Standardization of Dialect Comments in Social Networks in View of Sentiment Analysis : Case of Tunisian Dialect

Saméh Kchaou, Rahma Boujelbane, Emna Fsih, Lamia Hadrich Belguith

ANLP Research group, MIRACL Lab. FSEGS, University of Sfax, Tunisia

Introduction

■ Given the unlimited access to the internet, many languages of users written spontaneously which called Arabic dialect (AD) are presented on social networks (SN)

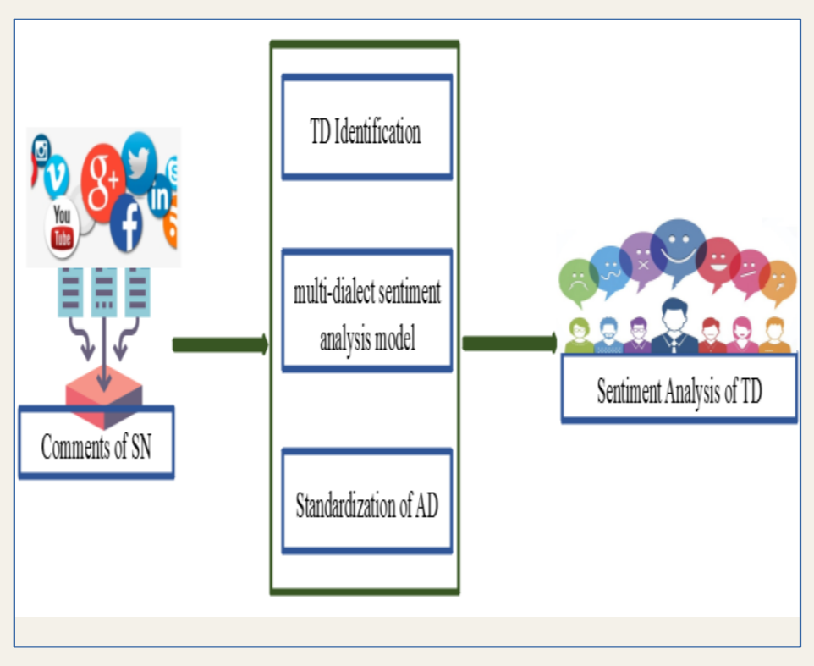
■ The need to process AD written on the SN allows to facilitate several NLP tasks such as opinions analysis.

=> Sentiment Analysis of the spoken Arabic Tunisian dialect (TD)

Problematic :

* TD written on SN are coupled by other languages(Foreign Languages, SMS language…

* Lack of resources for TD (parallel corpus, annotated corpus...)



=> Several solutions to adapt for the sentiment analysis of the TD:
1. Create an analysis model specified for the TD
2. Create a multi-dialect model applicable on TD
3. Create a standard template for a standard language.

Objectives :

=> Sentiment Analysis of TD using a standard model :
* 1. Identify the TD on SN
* 2. Translate the identified TD into a standard language: Modern Standard Arabic (MSA)
* 3. Apply a standard sentiment analysis model on the translated TD.

Overview

■ Several approaches are used for dialect identification:
  * Classical approaches :
    > (Kchaou et al., 2019)
    > (MADAR Shared Task 2019)

  * Deep learning approaches:
    > (Issa et al., 2021)
    > (NADI Shared Task 2020/2021)

■ Several approaches have been applied for Machine translation of TD:
  * Linguistic (Hamdi et all, 2013)
  * Statistical (Kchaou et all, 2020)
  * No works dealing with Neural Machine Translation until now besides our work on the translation of Tunisian transcriptions (Abida et all,2022)

Resources for Tunisian Dialect identification

■ **Corpus for Dialect Identification :** contains 95k annotated comments with 3 classes(Tun, MSA, Other), this corpus is collected from :

> **1**. (Kchaou et al., 2020): Parallel corpus containing 32k parallel sentences TD-MSA built using an augmentation method applied on :
  * MADAR : Includes 1.8k parallel sentences in travel domain of 25 Arabic dialects and the MSA,
  * PADIC : Containing 6.4k parallel sentences from everyday life and television programs in the Maghreb dialects, Levant dialects and MSA,
  * Tunisian constitution: Includes 500 parallel sentences,
  * 900 TD comments manually translated by native speakers into MSA.

