Impact of Normalization and Data Augmentation in NER for Algerian Arabic Dialect

Abdelhalim Hafedh Dahou[1] and Mohamed Amine Cheragui[2(✉)]

[1] Lorraine University and ATILF Lab, 54000 Nancy, France
adahou@atilf.fr
[2] Department of Computer Science and Mathematics, Ahmed Draia University, 01000 Adrar, Algeria
m_cheragui@univ-adrar.edu.dz

Abstract. Today, social media incorporate a considerable volume and variety of textual data constituting a first source for several NLP applications to take advantage of it in order to better understand and apprehend social phenomena. For this purpose, several approaches have been applied to exploit this mass of data, among them the deep learning approach, which has demonstrated its performance against classical machine learning algorithms since 2006. However, this approach inherited the same problems as those of classical machine learning, namely the heterogeneity of the resources (most of the time, the data are not directly exploitable, they must be sorted, filtered and normalized) dedicated to learning, but also in most cases the available resources are reduced in size for some languages and dialects which affects the performance of the models in training. The aim of this paper is to study the impact of normalization and data augmentation on named entity recognition (NER) task on Algerian text, using Arabic pre-trained models. For training and evaluation, we built a corpus dedicated to the recognition of named entities, called DzNER, based on Facebook's comments. To evaluate the models, we set up 4 scenarios, the first one without normalization and data augmentation, in this case, the *ARBERT* model outperformed the other models with an F1 score of 84.4%. The second scenario is to use normalization, which enabled the *Dziribert* model to get the highest F1 Score of 81.9%. The third scenario with data augmentation, where the *Arabert* v0.2 model yielded the best F1 score with 85.1%. The *Arabert* v0.2 model again obtained the best F1 Score with 86.2% in the last scenario combining normalization and data augmentation.

Keywords: Algerian dialect · NER · Deep learning · Normalization · Data augmentation

1 Introduction

It is clear that today, most of the work done in natural language processing is focused on social media platforms like Facebook for two major reasons The first one is related to the study of sociological phenomena and the second one is technical in relation to the

S. Chikhi et al. (Eds.): MISC 2022, LNNS 593, pp. 249–262, 2023.
https://doi.org/10.1007/978-3-031-18516-8_18

availability of the raw data (textual resources). One of the tasks for which social media have positively impacted on their development is the named entities recognition.

Named Entity Recognition (NER) is a sub-task of information extraction that attempts to identify and classify mentions of named entities in unstructured text into predefined categories, such as: names of persons, organizations, and geographic locations, as well as other concepts such as time, measures (money, percent, weight, etc.), email and addresses,… etc. [1].

Most of the work in NER has focused on English since 1990 [2, 3] A significant progress has made during the last decade in Arabic named entities recognition (ANER), despite the difficulties that this language presents due to complex linguistic phenomena such as: (i) Agglutination, which is part of morphological analysis, is present sometimes in context of NER appearance. (ii) Non-Diacritization complicates the lexical analysis because the text treated is highly ambiguous. (iii) The absence of capitalization makes the detection of NER more complicated. Actually, several works are interested in the NER from the dialects (a variants of the Arabic language) due mainly to the emergence of social media such as: Facebook and Twitter [4, 5], considered as an important source of information regarding the quantity of comments and messages published on these social media.

NER is a very active field of research, with a considerable number of studies, based on different approaches, which are: Extraction based on linguistic or symbolic approaches (rule-based approaches), statistical or machine learning approaches, and recently on deep learning approach.

Although the use of the deep learning approach has achieved satisfying results compared to classical machine learning techniques, for most tasks in natural language processing. This is due to its reliance on non-linear transformations while learning complicated features from data, in addition to reducing the time and effort required to design features by learning useful representations from raw data automatically [6]. It still faces the same problem as the classical machine learning such as: a heterogeneity of orthographic forms (several variants) in writing words (a persistent phenomenon in the Arabic dialect) and the lack of linguistic resources (corpus) in terms of availability and quantity. Faced with these two facts, solutions like normalization and data augmentation have been developed to address these issues.

