Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 5436–5443
Marseille, 20-25 June 2022

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
samehkchaou4@gmail.com, rahma.boujelban@gmail.com, emnafsih@gmail.com, lamia.belguith@fsegs.usf.tn

Abstract

With the growing access to the internet, the spoken Arabic dialect language becomes an informal languages written in social media. Most users post comments using their own dialect. This linguistic situation inhibits mutual understanding between internet users and makes difficult to use computational approaches since most Arabic resources are intended for the formal language: Modern Standard Arabic (MSA). In this paper, we present a pipeline to standardize the written texts in social networks by translating them to the MSA. We fine-tune at first an identification bert-based model to select Tunisian Dialect (TD) comments from MSA and other dialects. Then, the resulting comments are translated using a neural translation model. Each of these steps was evaluated on the same test corpus. In order to test the effectiveness of the approach, we compared two opinion analysis models, the first is intended for the Sentiment Analysis (SA) of dialect texts and the second is for the MSA texts. We concluded that through standardization we obtain the best score.

Keywords: Dialect Identification, Neural Machine Translation, Sentiment Analysis, Tunisian Dialect, Modern Standard Arabic.

1. Introduction

The Arab World is a collection of 21 countries covering Arabia, the Near East and North Africa, having in common an Arabic culture and the Arabic language. The Arabic language is marked by an important diglossia between the literal Arabic, a mainly written vehicular language, and Arabic Dialect (AD), a mainly oral vernacular language. Literal Arabic includes Classical Arabic (pre-Quranic, Quranic, and post-Quranic) and Modern Standard Arabic (MSA). During the last decade, the linguistic situation in the Arab world has been greatly changed. Indeed, with the increasing use of the internet, social media has been invaded by content written with the spontaneous spoken language of each user which is the dialect. This change has led to the appearance of written forms of the AD on internet platforms and has produced gigantic volumes of data written in various dialects. Each of these AD has its linguistic properties due to the absence of a standard norm for its writings. So, the treatment of AD is a bit difficult. In effect, MSA presents the formal language for all Arabic dialects, but they are different at all linguistic levels. The existing resources of MSA are rendered unable to process the AD. The main issue in this context, is how to understand and treat writing published in social networks to be able to use them in other tasks such as opinion analysis and multiple Natural Language Processing (NLP) related applications. Indeed, two solutions can be envisaged. Either consider a dialect as a whole language and develop for its own NLP tools or Migrate the AD to the official language (MSA) to be able to take advantage of its wealth of resources. In this work, we opt for the second solution by treating the case of the Tunisian dialect (TD). In fact, in social networks, TD comments are often mixed with other languages, including other AD, MSA and foreign languages. For this, it is essential to identify TD comments before translating them to MSA. Several works have focused on the treatment of Arabic dialects and in particular the Tunisian dialect. For example (Zribi et al., 2017) study the morphological analysis of TD. (Boujelbane et al., 2013) and (Hamdi et al., 2013) have proposed a rule-based method to translate oral TD transcripts to the MSA. These methods cannot be effective given the richness, the morphological lexical variability and also the spelling of the textual content in social networks. With regard to identification, existing works show that this task needs more works given the number of classes (i.e. 21 classes) or the lack of annotated resource. In this work, we contribute to build a TD identification model and a TD-MSA translation model. Then we use them to build a tool that generates, from the texts of social networks, a text in standard language. The remainder of this paper is structured as follows: In Section 2, we review the related work on AD identification and MT of TD. We study, in Section 3, the linguistic situation of TD on social media. Section 4 is intended for the description of TD-MSA identification model. We depict, in Section 5, the TD-MSA translation model. The description of the standardization tool is presented in sections 6. In Section 7, we conclude and we outline our future work.

2. Related works

In this section, we present a brief state of the art on **A**rabic **D**ialect **ID**entification (ADID). Then, we present related works on NMT of AD.

