Introducing A large Tunisian Arabizi Dialectal Dataset for Sentiment Analysis

Chayma Fourati **Hatem Haddad** **Abir Messaoudi**

Moez Ben Haj Hmida **Aymen Ben Elhaj Mabrouk** **Malek Naski**
iCompass, Tunisia
{chayma,hatem,abir,moez,aymen,malek}@icompass.digital

Abstract

On various Social Media platforms, people tend to use the informal way to communicate, or write posts and comments: their local dialects. In Africa, more than 1500 dialects and languages exist. Particularly, Tunisians talk and write informally using Latin letters and numbers rather than Arabic ones. In this paper, we introduce a large common-crawl-based Tunisian Arabizi dialectal dataset dedicated for Sentiment Analysis. The dataset consists of a total of 100k comments (about movies, politic, sport, etc.) annotated manually by Tunisian native speakers as Positive, Negative, and Neutral. We evaluate our dataset on sentiment analysis task using the Bidirectional Encoder Representations from Transformers (BERT) as a contextual language model in its mul-tilingual version (mBERT) as an embedding technique then combining mBERT with Convolutional Neural Network (CNN) as classifier. The dataset is publicly available[1].

1 Introduction

While being so diverse and rich, Arabic language and particularly Arabic dialects, are still under represented and not yet fully exploited. Arabizi is a term describing a system of writing Arabic using English characters. This term comes from two words “arabi” (Arabic) and “Engliszi” (English). Arabizi is the representation of Arabic sounds using Latin letters and numbers to replace the non existing equivalent ones (Mulki et al., 2017). Particularly in Tunisia, this way of writing was introduced in (Fourati et al., 2020) as ”Tunizi”.

In this paper, we introduce a large Tunizi dataset available freely in order to help Tunisian and other researchers interested in the Natural Language Processing (NLP) field. This dataset contains 100k Tunizi comments collected from crowd-sourcing and built for the sentiment analysis task. Indeed, the dataset’s comments are manually annotated by Tunisian native speakers as positive, negative, or neutral. This dataset can be also used for other NLP subtasks such as dialect identification, named entities recognition, etc.

In this paper, we present the motivation of our work, the Tunizi way of writing, previous related work, data composition including the data collection, preprocessing, annotation, and the data statistics. Then, we present experiments performed using our dataset for the Sentiment Analysis (SA) task. Finally, we present the conclusion and future work.

2 Tunizi

Tunizi was referred in (Fourati et al., 2020) as ”the Tunisian Arabic text written using Latin characters and numerals rather than Arabic letter”. This type of writing is most used on social media platforms since it presents an easier way to communicate. Also, since the Tunisian dialect already contains French and English words, people tend to use Tunizi for easier typing of non formal texts. However, not all Arabic letters have a Latin equivalent representation such as the character ”ق” or ”ح”. These characters are expressed using numbers instead in the Tunizi representation. Tunisians have high contact culture where online social networks play a key role in facilitating social communication (Skandrani et al., 2011). For instance, due to the linguistic situation specific to each country, Tunisians generally use the french background when writing in Tunizi, whereas, other Arabic countries like Egypt would use English. For example the word اشتريت would be written in Tunizi as ”chrit” whereas as ”ishtareet” in Arabizi.

In Table 1, some examples of Arabic letters and their Tunizi equivalent are presented.

[1] https://zenodo.org/record/4275240.YDew1-oo9H4

Proceedings of the Sixth Arabic Natural Language Processing Workshop, pages 226–230
Kyiv, Ukraine (Virtual), April 19, 2021.

Arabic	Arabizi	Tunizi
ح	7	7
خ	5 or 7’	5 or kh
ذ	d’ or dh	dh
ش	$ or sh	ch
ث	t’ or th or 4	th
غ	4’	gh or 8
ع	3	3
ق	8	9

Table 1: Arabic letters and equivalent Arabizi and Tunizi representation

In (Abidi, 2019), a study was conducted on 1,2M social media Tunisian comments including 16M words and 1M unique words states the following: 53% of the comments used Romanized alphabet, 34% used Arabic alphabet, and 13% used script-switching. The study states, also, that 87% of the comments based on Romanized alphabet are Tunizi, while the rest are French and English. Some examples of Tunizi sentences with different polarities (positive, negative, and neutral) are presented in Table 2.