The normalization in NER consists mainly in creating a standardization (unification) in the writing of the entities. It is clear that the texts coming from the social media present a multitude of anomalies, especially in dialects, in particular, Algerian dialect, such as: the switching code (Example: مريض grave اك/ you are seriously ill) [7], the exaggeration, which consists in duplicating the character several times in the word (Example: شووووف/ look at), the use of the numbers to replace characters (Example: 3ndek? / you, have it?), as well as other phenomena. Therefore, several works have been done to deal with these issues through the implementation of several techniques including: the deletion of duplicated characters exceeding two times in the case of exaggeration, to more structured processes, such as the creation of a writing convention like CODA [8].

Data Augmentation allows, in some cases, to overcome the problem of SMALL quantity of data by increasing the limited initial dataset before the training phase [9]. In the literature, the concept of data augmentation has grown with deep learning, where

several techniques have emerged: unigram noising [10], random swap and deletion [11], Synonym Replacement: geometric distribution [12], Embedding Replacement [13] and the Replacement by Language Models [14]. Still the disadvantage of those techniques is that they may change the label of the entity or the meaning of the sentence that includes the entity.

The aim of this paper is to observe the impact of these two processes on NER in Algerian dialect, using deep learning approach by evaluate this impact on the well-known Arabic pre-trained models which are: Dziribert, ARBERT, MARBERT, mBERT, Arabert base and Arabert Twitter trained and tested on a corpus that we have developed called DzNER.

This paper is organized as follows: in Sect. 2, we summarized the related works that exist in this field. Section 3, we discuss the methodology of the work. The experiments and dataset are presented in Sect. 4. Before ending, results and their discussion are introduced in Sect. 5. Finally, in Sect. 6, we conclude with the study's outcomes and future perspectives.

2 Related Work

Based on our research, there is limited work on the Arabic dialect, and especially for the Algerian dialect. Most of the research that can be found in the literature have focused on Modern Standard Arabic (MSA). For this reason, this section will be divided into two parts, the first will list some works on NER task on Arabic dialect and the second includes works done on normalization and data augmentation in the case of Algerian dialect.

For NER in the Arabic dialect, we can mention two works, the first one by [15], built an NER system for Egyptian dialect (social media). The designed tool is based on a machine learning approach using the CRF (Conditional Random Fields) approach, to recognize two types of entities: persons and locations. The system developed uses a number of features: Lexical features, Contextual Features, Gazetteers, Morphological Features, Distance from specific keywords and Brown Clustering. Their system yields a performance of 49.18% for PER and 91.43% for LOC using the F-measure and based on a corpus built by the authors and manually annotated. The second work [16], presented a system that recognizes and translates Tunisian named entities into MSA using: bilingual dictionaries, gazetteers, syntactic grammars, and finite-state transducers, implemented in the NooJ linguistic platform. The developed tool can recognize three categories of entities: ENAMEX, TIMEX and NUMEX. For the experimentation, the authors built a Tunisian dialect corpus collected from social media and Tunisian novels. Their system yields a performance of 86.81% F1 measure for ENMAX, 94.98% for TIMEX and 94.37%.

For the normalization, we have noted the existence of several works concerning the Algerian dialect.

Saadane and Habash [17] have developed a set of orthographic guidelines for the Algerian dialect (ALG CODA) and more specifically a sub-variant of it, which is the Algiers dialect. This normalization is based on the study done by [17] concerning Arabic dialect. For this purpose, the authors consider several exceptions and extensions specific

to the Algerian dialect including phonological extensions, phono-lexical exceptions, morphological extensions and lexical exceptions in order to deal with words that have specific spelling. [18], proposed during their development of a sentiment analyzer, for Facebook posts written in Algerian dialect, a normalization process in 05 steps: step 1, delete any letter that appears twice in the sequence, remove punctuation marks except the question mark (?) and the exclamation mark (!) and replacing the letters with the word " رقم/ Number". Step 2, keeping emoticons in the comments. Step 3, making sure that the word and its abbreviation exist in the corpus. Step 4, Identification of the French Language Words. The last step, remove the diacritics, normalization of "Alif", remove the elongation and normalization of Hamza. [19], presented an Algerian Arabizi-French code-switching corpus (DZDC12). The corpus is manually constructed from Facebook comments (2400 comments) representing 12 cities of Algeria (12 Sub Dialects). The Corpus is designed to be a benchmark for several NLP tasks including the NER task (annotated according to Persons and Locations).

