TUNIZI: A TUNISIAN ARABIZI SENTIMENT ANALYSIS DATASET
Dataset Datasheet

Elaborated By:
iCompass

Tunis, 2020

1 Abstract

On social media, Arabic speakers tend to express themselves in their own local dialect. To do so, Tunisians use "Tunisian Arabizi", which consists in supplementing numerals to the Latin script rather than the Arabic alphabet. Analytical studies seek to explore and recognize online opinions aiming to exploit them for planning and prediction purposes, such as measuring customer satisfaction, and establishing sales and marketing strategies. However,in the African continent, analytical studies based on Deep Learning are data hungry. To the best of our knowledge, no annotated Tunisian Arabizi dataset exists.

In this datasheet, we introduce TUNIZI, a sentiment analysis Tunisian Arabizi dataset collected, preprocessed and annotated. The dataset will be made public in order to help the African NLP community in further research activities.

2 Introduction

Twitter, Facebook and other micro-blogging systems are becoming a rich source of feedback information in several vital sectors, such as politics, economics, sports and other matters of general interest. Our dataset is taken from people expressing themselves in their own Tunisian Dialect using Arabizi.

In this datasheet, we present the objectives of TUNIZI dataset, the motivation of this work, the composition of the dataset, the process of collecting data, pre-processing and annotating, uses of this dataset for future consumers, its distribution, and finally, how this dataset will be maintained.

3 Motivation

In [1], a survey was conducted to address the availability of Tunisian Dialect datasets. The authors concluded that all the existing Tunisian datasets are using Arabic letters and that there is a lack of Tunisian Arabizi annotated datasets. [2] presented a multi-dialectal parallel corpus of five Arabic dialects: Egyptian, Tunisian, Jordanian, Palestinian and Syrian in order to identify similarities and possible differences among them. The Overlap Coefficient results, representing the percentage of lexical overlap between the dialects, revealed that the Tunisian dialect has the least overlap with all other Arabic dialects. These results highlight the problem that the Tunisian Dialect is a low resource language and there is a need to create Tunisian datasets for analytical studies. In this paper, we present TUNIZI, a Tunisian Arabizi dataset for sentiment analysis studies. This dataset was created by the iCompass team, a Tunisian Startup speciallized in Artificial Intelligence and Deep Learning.

4 TUNIZI dataset composition

TUNIZI is composed of one instance presented as text comments collected from YouTube videos. Table 1 presents examples of comments with the translation to MSA and English where the first comment was annotated as positive and the second as negative.

Table 1: Tunisian Arabizi comments translated to MSA and English

Tunisian Arabizi	MSA	English
lkolna m3ak w msendinek	كلنا معك و نساندك	We are all with you and supporting you
nakrhek 5atrek kadheb	أكرهك لأنك كاذب	I hate you because you are a liar

This dataset is an annotated sample from a larger annotated dataset that contains 10k comments. The chosen sample includes 3k comments randomly taken from a balanced and representative dataset. The sample is also balanced including exactly the same amount of negative and positive comments. All comments are annotated as positive or negative. This data does not include any confidential information since it is collected from comments on public YouTube videos. However, negative comments may include offensive or insulting content. This dataset relates directly to people from different regions, different ages and different genders, since all the gathered data are comments on YouTube videos.

As a result, the dataset is balanced, containing exactly the same amount of positive comments and negative comments. Statistics, after pre-processing, including the total number of comments, number of positive and negative comments, number of words and number of unique words are stated in Table 2.

Table 2: Dataset Statistics

Characteristic	Number
#Comments	3000
#Negative comments	1500
#Positive comments	1500
#Words	26073
#Unique words	11244

5 Collection process

TUNIZI is collected from comments on Youtube social network.
All data was directly observable and did not require other data to be inferred from.
The latest comment existing in TUNIZI was published in 30/08/2019 and the most recent comment in 08/01/2020.
Comments were collected using scraping to extract our data from YouTube videos.
Scraping costs only time and computational resources since it is an automatic process.
During the collection, any non arabizi comment is removed.

We present a sample of size 3k out of a larger dataset containing 10k comments where 1500 positive comment and 1500 negative ones were taken randomly from the original dataset.
Data was collected by the iCompass team working on this project.

6 Preprocessing and annotation

TUNIZI was preprocessed by removing links, emoji symbols and punctuation. Two examples of comments before and after preprocessing are presented in Table 3.

Table 3: preprocessing examples

Before preprocessing	After preprocessing
Ana n7ebou ywali ra2is #kaïs_saied #17 ♡♡♡	ana n7ebou ywali ra2is kaïs saied 17
Vive mon cher président kais said!!!! ☺☺	vive mon cher président kais said

Annotation was then performed by five Tunisian native speakers, three males and two females at a higher education level (Master/PhD). Two male PhD holders, aged 43 and 42; one female and one male, both aged 25 working as research and development engineers at iCompass; and one female aged 23, an engineering student.

7 Uses

TUNIZI is new to the market, where few data exist and no projects on Tunisian Arabizi were conducted at earlier stages.
TUNIZI helps the African NLP community in further research activities in Sentiment Analysis projects.

8 Distribution

TUNIZI dataset will be made public for all upcoming research and development activities. TUNIZI will be published publicly on Github.

9 Maintenance

Since the Tunisian Dialect is not a standard language with written rules, dataset maintenance is crucial, especially updating the dataset with new vocabulary. An example is the appearance of new words and expressions in songs, series, TV shows, etc. An example of a new word and a new expression with English translation, the source and date of appearance are presented in Table 4.

Table 4: Examples of new Tunisian vocabulary

Comment	English translation	Source	Date
7oumeni	Neighbor	Rap song	2014
"akther m zhe men ghadi"	Very happy	TV show	2019

TUNIZI is maintained by iCompass team that can be contacted through emails or through the Github repository. Updates will be available on the same Github link

10 Conclusion

We presented TUNIZI, the first Tunisian Arabizi Dataset including 3K sentences, balanced, covering different topics, preprocessed and annotated. As the interest in Natural Language Processing, particularly for African languages is growing, a natural future step would involve building Arabizi datasets for other underrepresented north african dialects such as Algerian and Moroccan.

References

[1] Author Jihene Younes, Hadhemi Achour, Ahmed Frechichi, Emna Souissi *Survey on Corpora Availability for the Tunisian Dialect Automatic Processing*. [*2018 JCCO Joint International Conference on ICT in Education and Training, International Conference on Computing in Arabic, and International Conference on Geocomputing* (*JCCO: TICET-ICCA-GECO*)]. Hammamet, Tunisia, 2018.

[2] Houda Bouamor, Nizar Habash, Kemal Oflazer *A Multidialectal Parallel Corpus of Arabic*. [*Ninth International Conference on Language Resources and Evaluation* (*LREC-2014*)]. Reykjavik, Iceland, 1240-1245, 2014