3 Related Work

In (Mdhaffar et al., 2017), a Tunisian dataset entitled TSAC (Tunisian Sentiment Analysis Corpus) was presented. This dataset combines 17366 positive/negative Tunisian Facebook comments written in Arabic and Tunizi. TSAC includes only 9196 comments written in Tunizi. However, Tunizi comments are not balanced with 3856 negative comments (41%) and 5340 positive comments (59%). The doc2vec algorithm was used to produce document embeddings of Tunisian Arabic and Tunizi comments. The generated embeddings were fed to train a Multi-Layer Perceptron (MLP) classifier performing a F-measure values of 78%.

In (Chader et al., 2019), authors proposed a transliteration method from Arabizi to MSA in order to evaluate sentiment analysis of Arabizi Algerian dialect. The best achieved performances was 87% of F-score using the Support Vector Machine.

In (Taha, 2020), a work on Lebanese Arabizi was performed. In order to create Sentiment Analysis experiments, a dataset was collected from Twitter and manually labelled as positive and negative. The dataset resulted in 2.9K Tweets: 801 positive, 881 negative, and 1.2K neutral Arabizi tweets. Comments used for Sentiment analysis experiments are of size 1.6K balanced as 800 positive, and 800 negative tweets. A Sentiment Analysis task is, then, performed using a lexicon-based approach. The lexicon created was first of size 1.6K words and was expanded until achieved 80K words. The lexicon-based approach using the original lexicon classified only 23% of the tweets. Unclassified tweets are classified as positive or negative randomly. The classification using the original lexicon achieved 60% F1-score. Classification achieved a coverage of 63% of the tweets with an F1-score of 78% in one of the expanded versions of the lexicon.

In (Fourati et al., 2020), a Tunisian Arabizi Sentiment Analysis dataset called ”Tunizi” was introduced. This work provided 9k sentences annotated as positive and negative. TUNIZI is a balanced and representative dataset that covers different topics collected from Youtube social network. However, the dataset is not enough large to be used for deep learning approaches.

In (Baert et al., 2020), a work focuses on the Arabizi form of writing. The work creates a large dataset (LAD) of size 7.7M written in Arabizi collected from Arabizi tweets using Egyptian keywords as seeds. LAD was used for pretraining a BERT language model called ”BAERT”. A subset of LAD entitled SALAD of size 1700 was annotated for sentiment analysis. The subset was used for fine-tuning BAERT. SALAD contains tweets annotated as Positive, Negative, Neutral, Conflict and ConflictTxtvsEm (tweets where the comments have a conflict in the text and the emoji symbol) . However, the dataset is not balanced and very limited to be used for deep learning approaches. Tunizi dataset (Fourati et al., 2020) was also used for fine-tuning BAERT model. Fine-tuning on Tunizi and SALAD (3 classes) performed 83.8% and 74.4% of F1-score respectively. In order to perform better performances, such experiments would require larger annotated datasets.

In (Gugliotta et al., 2021), the authors present a multi-task sequence prediction system for Tunisian Arabizi. A dataset was annotated for several subtasks: text classification, tokenization, PoS tagging and encoding of Tunisian Arabizi into CODA* Arabic orthography. Nevertheless, the dataset was not annotated for sentiment analysis task.

Previous work proposed relatively small datasets to be used for deep learning approaches. Hence, our contributions can be summarized as: the first release of a Tunizi large-scale Common-Crawl-based

Tunizi Comment	MSA Translation	English Translation	Polarity
ma7leha lbo93a	ما احلى هذا المكان	This place is nice	Positive
5ayeb el loun	اللون بشع	The color is bad	Negative
bech nemchi nenta5eb	سوف انتخب	I will go vote	Neutral
5edma 3alamiya 5ouya	عمل رائع يا أخي	Great job bro	Positive

Table 2: Tunizi example comments and their Modern Standard Arabic (MSA) and English translations.

dataset[2], annotated manually for sentiment analysis task and training the dataset using neural networks and attention mechanisms for the Sentiment Analysis.