Guellil and Azouaou [20] build a synthetic analyzer of Algerian dialect, the authors made a normalization of certain phenomena which persists in their corpus, as an example: code switching using the frequency, treated the exaggeration by using regular expressions and the appearance of numbers in some words (for example: wa9tach/when) by a substitution of the numbers by the corresponding letters. [21], presented a hybrid opinion mining approach to classify social media messages written in the Algerian dialect with Latin letters. They propose a normalization by adding a regrouping step that is split into two sub-steps: Phonetic regrouping (by using 'soundex' algorithm [22]), and similarity regrouping (using the Levenshtein distance [23]). The various tests carried out have shown that the hybrid approach with regrouping (overall F score of 87,74%) outperforms the hybrid approach without regrouping (Overall F score of 85,23%).

Abidi and Smaili [24] built an Algerian dialect lexicon, which covers the phenomenon of code switching in Facebook comments. For this purpose, the authors performed two actions: Transliteration of foreign words and word embedding, in order to identify words (using Word2Vec). The objective is to find a list of words that could be correlated to a lexical entry whatever the language. The proposed approach achieved a F-measure score of 73%, on a test lexicon.

For the data augmentation in the case of the Algerian dialect, we refer two works: [25], proposed a new dataset for code switched NER for Algerian dialect, the authors observed the impact of multi-task learning between four tasks Code-Switch Detection (CSD), NER, Spelling Normalization and identifying users' sentiments, by adopting a Deep learning approach. From the different Experiments, 02 conclusions emerge: the first one, the accuracy indicates that learning NER task jointly with CSD (99.82%) improves its performance over learning the task separately (99.80%). Moreover, data augmentation proved a positive effect at the very beginning of learning (before epoch 6), but it is outperformed by non-augmented data after that.

Adouane et al. [26] presented a manually annotated corpus dedicated to sentiment analysis in Algerian dialect. According to the authors, the annotation process takes into consideration 04 tags which are: POS (Positive), NEG (Negative), NEU (Neutral) and MIX (when the comment combines between Positive and Negative and/or Neutral). The authors used the corpus through a comparative study between 04 models, which are:

SVM, CNN, LSTM, and BiLSTM. To deal with the unbalanced data, the authors added experimentation by augmenting the number of minority classes without changing the test set. The augmented CNN obtained a gain of 10.54 points on the F-score.

3 Methodology

As previously stated, we used Algerian social media data to conduct an analysis study on the impact of normalization and data augmentation on the NER task. The goal of this research is to learn more about the impact of normalizing text before feeding it to the model, as well as the impact of text quantity and entities diversity on the model's accuracy. We also contribute additional data to the community that can be utilized in other NLP tasks, as well as a lexicon to aid researchers working with the Algerian script.

3.1 Normalization

Our procedure to normalize the Algerian dialect corpus is based on three steps, calculate the frequency of each word in the corpus, cluster the most similar words based on semantic and lexical similarity and finally, apply the normalization by choosing the most frequent word in the cluster. As mentioned above, there is no existing research done on this pre-processing step, so we start the normalization from scratch without any existing lexicon for entities. Before starting building our strategy, we tested different word embedding representations and algorithms in order to find the similarity between the words but the results are less than expected because there is a variation in the style of writing. As solutions, we tested the fasttext [27] by applying a training phase in the whole corpus using both architectures (cbow and skipgram) and request it to find the similar words with the get_nearest_neighbors method but the results doesn't covers all the desirable matches, for example, the model can't find the similarity between "**الجزاير**"and "Dz" or "**لالجيري**"which means Algeria. The same for *Dziribert* [28] which is pre-trained on the Algerian dialect, especially in the terms of semantic similarity in the case of "**البلاد**"where acts as "**الجزائر**"when someone is out of Algeria and said that I want to go to the country, and also we tested the algorithms that calculates the differences between words in terms of characters and context. For that, we decided to build our strategy and in the next sections, we will describe the processing steps in detail.