> 2. NADI corpus: An annotated corpus at country-level,

Corpus Name	#TUN comments	#MSA comments	#Other comments
Corpus of (Kchaou et al., 2020)	32k	32k	0
NADI corpus	1k	0	30k
All corpus	33k	32k	30k

corpus statistics

■ Tunisian Arabic dialect identification (TADID) model :

> Traditional approaches:

	NB classifier	SVM classifier	MLP classifier
Score on DEV	81	80.01	70
Score on Test	80.15	79.60	71.3

	N-Gram Features		F1 score	
	Word	Char	Dev	Test
1.	1	-	81	80.15
2.	-	1	20.02	19
3.	1	1	80	80.95
4.	-	1→3	60.28	60.84
5.	1	1→3	82.50	**83.10**

Evaluation classifiers using Word feature.

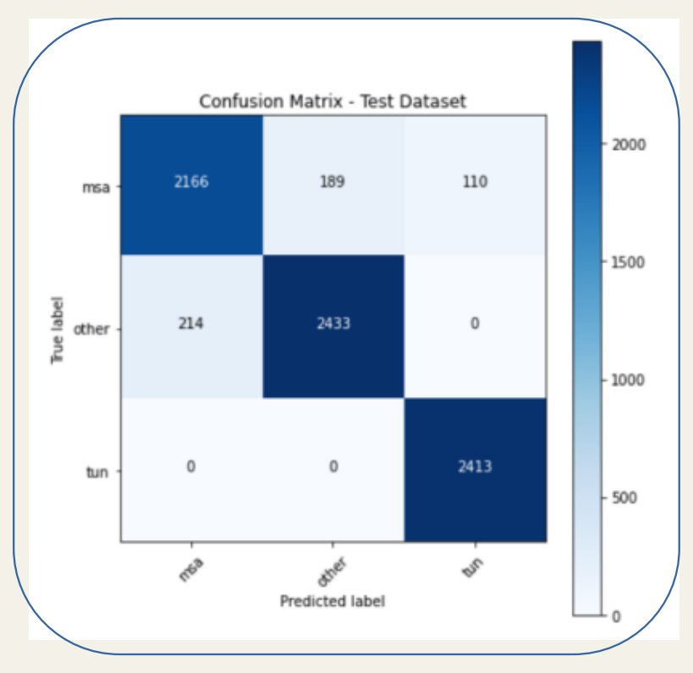
> Transfer learning approaches: TADID model is developed using the Multi-dialect-Arabic-BERT

Tunisian Dialect identification resources(2)

	Model Name	% Accuracy Score
1.	bert-base-arabic	86.50
2.	bert-large-arabic	75.20
3.	albert-base-arabic	67.10
4.	Multi-dialect-Arabic-BERT	**88.82**

Transformer classifier evaluation using BERT family.

=> Multi-dialect-Arabic-BERT = 88,82%



Output confusion matrix of TADID system

Resources for TD-MSA Translation

■ **TD-MSA parallel corpus:**
  > 32k parallel sentences (Kchaou et al., 2020)
> Social media corpus: 2k comments predicted using TADID model and manually translated by native speakers,

=> Given the lack of parallel corpus for TD data, the translated comments will be freely distributed for the research community,

■ **Substitution augmentation method for TD corpus:**
  > The used augmentation method consists to generate different TD sentences from the original TD sentences of the corpus without losing meaning of sentence by substituting words with their synonyms. The chosen words for substitution are selected randomly.
  > The multidialect-bert-base-arabic language model (Talafha et al., 2020b) is used to generate the synonyms of words

NMT corpus	#Lines	#Tokens	#Vocabulary
TD sentences	68k	199k	23.8k
MSA sentences	68k	201.5k	21.3k

Statistics of the created TD-MSA corpus for NMT model

■ **TD-MSA NMT model:**
  > Model based on Transformer encoder/decoder with self-attention mechanism
  > To configure the input for training model, data is encoded into token ID sequences using the tokenization of the multidialect-bert-base-arabic model.
    => BLEU score = 20,88%

> To translate words out of vocabularies, another input configuration has been adapted: segment data into subword units using Byte Pair Encoding (BPE).
  => BPE improves the BLUE score up to 22.76%

