments crawled from a newspaper website and a corpus of lyrics of Algerian songs. The paper reported results of offensive language detection experiments using BERT-like models. However, since the dataset is

[13] Riabi et al. work was published in July 2023 while we were conducting our experiments.

small and not collected from social media, it would be difficult to draw solid conclusions about the performance of those models in a real-world scenario.

B Error Analysis

To better understand the behaviours of the models in the in-domain experiments, we calculated the percentage of the misclassified cases by each model in each class based on the annotation of the source datasets composing our dataset (*cf.* Table 1. It displays the labels of the source annotation). Since all the source datasets adopted a ternary or multiclass annotation, it would provide a more precise description of the text than the binary classes we have adopted. Then, we averaged the misclassification percentages of the 4 models for each class.

Table 6 shows that the easiest texts to predict as non-offensive are the ones labelled as known fact from the LBSA dataset. All the 21 cases with this label have been classified correctly by all the models (see example 1 in Table 8)[14]. On the other hand, the easiest class to predict as offensive is the class Abusive from the DZOFF dataset, which is constituted of texts with obscene and swear words (ex. 2).

The non-offensive class with the highest ratio of false positives is Comment from the DZREF dataset. As shown in the table, more than 7% (on average) of the examples in this class, which is superposed to contain neutral discourse, have been flagged as offensive. After the examination of the cases that were marked as offensive by at least one model (totalled 32 cases), it turns out that nearly one-third of these cases are effectively offensive, meaning they were mis-annotated. Some of them involve untargeted swearing (ex. 3) and others involve hate speech but the targets were not refugees or migrants (ex. 4). This may explain why the annotators did not consider them as positive cases (e.g., hate speech), given that DZREF dataset specifically concerns hate speech directed at refugees and migrants.

We can also observe in Table 6 that the proportions of misclassification in the offensive classes (i.e., the false negatives) are higher than the proportion of misclassification in the non-offensive classes (i.e., the false positives). Furthermore, almost all the classes with the highest ratio (of false negatives) belong to the Lebanese dataset. For instance, the unique example that constitutes the class Racism (ex. 5) was predicted by the 4 models as non-offensive, resulting in a ratio of misclassification equal to 100%, followed by the class Foul language with an average of 42.3% of misclassification.

Since the number of cases misclassified by SVM is high, we examined only the cases predicted as non-offensive by the 3 BERT models. We noticed 3 kinds of cases:

- Error in the annotation or challenging examples.
- Texts with an implicit offence that employ terms very specific to the context and the culture of the country.
- Texts comprising well-known insults and obscene words.

The first two cases were present in examples in both Algerian and Lebanese dialects (ex. 6-8). However, interestingly, the last kind of errors was observed only among the texts in the Lebanese dialect comprising obscene words not used in the Algerian dialect. For example, the texts involving the obscene terms *ke**m* and Cha**ta (see the full texts in ex. 9, 10) have not been marked as offensive. This means that the classification models failed to identify some of the well-known swear words in the Lebanese dialect. Conversely, this is not the case in the Algerian dialect: as mentioned previously, the texts comprising swearing in the Algerian dialect were the easiest to identify as offensive). Therefore, it would be reasonable to attribute this discrepancy to two reasons:

1. the fact that two of the used models are pre-trained on the Algerian dialect or the Moroccan (which is similar to the Algerian), but not pre-trained on the Lebanese dialect,
2. the limited number of the offensive examples in Lebanese used to fine-tune the models (only 194 examples), which is not the case for the Algerian dialect (more than 1000 examples).

[14] All the examples from the datasets are listed in Table 8. To avoid the repetition of the table number, we will refer to the next examples only by mentioning the example number between parentheses.

In other words, the Lebanese data used to fine-tune the models, pre-trained on the Maghrebi dialects, were not diverse enough to effectively extend the applicability of these models to detect offensive language specific to the Lebanese dialect.

C Tables

Total # of examples	Dialects (ratio)	Platforms	# Offensive examples (%)
7383	DZ (57.5%) LB (42.5%)	Facebook YouTube Twitter	1526 (20.7%) DZ: 1332 LB : 194

Table 2: Statistics of the constructed dataset.