Words Frequency: Before the extraction of the similar words in the corpus, we need to find firstly the canonical forms and then the similar words to those canonical forms. As mentioned above, there are no previous canonical forms in the literature for the entities in the Algerian dialect or any lexicon for the normalization, for that, we decided to create our lexicon based on the data that we have. The first step is to calculate the frequency of each word in the corpus, and based on the word frequency, we take the most frequent word and consider it as a canonical form. After analyzing the corpus, we found that the Algerian users used different words in the social media compared to the words used in their life discussion, for example, the Algerians when speak about Algeria, they said "**دزاير**"but in the social media, they used "**الجزاير**"or "**الجزائر**", we know that "**دزاير**"is the most word used to express Algeria, but in our case with this data, we took "**الجزائر**"as the canonical form.

Clustering: Due to the weakness of the automatic tools and models to find the most similar words in the Algerian corpus and cluster them, we decided to do this part of the process manually. An expert in the Algerian dialect and also in the user generated text in social media conducted this part and built a lexicon that groups canonical words with their transliterations that exist in the corpus. This operation resulted in a normalization lexicon of 711 entries and Fig. 1 represents part of this lexicon.

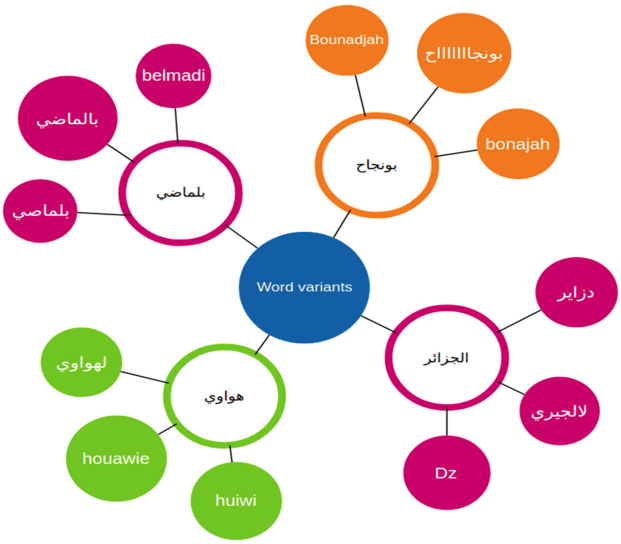

Fig. 1. Examples from the normalization lexicon.

Apply the Normalization: The last step concentrates on the normalization by applying a conversion in the Algerian corpus by modifying the words that exist in the lexicon by the canonical form in order to obtain a significant corpus and the same writing style without ambiguities.

3.2 Data Augmentation

As known in the literature, training a deep learning model needs a large number of texts to increase the chances that the model does the task with very good accuracy. But usually, there are cases when we do not have much data for model training. This problem can be solved by doing data augmentation. It is the technique through which one can increase the size of the data for the training of the model, with adding the new data which will take time and cost highly compared to the images data and sometimes the techniques can produce words out of the context or change definitely the context of the sentence. In our case, and to prevent any errors or text context modification, we decided to increment the size of the data by adding new data collected from social media. The data set used in this

study named *DzNER* contains three entities: Person, Location and Organization in totale of 578 person, 548 location and 186 organization collected from Algerian Facebook pages.

To do the data augmentation, our strategy passed through several steps including: select the Facebook pages, collect all the comments in the posts, pre-process the comments and finally annotate the new data by two expert annotators.

Starting with the selection of the Facebook pages, we chose those Facebook pages based on the number of followers and comments. Not all the posts are selected, but just the posts that have a large number of comments that contain entities. After the selection of the Facebook pages and their posts, we built a python script that takes as an input the link of the post and produces as an output a CSV file that contains all the comments that exist in the current post without any other information like user's ID, user's name, or comment's time. Moving on, all the collected sentence passed by a pre-processing phase in order to:

- Remove spaces and emojis.
- Remove the sentences that contain classical Arabic or more MSA than Algerian dialect manually.
- Delete links and hashtags.
- After all the above notes, we tokenized the sentences into tokens to be ready for annotation.

Finally, the annotation process applied by two expert annotators in the Algerian dialect and also in the social media user generated by following the same guidelines to produce a high-level corpus in terms of credibility. Table 1 shows the differences between the *DzNER* corpus and the new data collected in terms of entities.