Transformer model		
	Words sequence	Subwords of of words
Development set	21.08	**24.07**
Test set	20.88	**22.72**

Learning model with the Vocabulary of sub-words surpasses that trained by the tokenized sequence

MAGES: Modern standard Arabic texts GEnration tool from Social media

■ MAGES is a tool that combines the developed TADID model and TD-MSA NMT model: Given a corpus taken from social networks, TADID model makes it possible to identify the MSA and the TD comments and it filters the tag other for other dialects. It translates the TD comments to the MSA and leaves the comments written in MSA intact.

■ To evaluate the MAGES tool, another test set was used: It contains 1406 comments: 500 parallel sentences TD-MSA used in (Kchaou et al., 2020) and 406 comments in other languages.

■ From the 1406 comments, MAGES generates 444 sentences in MSA among 500 MSA comments, i.e. an accuracy of 93%.
It has correctly identified 410 TD comments,

■ The tagged comments with the MSA class are passed to the output of the system, and the identified TD comments are passed to NMT model.

=> The MAGES tool allows to standardize Tunisian comments in MSA whether written in MSA or TD.
Comments written with other languages are eliminated.

MAGES evaluation on the application of sentiment analysis

■ The main objective of MAGES tool is to :
  > facilitate the creation of parallel corpus,
  > allow the application of MSA linguistic resources such as sentiment analysis

■ Effect of MAGES on sentiment analysis of dialect textual content in social networks:

> The 1406 comments of the test set are tagged by 3 classes(Positive, Negative, Neutral):

Test Data	#Positive	#Negative	#Neutral
TD	69	90	341
MSA	83	74	343
Other	30	50	326
TOTAL	182	223	1010

Distribution of sentiment classes in test corpus

> Two pre-trained models CAMeLBERT (Inoue et al., 2021) are used in order to compare sentiment analysis of TD comments and their correspondence in MSA:
* CAMeLBERT-AD for sentiment analysis of TD (or Mages system input)
* CAMeLBERT-MSA model for sentiment analysis of MSA texts (Mages system output)

F-mesure score of the Sentiment analysis model		
	CAMeLBERT-AD	CAMeLBERT-MSA
System input	**33.92**	29
System output	43	**49.10**

Evaluation of CamelBert model on the test corpus

=> F-measure on the system output in MSA is more efficient than the F-measure on the system input in TD

=> Whatever the BERT model used, the best result is obtained on the MSA data in the output of the MAGES system.

=> Results show that the approach of standardization of dialect content is better than that of independent treatment of Arabic dialects,

Conclusion

■ Parallel corpus containing 64k parallel sentences is created in which 2k parallel sentences TD-MSA are manually built and are made available to researchers.

■ An identification model for TD and MSA from a corpus scraped from social networks is proposed.

■ A model to standardize the written TD texts in social networks in order to facilitate the computational analysis of poorly endowed languages is proposed in this work,

■ An MSA text generation tool (MAGES) is created in order to develop a sentiment analysis model for TD.

Future work:

■ Introduce the written comments in Arabizi: Arabic dialect written in Latin script.

■ Exploit other advanced pretraining methods, in order to translate TD into a foreign language like English or French.

■ Investigate the effectiveness of the proposed techniques on other Arabic dialects.

Bibliographical References

1, Kchaou, S., Bougares, F., and Hadrich-Belguith, (2019). LIUM-MIRACL participation in theMADAR Arabic dialect identification shared task. In Proceedings of the Fourth Arabic Natural Language Processing Workshop, pages 219–223, Florence, Italy.

2, Issa, E., AlShakhori1, M., Al-Bahrani, R., and Hahn Powell, G. (2021). Country-level Arabic dialect identification using RNNs with and without linguistic features. In Proceedings of the Sixth Arabic Natural Language Processing Workshop, pages 276–281, Kyiv, Ukraine (Virtual)

3, Kchaou, S., Boujelbane, R., and Hadrich-Belguith, L. (2020). Parallel resources for Tunisian Arabic dialect translation. In Proceedings of the Fifth Arabic Natural Language Processing Workshop, pages 200–206, Barcelona, Spain (Online),