2.1. Arabic Dialect identification

Although AD is considered as an under-resourced language, there are several ADID systems built using both classical approaches and deep learning approaches. In this context, multiple shared tasks are created such as the VarDial evaluation campaign 2017/2018, MADAR Shared Task 2019 and NADI 2020/2021. Among the works of these evaluation campaigns which used classical approaches, we cite the work of (Nayel et al., 2021) who learned 5 classical models using Naïve Bayes, Support Vector Machine, Decision Tree, Logistic Regression and Random Forest classiers. On NADI 2021, dataset[1] have been labeled at the level of 21 Arab countries and 100 provinces. The best score was obtained by Naïve Bayes classifier and TF/IDF with unigram word as features extraction method. (Aliwy et al., 2020) combined three types of classifiers (Naïve Bayes, Decision Tree and Logistic Regression) to build an ADID system using data of NADI 2020[2] grouped on 21 classes. (Kchaou et al., 2019) performed several ADID experiments on MADAR corpus[3] which is composed by 26 AD. Authors, trained a Multinomial Naive Bayes classifier using Word and character n-gram features as a baseline. Then, they evaluated the performance of the ADID model using only n-gram word and character level Language Models (LM). After this, they integrated LM scores as an extra feature to the baseline system. The best score was obtained using Word and character 5-gram features. On the same data of MADAR corpus, (Talafha et al., 2019) proposed a simple model with LinearSVC classifier (SVM), and they showed that SVM can surpass any deep learning model with a well-selected set of descriptors such as with the incorporation of certain words and using the Bag of word technique TF/IDF. Moreover, deep learning technology has recently been well used and has shown efficient results. Several works have explored the potential of this technology to develop their identification model. For example, (Issa et al., 2021) built two models, the first using pretrained CBOW word embedding with an LSTM architectures, and the second using linguistic features as low-dimensional feature embedding fed through a simple feed-forward network. These two models were intended to classify Arabic tweets from NADI on country level. The experience detected that rare linguistic features do not enrich the efficiency of an LSTM with pretrained CBOW word embedding. In general, most of the classic deep learning techniques have presented limitations when considering the classification issues of Arabic dialects. This can be due to the lack of huge annotated data. For this, several works have opted for transfer learning methods that have proven to be an efficient approach in classification tasks, especially when there is not enough training data. For example, we cite the work of (AlKhamissi et al., 2021) for NADI shared task 2021 where authors developed a model with a set of variants built by MARBERT transformers on the country level dataset. (Beltagy et al., 2020) showed that fine-tuning pretrained model to specific domain is also an efficient solution. In this context,(Talafha et al., 2020a) trained a transformer model baptized Multi-dialect-Arabic-BERT using a fine tuning of Arabic BERT pre-trained language model used in (Safaya et al., 2020). This solution has obtained as the first in the NADI competition 2020.

2.2. Neural machine translation of Arabic Dialect

Different architectures have been proposed in order to build NMT models. Indeed, on a large size of a parallel corpus, the deep learning methods show good results compared to other statistical or linguistic methods for the task of machine translation. Seeing the lack of resources for poorly endowed languages, few works are in line with the AD translation trend. For example, (Al-Ibrahim and Duwairi, 2020) applied a Recurrent Neural Networks (RNN) encoder-decoder model on a small corpus manually prepared in order to translate Jordanian Arabic dialect towards MSA. (Almansor and Al-Ani, 2017) transformed the Egyptian dialect to MSA using Skip-gram and Continuous Bag of Words (CBOW) translation models. (Baniata et al., 2018) presented a multi-task learning model using an RNN which shares one decoder among language peers. They used a limited corpus of the Levantine dialects, such as Jordanian, Syrian and Palestinian, from **P**arallel **A**rabic **DI**alectal **C**orpus (Karima et al., 2015) (PADIC) and from a **M**ultidialectal **P**arallel **C**orpus of **A**rabic (MPCA) (Bouamor et al., 2014). For TD,to our knowledge there is until now no neural translation model available. Existing approaches for TD translation are linguistic approaches like that presented in(Hamdi et al., 2013) and statistical approaches like that presented in (Kchaou et al., 2020).

3. Tunisian Dialect and its Challenges in Social Media

Tunisian Dialect (TD) is part of the Maghrebi dialects, generally known as the ”Derja Tounsi” to distinguish it from MSA. The latter has a dynamic status and are not used for official writing. It presents a form of informal Arabic even though it is written with Arabic letters, and it presents several challenges for NLP applications in general and in MT in particular. Among these challenges, we cite:

- **Arabizi scripts:** The TD written by social media users is influenced by other non-Arabic languages and the SMS language. In fact, each user writes in his own way. When writing, Internet users tend to use Latin letters and substitute some letters with

[1] https://sites.google.com/view/wanlp2021

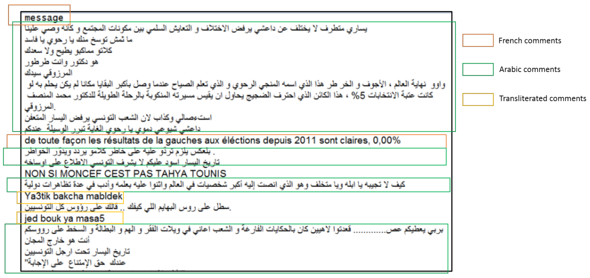
[2] https://sites.google.com/view/wanlp-2020/home

[3] https://camel.abudhabi.nyu.edu/madar-parallel-corpus/

numbers For example, the Arabic letter ”ه” (”h”) is written as 7 by some users and with the Latin letter h by others.