Table 1. DzNER corpus and the new data collected in numbers.

Categories	DzNER	DzNERAG	Difference
Persons	578	2219	1641
Locations	548	1374	826
Organization	*186*	500	314

4 Experimentation

The aim of this study as mentioned above, is to evaluate the effect of normalization and data augmentation on the Algerian text for the NER task. In addition, examine the quality of the normalization lexicon. Since this is the first work on Algerian dialect normalization in NER, we have built a new lexicon of 711 entries and the *DzNER* corpus to evaluate the performance of the produced lexicon, and for the impact of the data augmentation, we collected and injected more data to the *DzNER* corpus, in order

to explore the differences. The evaluation was done by using five Arabic pre-trained models in different Arabic dialects and MSA which are: *Arabert* [29], *ARBERT* [30] and *MARBET* [30]. For the Algerian dialect we used the *Dziribert* [28] model, and the last model is the multilingual Bert known by *mBERT* [31], which was trained on different languages including the Arabic and its dialect.

We first started by evaluating the models on the *DzNER* corpus without any normalization or augmentation. Secondly, we normalized the *DzNER* corpus and applied an evaluation on the five models in order to test the impact of normalization on small data. The third experiment focused on discovering the impact of data augmentation and this is by evaluating the five models on the *DzNERAG*. The last experiment consisted on applying a normalization process on the *DzNERAG* and made the same evaluation in order to figure the impact of normalization in augmented data and compared it with the second experiment.

4.1 Dataset

The used data was acquired from Facebook posts comments using a python script and Google Chrome extension named instant Data Scraper, as described in the data augmentation section. In Algeria, Facebook is the most popular social media network, with more users than Twitter. The Facebook pages and posts were chosen based on the number of followers and comments received in various categories such as sports, travel, electronics and others. We found a lot more entities for person in the sports topic, the traveling topic was specialized in the location entity, and the electronics for the organization entity. The goal is to gather reliable data that can be used to train models and to cover all entities with the same distribution. The two corpora are described in Table 2 in terms of number of entities and also the test part that was extracted from the same data collected used for the comparison between the models.

Table 2. Statistics about the two corpora and the test part.

Corpus	Tokens	Persons	Locations	Organization
DzNER Train	22320	578	548	186
DzNERAG Train	55227	2219	1374	500
Test	3504	86	113	20

4.2 Pre-trained Models

In order to evaluate the impact of normalization and data augmentation, we evaluate the performance of the most well-known Arabic and multilingual pre-trained models on the NER task using Algerian dialect dataset. Table 3 presents a comparison between those models in terms of size, dataset and versions. All models have been fine tuned by following the standard architecture illustrated in Fig. 2. The implementation script as well as the dataset utilized will be available on Github.

Table 3. Used Arabic pre-trained models.

Models	Size (params)	DataSet (nwords)	Vocab size
Arabert v0.2-base	136M	8.6 Billion	64K
Arabert v0.2-Twitter-base	136M	8.6 Billion + 60 Million Multi-Dialect Tweets	/
Dziribert	124M	1 million tweets	50K
ARBERT	163M	6.2 Billion	100K
MARBERT	163M	6.2 Billion	100K
mBERT	110M	1.5 Billion	106K

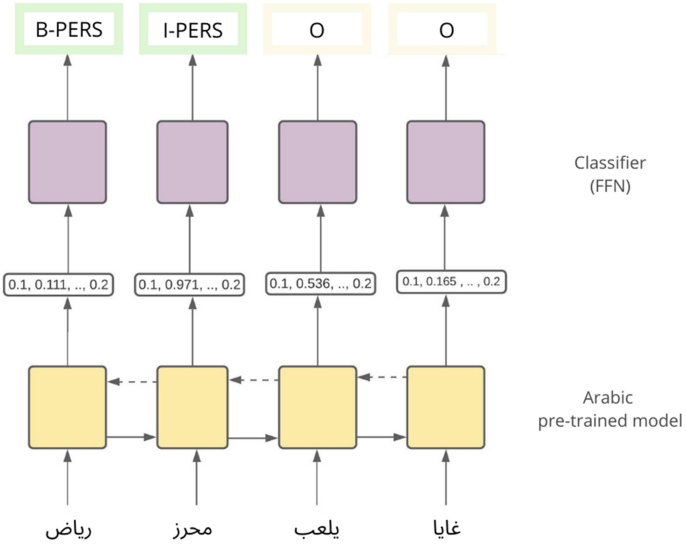

Fig. 2. Model fine-tuning architecture.