Learning rate	1e-5
Batch size	16
Number of epochs	3

Table 3: The used hyperparameters for the BERT models.

	F1			**Acc.**
	Non-Off.	**Off.**	**Macro**	
DziriBERT	**0.98**	0.93	**0.96** (0.93)	**0.97**
DarijaBERT-arabizi	**0.98**	**0.94**	**0.96** (0.93)	**0.97**
mBERT	0.97	0.87	0.92 (0.86)	0.95
SVM	0.90	0.29	0.60 (0.60)	0.83
Majority Class	0.88	0.00	0.44	0.79

Table 4: In-domain evaluation results using 5-fold cross-validation.

		F1			**Acc.**
		Non-Off.	**Off.**	**Macro**	
DziriBERT	DZ	0.97	0.94	0.96	0.96
	LB	**0.99**	**0.88**	**0.94**	**0.99**
DarijaBERT-arabizi	DZ	**0.98**	**0.95**	**0.97**	**0.97**
	LB	**0.99**	0.83	0.91	0.98
mBERT	DZ	0.95	0.89	0.92	0.93
	LB	0.98	0.74	0.86	0.97
SVM	DZ	0.84	0.32	0.58	0.75
	LB	0.97	0.06	0.51	0.94
Majority Class	DZ	0.81	0.00	0.41	0.69
	LB	0.97	0.00	0.48	0.94

Table 5: Dialect-specific Performance. The best results in the Algerian dialect (DZ) are highlighted in bold, while the best results in the Lebanese dialect (LB) are both bold and underlined.

Generic Class	Source dataset	Source class	# examples	% misclassifi-ations
Non-offensive	LBSA	Known fact	21	**0.00%**
	LBSA	Sarcasm	111	0.23%
	LBSA	Joke	112	0.45%
	LBSA	None	2631	0.50%
	DZMP	Not abusive	2119	1.12%
	LBSA	Courtesy words	32	1.56%
	LBSA	Saying	33	2.27%
	DZOFF	Normal	556	3.15%
	DZREF	Sympathetic	65	6.92%
	DZREF	Comment	177	7.20%
Offensive	DZOFF	Abusive (swearing and obscene content)	363	**14.94%**
	DZOFF	Offensive	496	26.46%
	DZREF	Incitement	8	28.13%
	DZREF	Hate	138	31.52%
	DZREF	Refusing with non-hateful words	46	33.15%
	LBSA	Sexism	2	37.50%
	DZMP	Abusive	281	38.52%
	LBSA	Sectarianism	14	41.07%
	LBSA	Bullying	60	41.25%
	LBSA	Foul language	117	42.31%
	LBSA	Racism	1	100%

Table 6: Average misclassification ratio of the 4 models.

		DZTRB$_{test}$				EGMHC (Arabizi part)
		F1			**Acc.**	**Acc.**
		Off.	**Non-Off**	**Macro**		
Fine-tuned on our dataset	DziriBERT	0.19	0.90	0.54	0.82	0.04
	DarijaBERT-arabizi	0.22	0.89	0.56	0.81	0.12
	mBERT	0.26	0.86	0.56	0.77	0.15
	SVM	0.00	0.90	0.45	0.81	0.00
Fine-tuned on DZTRB's training set	DziriBERT (Riabi et al., 2023)	0.37	0.85	0.61	-	-
	mBERT (Riabi et al., 2023)	0.00	0.90	0.45	-	-
	CharacterBERT (Riabi et al., 2023)	0.25	0.80	0.52	-	-

Table 7: Performance scores of testing the models on two out-of-domain datasets DZTRB$_{test}$ and EGMHC.