4 Tunisian Arabizi dataset

This dataset is composed of one instance presented as text comments, and labeled as positive, negative, or neutral. In this work, we use crowd-sourcing from social media until we reached 100k comments.

4.1 Data Collection

Because of the lack of available Tunisian dialect data (books, wikipedia, etc.), we use a Common-Crawl-based dataset extracted from social media. It is collected from comments on various social networks. The chosen posts included sports, politics, comedy, TV shows, TV series, arts and Tunisian music videos such that the dataset is representative and contains different types of ages, background, writing, etc. This data does not include any confidential information since it is collected from comments on public Social Media posts. However, negative comments may include offensive or insulting content. This dataset relates directly to people from different regions, different ages and different genders.

A filter was applied to ensure that only Latin-based comments are included. The extracted data was preprocessed by removing links, emoji symbols, and punctuations. Examples of comments before and after preprocessing are in Table 3.

4.2 Annotation Policy

The collected comments were manually annotated using an overall polarity: positive, negative and neutral. The annotation task was divided equally on ten Tunisian native speakers: two males PhD holders, aged 43 and 42; two females and one male, one aged 24 and two aged 26 working as research and development Artificial Intelligence engineers and five engineering students: four females, and one male aged 23.

Sentiment is an extremely multi-faceted phenomenon. For this reason, “light” comprehensive guidelines covering the most frequent potentially ambiguous cases were presented to the annotators with the aim of ensuring high enough agreement.

4.3 Dataset Statistics

We divided the dataset into separate training, validation and test datasets, with a ratio of 7:1:2 with a balanced split where the number of comments from positive class and negative class are almost the same. Table 4 presents statistics of the dataset including number of positive, number of negative, and number of neutral comments. Statistics of the train, validation, and test datasets for each label are presented in Table 5.

5 Experiments

To evaluate the dataset, two experiments were conducted:

- MBERT (Devlin et al., 2018) as language model and a classifier:

  We decided to use the Bidirectional Encoder Representations from Transformers (BERT) as a contextual language model in its multilingual version (MBERT) as an embedding technique. MBERT contains more than 10 languages including English, French and Arabic. Using the BERT tokenizer, we map words to their indexes and representations in the BERT embedding matrix. These representations are used to feed the MBERT classifier. Filter sizes used for all experiments are 2,3,4,5,6.

- MBERT as language model and Convolutional Neural Network (CNN) as classifier: Using the BERT tokenizer, we map words to their indexes and representations in the BERT embedding matrix. These representations are used to feed a CNN classifier.

[2] https://zenodo.org/record/4275240.YDew1-oo9H4

Before preprocessing	After preprocessing
3saall to9telll nmout 3liha!!!	3saall to9telll nmout 3liha
odkhlou lel google www.google.com	odkhlou lel google
ma3ejbetnich el 7al9a lyoum ☹	ma3ejbetnich el 7al9a lyoum

Table 3: Examples before and after preprocessing

Charactertistic	Statistic
#positive	52 711
#negative	43 767
#neutral	3 522
Total Comments	**100K**
#words	996 782
#unique words	189 257

Table 4: Dataset Statistics

Charactertistic	Train	Validation	Test
#positive	38 239	4 824	9 648
#negative	29 295	4 824	9 648
#neutral	2 466	352	704
Total Comments	**70K**	**10K**	**20K**

Table 5: Train, Validation and Test Datasets Statistics

Since the neutral class have less comments than the positive and negative ones, the same experiments were conducted on only two classes of the dataset: positive and negative. Hence statistics of the new sub-dataset are presented in table 6. The sub dataset was also divided for train, dev, and test as 7:1:2 where the train, dev, and test dataset are balanced.