4.3 Experimentation Setup

All of the training and testing trials were carried out using the Google Colab platform with a GPU Tesla P100-PCIE-16GB. To make the annotation process easier, we used Google Sheets to facilitate the work and to share the sheet between the annotators. Using the test dataset, we fine-tuned the hyper-parameter to find the optimal configuration for each pre-trained model in order to achieve the best results. The hyper-parameters for each model are listed in Table 4.

5 Results and Discussion

The models are evaluated by calculating the precision, recall, and F1-score of their output on the test dataset. Precision and recall are often used metrics to provide more

Table 4. Hyper-parameters value for each used model.

Models	Epochs	Learning rate	Warmu setps
Arabert	20	5e−5	42
Dziribert	20	5e−5	0
MARBERT	15	3e−5	42
ARBERT	30	3e−5	42
mBERT	15	5e−5	42

accurate outcomes as well as to provide more information to the expert about the model's behavior, particularly in multi-class classification.

Table 5. Results of the first experiment using *DzNER* corpus.

Metrics	Arabert v0.2-base	Arabert v0.2-Twitter-base	mBERT	ARBERT	MARBERT	Dziribert
P	0.875	0.846	0.776	**0.888**	0.799	0.831
R	0.803	0.803	0.698	**0.803**	0.780	0.767
F1	0.838	0.824	0.735	**0.844**	0.789	0.798

Table 5 shows the results of the five models which trained on the *DzNER* corpus. We can observe that the *ARBERT* model outperformed the other models in terms of F1 score and this is because the *ARBERT* is already pre-trained on different sources of data writing in dialectical style and MSA, also the large vocab helps the model to boost its performance. *Dziribert* outperformed the *MARBERT* and *mBERT* in terms of precision and F1-score and both were outperformed by *Arabert* due to the quantity of data seen in the training before. *mBERT* is the latest model in terms of results compared to the others.

Table 6. Results of the first experiment using normalized *DzNER* corpus.

Metrics	Arabert v0.2-base	Arabert v0.2-Twitter-base	mBERT	ARBERT	MARBERT	Dziribert
P	**0.855**	0.843	0.771	0.839	0.806	**0.855**
R	0.781	0.781	0.690	0.781	**0.795**	0.785
F1	0.817	0.811	0.729	0.809	0.800	**0.819**

Table 6 shows the normalization impact on the performance of the five models that were trained on the normalized *DzNER* corpus. We can observe that there is an

amelioration for the MARBERT and *Dziribert* models in which they outperformed the rest of models in terms of all metrics which prove that training small models on significant data will boost their performance. Moreover, we can see that the method used for normalization decreased the performance of *ARBERT*, *Arabert* and also *mBert*.

Table 7. Results of the first experiment using *DzNERAG* corpus.

Metrics	Arabert v0.2-base	Arabert v0.2-Twitter-base	mBERT	ARBERT	MARBERT	Dziribert
P	**0.991**	0.880	0.803	0.906	0.857	0.852
R	0.799	**0.803**	0.748	0.794	0.794	0.767
F1	**0.851**	0.840	0.775	0.846	0.824	0.807

Table 7 shows the data augmentation impact on the performance of the models that were trained on the DzNERAG corpus. We can observe that data augmentation boost the performance of all models and Arabert base version outperformed the others models in terms of precision and F1-score but in terms of recall, we see that the twitter version achieved better results due to the large data seen in the training phase that contains much more dialectical texts. As we can see, the data augmentation gives all the models a boost in terms of performance compared to the normalization and this is due to the diversity and quantity of data.

Table 8. Results of the first experiment using normalized *DzNERAG* corpus.