1	A true negative example from the source class Known fact in LBSA dataset (a class with 0% misclassification)	
	Arz	Absha3 shi bl safar huwe dab L shenat □□♀□😃
	Ar	أبشع شي بالسفر هو ضب الشنط □□♀□😃
	En	The worst thing about traveling is to pack suitcases.
2	One of the examples that was correctly predicted as offensive by the 4 models. It belongs to the class Abusive in DZOFF dataset, which is the offensive class with the smallest misclassification ratio. It contains text with obscene and swear words	
	Arz	Roh ta3**i ya n**ch
	Ar	روح تع**ي يا ن**ش
	En	Go get fu**ed, passive gay man.
3	An untargeted obscene popular word in the Algerian dialect, which was mis-annotated in DZREF dataset. Interestingly, it was predicted as offensive by mBERT in addition to DziriBERT.	
	Arz	tn**et
	Ar	تن**ت
	En	It's fu**ed
4	A mis-annotated offensive example from the DZREF dataset, which does not target African refugees or migrants.	
	Arz	bravo sahafi bravo france tfou lik
	Ar	برافو صحافي برافو، فرنسا تفو عليك
	En	Bravo, journalist, bravo! France spit on you.
5	The unique example in the Racism class from the LBSA dataset. It was marked as non-offensive by the 4 models.	
	Arz	plz plz gebran 5alik mtebe3 lmawdo3 ma ba2 badna phalesteneye wsoreyen 3ena el mawjoden bykafo
	Ar	بليز بليز جبران، خليك متابع الموضوع ما بقا بدنا فلسطينيّ وسوريين عنا الموجودين بيكفوا
	En	Please, please Gebran, stay tuned to the topic. We don't want Palestinians and Syrians here; the ones we have are enough.
6	An example of mis-annotation from the DZOFF. It was erroneously annotated as offensive but predicted as non-offensive by the 4 models. This expression is typically said by someone who feels wronged, directing it towards the wrongdoer. It conveys a sense of reliance on God's justice and intervention.	
	Arz	Hasbiya Allah wa ni3ma.el wakil fik
	Ar	حسبي الله ونعم الوكيل فيك
	En	God suffices me, and He is the best disposer of affairs concerning you.
7	An example in the Foul language class from the LBSA dataset. It was classified by the 4 models as non-offensive. This example is not inherently a foul language but it can becomes offensive if directed at a person a disrespectful manner. This is indeed a challenging case for annotation and classification if the context is unknown.	
	Arz	chou hal habel hayda man :p ma32oul
	Ar	شو هالهبل هيدا مان. معقول
	En	What is this nonsense, man :p I can't believe it.
8	False negative from DZOFF dataset: subtle offence that employ the term "sahib l kachir" translated literally to "people of sausage", which is very specific to the context of some political events in Algeria. This term refers to individuals who are perceived as being supportive of the government and are brought to governments' rallies with the incentive of receiving a sandwich containing sausage.	
	Arz	Hadou ysemhoum sehab l kachir […]
	Ar	هادو يسموهم صحاب الكاشير[…]
	En	Those are called people of sausage […]
9	An example of false negative from the LBSA although comprising a common swear word in the oriental Arabic dialects such as the Levantine and Egyptian.	
	Arz	Chou hal cha***ta
	Ar	شو هالش**طة
	En	What is this bi**h
10	Another example of a false negative although comprising a well-known swear word: "K**m". This is a highly offensive term in oriental Arabic dialects. It literally translates to derogatory terms related to female genitalia. The intended meaning is a strong curse directed at someone, often expressing extreme anger or disdain. The English translation below is not literal.	
	Arz	Mitl a ade bi Beirut.... k***m Nasrallah!
	Ar	متل العادة ببيروت...ك***م نصرالله
	En	Like usual in Beirut, curse on Nasrallah!
11	An example from $\text{DZTRB}_{\text{test}}$ dataset annotated as offensive. This could be considered a subjective annotation, illustrating the challenge of classifying the cases of this dataset.	
	Arz	nhab bladi w dima lalgerie w makra fl 3adyan
	Ar	نحب بلادي ودائماً للجزائر ومكرا في العديان
	En	I love my country and always for Algeria to spite enemies

Table 8: Examples from the used datasets. Each Arabizi example (Arz) is transliterated to the Arabic script (Ar) and translated to English (En)

D Performance on DZRTB dataset

An examination of a sample from DZTRB$_{test}$ confirmed the remark of its authors that harmful cases are indeed challenging even for annotators, which is also illustrated by the moderate inter-annotator agreement of 0.54. Most cases are about football, do not involve swear words, and the annotation seems subjective (see ex. 11 in Table 8). The difficulty of this dataset could be also illustrated by the low performance of the models trained and tested on it as reported in (Riabi et al., 2023): the best model (DziriBERT) yielded an F1 of 0.61 and mBERT did not detect any offensive case.