Charactertistic	Statistic
Total Comments	**96 478**
#words	971 932
#unique words	208 464

Table 6: Binary labeled dataset

Experiments results on the binary and multi-label classification, are presented in table 7 and 8 respectively. Each table presents results of two experiments where MBERT is used as an embedding and a classifier and MBERT as an embedding and CNN as a classifier.

Model	Acc.	F1-Mic.	F1-Mac.
MBERT	**0.829**	**0.829**	**0.829**
MBERT & CNN	0.822	0.822	0.821

Table 7: SA results on binary classification

Model	Acc.	F1-Mic.	F1-Mac.
MBERT	0.79	0.79	0.558
MBERT & CNN	**0.803**	**0.803**	**0.580**

Table 8: SA results on multi-label classification

Our dataset showed that the Bidirectional encoder representations from transformers (BERT) as a contextual language model in its multilingual version (Devlin et al., 2018) as an embedding technique followed by CNN for classification outperforms the BERT as an embedding and a classifier when tested on the whole dataset for multi-label classification. The best result for Sentiment Analysis performed 80.3% accuracy and F1 micro score. However, F1 macro score achieved 58.0% because the neutral class contains less comments than the positive and negative ones. However, MBERT as an embedding and a classifier outperforms MBERT as an embedding and CNN as a classifier for binary classification when omitting the neutral class. Results achieved better performances and have led to an increase in all performance measures. Since this sub-dataset is balanced. Hence, accuracy, F1-Macro, and, F1-Micro metrics achieved 82.9%.

6 Conclusion

In this work, we present the largest dataset dedicated for the Tunisian Arabizi, annotated as positive, negative, and neutral and preprocessed for Sentiment Analysis subtask. Such datasets can be used for further NLP activities. We also experiment deep learning models using our dataset. This dataset could improve performances of several works dedicated for the Tunisian dialect in the NLP field. As a future work, we plan to build and release other large underrepresented African and Arabic datasets in order to help the NLP community further build machine learning and deep learning models.

References

Karima Abidi. 2019. *Automatic building of multilingual resources from social networks : application to Maghrebi dialects (Doctoral dissertation)*. Ph.D. thesis, University of Lorraine, Nancy, France.

Gaétan Baert, Souhir Gahbiche, Guillaume Gadek, and Alexandre Pauchet. 2020. Arabizi language models for sentiment analysis. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 592–603, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Asma Chader, Lanasri Dihia, Leila Hamdad, Mohamed Belkheir, and Wassim Hennoune. 2019. Sentiment analysis for arabizi: Application to algerian dialect. pages 475–482.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Chayma Fourati, Abir Messaoudi, and Hatem Haddad. 2020. Tunizi: a tunisian arabizi sentiment analysis dataset. In *AfricaNLP Workshop, Putting Africa on the NLP Map. ICLR 2020, Virtual Event*, page arXiv:3091079.

Elisa Gugliotta, Marco Dinarelli, and Olivier Kraif. 2021. Multi-task sequence prediction for tunisian arabizi multi-level annotation.

Salima Mdhaffar, Fethi Bougares, Yannick Estève, and Lamia Hadrich-Belguith. 2017. Sentiment Analysis of Tunisian Dialects: Linguistic Ressources and Experiments. In *Third Arabic Natural Language Processing Workshop (WANLP)*, pages 55–61, Valence, Spain.

Hala Mulki, Hatem Haddad, and Ismail Babaoglu. 2017. Modern trends in arabic sentiment analysis: A survey. *Traitement Automatique des Langues*, 58(3):15–39.

Hamida Skandrani, Abdelfattah Triki, and Boudour Baratli. 2011. Trust in supply chains, meanings, determinants and demonstrations: A qualitative study in an emerging market context. *Qualitative Market Research: An International Journal*, 14:391–409.

Tobaili Taha. 2020. *Sentiment Analysis for the Low-Resourced Latinised Arabic "Arabizi"*. Ph.D. thesis, TheOpen University.