Metrics	Arabert v0.2-base	Arabert v0.2-Twitter-base	mBERT	ARBERT	MARBERT	Dziribert
P	**0.896**	0.857	0.846	0.873	0.823	0.858
R	0.831	**0.853**	0.757	0.821	0.831	0.803
F1	**0.862**	0.855	0.800	0.847	0.827	0.830

Table 8 shows the data augmentation and normalization impact on the performance of the models that were trained on the normalized DzNERAG corpus. As we said above, in this experiment, we want to see the impact of both techniques at the same time. We can observe that data augmentation and normalization boost the performance of all models and demonstrate that the hybrid of those techniques lead to achieve good results compared to applying each technique individually. *Arabert* base still outperformed the others models in terms of precision and F1-score and the twitter version outperformed the other models in terms of recall due to the quantity of dialectal data seen. Compared to Table 7, we see the positive impact of the normalization and we can conclude that the normalization works positively on large data rather than small data.

6 Conclusion

Algerian users in social media don't pay attention to their writing style or follow any grammar or orthography because they don't exist for the Algerian dialect, users write a word as they want or such as it is pronounced. This behavior will create several transliterations for a single word and also include different languages in one sentence and lead several NLP tasks, particularly NER, to be affected by this issues. The influence of normalization and data augmentation on the Algerian dialect was investigated in this article. The process of normalization was carried out by combining related transliterations into a single canonical form, which is the most common transliteration in the corpus. Data augmentation was accomplished by adding new data to a tiny dataset that we created for this project. Normalization had a beneficial influence on *Dziribert* and *MARBERT* but a negative impact on the others, according to the results of the experiment. Data augmentation improves the performance of all models that have been evaluated. The last experiment shows that combining the two methods always improves model performance. This research provides fresh insights on data normalization and augmentation in Algerian dialect, particularly for the NER task, as well as a new dataset and lexicon to be exploited. As a result, we'll expand our lexicon's scope to cover more entities, as well as look into the impact of these techniques on the MSA script.

References

1. Balla, H., Delaney, S.J.: Exploration of approaches to Arabic named entity recognition. In: CEUR Workshop Proceedings, vol. 2611, pp. 2–16 (2020). https://doi.org/10.21427/ETJH-KF40
2. Ehrmann, M., Hamdin, A., Pontes, E.L., Romanello, M., Doucet, A.: Named entity recognition and classification on historical documents: a survey. arXiv: 2109.11406. https://arxiv.org/abs/2109.11406 (2021)
3. Shaalan, K.: A survey of Arabic named entity recognition and classification. Comput. Linguist. **40**, 469–510 (2014)
4. Ritter, A., Clark, S., Etzioni, O., Etzioni, M.: Named entity recognition in tweets: an experimental study. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 1524–1534. (2011)
5. Xiaohua, L., Shaodian, Z., Furu, W., Ming, Z.: Recognizing named entities in tweets. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pp. 359–36 (2011)
6. Alsaaran, N., Alrabiah, M.: Arabic named entity recognition: a BERT-BGRU approach. Comput. Mater. Contin. **68**(1), 471–485 (2021). ISSN: 1546-2226. https://doi.org/10.32604/cmc.2021.016054. http://www.techscience.com/cmc/v68n1/41836
7. Sabty, C., Elmahdy, M.S., Abdennadher, S.: Named entity recognition on Arabic-English code-mixed data. In: 2019 IEEE 13th International Conference on Semantic Computing (ICSC), pp. 93–97 (2019)
8. Habash, N., Diab, M., Rambow, O.: Conventional orthography for dialectal Arabic. In: Proceedings of the 8th Language Resources and Evaluation Conference (LREC) (2012)
9. Feng, S.Y., et al.: A survey of data augmentation approaches for NLP. In: Findings of the Association for Computational Linguistics: ACL-IJCNLP, pp. 968–988 (2021)
10. Xie, Z., et al.: Data noising as smoothing in neural network language models. ArXiv abs/1703.02573 (2017)

11. Wei, J., Zou, K.: EDA: easy data augmentation techniques for boosting performance on text classification tasks. arXiv: 1901.11196 (2019)
12. Mosolova, A., Fomin, V., Bondarenko, I.: Text augmentation for neural networks. In: AIST (2018)
13. Bayer, M., Kaufhold, M., Reuter, C.: A survey on data augmentation for text classification. ArXiv abs/2107.03158 (2021)
14. Marivate, V., Sefara, T.: Improving short text classification through global augmentation methods. In: Holzinger, A., Kieseberg, P., Tjoa, A., Weippl, E. (eds.) Machine Learning and Knowledge Extraction. CD-MAKE 2020. LNCS, vol. 12279, pp. 385–399. Springer, Cham (2020). ISSN: 1611-3349. https://doi.org/10.1007/978-3-030-57321-8_21
15. Zirikly, A., Diab, M.: Named entity recognition for dialectal Arabic. In: Proceedings of the EMNLP Workshop on Arabic Natural Langauge Processing, pp. 78–86 (2014)
16. Torjmen, R., Haddar, K.: The automatic recognition and translation of tunisian dialect named entities into modern standard Arabic. In: Bekavac, Božo, Kocijan, K., Silberztein, M., Šojat, K. (eds.) NooJ 2020. CCIS, vol. 1389, pp. 206–217. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-70629-6_18
17. Saadane, H., Habash, N.: A conventional orthography for Algerian Arabic. In: Proceedings of the Second Workshop on Arabic Natural Language Processing, pp. 69–79 (2015)
18. Soumeur, A., Mheni, M., Guessoum, A., Daoud, A.: Sentiment analysis of users on social networks: overcoming the challenge of the loose usages of the Algerian dialect. In: The Fourth International Conference on Arabic Computational Linguistics (2018)
19. Abainia, K.: DZDC12: a new multipurpose parallel Algerian Arabizi–French code-switched corpus. Lang. Resour. Eval. **54**, 419–455 (2020)
20. Guellil, I., Azouaou, F.: ASDA: Analyseur Syntaxique du Dialecte Algérien dans un but d'analyse sémantique. arXiv: 1707.08998. http://arxiv.org/abs/1707.08998 (2017)
21. Bettiche, M., Mouffok, M.Z., Zakaria, C.: opinion mining in social networks for Algerian dialect. In: Medina, J., Ojeda-Aciego, M., Verdegay, J., Perfilieva, I., Bouchon-Meunier, B., Yager, R. (eds.) Information Processing and Management of Uncertainty in Knowledge-Based Systems. Applications. IPMU 2018. Communications in Computer and Information Science, vol. 855. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-91479-4_52
22. Holmes, D.O., McCabe, M.C.: Improving precision and recall for Soundex retrieval. In: Proceedings International Conference on Information Technology: Coding and Computing, pp. 22–26 (2002)
23. Levenshtein, V.I.: Binary codes capable of correcting deletions, insertions, and reversals. In: Soviet physics. Doklady, vol. 10, pp. 707–710 (1965)
24. Abidi, K, Smaili, K.: An automatic learning of an Algerian dialect lexicon by using multilingual word embeddings. In: Proceedings of the 11th Language Resources and Evaluation Conference (LREC) (2018)
25. Adouane, W., Bernardy J.P.: When is multi-task learning beneficial for low-resource noisy code-switched user-generated Algerian texts? In: The 4th Workshop on Computational Approaches to Code Switching (2020)
26. Adouane, W., Touileb, S., Bernardy, J.P.: Identifying sentiments in Algerian code-switched user-generated comments. In: Proceedings of the 12th Language Resources and Evaluation Conference (LREC), pp. 2698–2705 (2020)
27. Bojanowski, P., Grave, E. Joulin, A., Mikolov, T.: Enriching word vectors with subword information. Trans. Assoc. Comput. Linguist. **5**, 135–146 (2017). Editor: Hinrich Schutze
28. Abdaoui, A., Berrimi, M., Oussalah, M., Moussaoui, A.: Dziribert: a pre-trained language model for the Algerian dialect. arXiv:2109.12346 (2021)
29. Antoun, W., Fady, B., Hazem, H.: Arabert: transformer-based model for Arabic language understanding. arXiv preprint arXiv:2003.00104 (2020)

30. Abdul-Mageed, M., AbdelRahim, E., El Moatez Billah, N.: ARBERT & MARBERT: deep bidirectional transformers for Arabic. arXiv preprint arXiv:2101.01785 (2020)
31. Pires, T., Schlinger, E., Garrette, D.: How multilingual is multilingual BERT?. arXiv preprint arXiv:1906.01502 (2019)