- **Linguistic ambiguity:** The morphology, syntax and vocabulary of TD are quite different from MSA. Indeed TD is characterized by non-standard syntax rules. In addition there is no spelling convention for their writing. For example, the Arabic sentence ”لم يدهب اطفل الي المدرسة” [4] becomes in TD side ”اطفل مَشَّاش لّكتب”. We notice that the adverb لم associated with the MSA verb “يدهب”, becomes in the verb TD an adverb ”م” and followed by the adverb ”ش”. We also notice a difference in syntax: in the MSA sentence, the verb is followed by a noun while in the TD sentence, the noun is followed by the verb.

According to these ambiguities, the same comments can be written in different ways. We tag in Figure 1 TD comments taken from social media with the the ambiguities mentioned above.



message
يساري متطرف لا يختلف عن داعشي يرفض الاختلاف و التعايش السلمي بين مكونات المجتمع و كانه وصي علينا
ما ثمش توسخ منك يا رحوي يا فاسد
كلاتو مماكبو يطيح ولا سعدك
هو دكتور وانت طرطور
المرزوقي سيدك
واوو نهاية العالم ، الأجوف و الخر طر هذا الذي اسمه المنجي الرحوي و الذي تعلم الصياح عندما وصل بأكبر البقايا مكانا لم يكن يحلم به لو كانت عتبة الانتخابات 5% ، هذا الكائن الذي احترف الضجيج يحاول ان يقيس مسيرته المنكوبة بالرحلة الطويلة للدكتور محمد المنصف المرزوقي.
استءصالي وكذاب لان الشعب التونسي يرفض اليسار المتعفن
داعشي شيوعي دموي يا رحوي الغاية تبرر الوسيلة عندكم
de toute façon les résultats de la gauches aux éléctions depuis 2011 sont claires, 0,00%
بلعكس يلزم تردو عليه على خاطر كلامو يردد ويدور الخواطر .
تاريخ اليسار اسود عليكم لا يشرف التونسي الاطلاع على اوساخه
NON SI MONCEF CEST PAS TAHYA TOUNIS
كيف لا تجبيه يا ابله ويا متخلف وهو الذي انصت إليه أكبر شخصيات في العالم واثنوا عليه بعلمه وأدب في عدة تظاهرات دولية
Ya3tik bakcha mabldek
سطل على روس البهايم اللي كيفك .. فالك على رؤوس كل التونسيين.
jed bouk ya masa5
بربي يعطيكم عص............ قعدتوا لاهيين كان بالحكايات الفارغة و الشعب اعاني في ويلات الفقر و الهم و البطالة و السخط على رووسكم
أنت هو خارج المجان
تاريخ اليسار تحت ارجل التونسيين
"عندك حق الإمتناع على الإجابة

Figure 1: Forms of comment in social media.

4. Resources to build Dialect-MSA Identification Model

In this work, the construction of a Tunisian Arabic dialect identification **(TADID)** model has two main goals: the first aim is to build annotated resources in TD. The second is intended to locate the TD sentences to be able to translate them to the MSA.

4.1. Corpus for Dialect Identification

In order to train a TADID model, we exploit the existing corpora, and adjust them with the desired classes. Indeed, we first use the corpus proposed in (Kchaou et al., 2020). The latter is a parallel corpus containing 9.7k parallel sentences TD-MSA built from the MADAR[5], Padic[6] corpora and the Tunisian constitution (CONT-TUN). It also includes 900 TD comments manually translated by native speakers into MSA. This size of the corpus was increased to 32k parallel sentences using an augmentation method based on sentence segmentation at stops words level. Since comments on social networks include, apart TD, other languages such as Magrebien dialect, Algerian dialect, English and French language, we add monolingual AD data from NADI corpus, which is an annotated corpus at country-level, to include more dialects. We preprocess the class column of this corpus as following: we keep the annotated tweets with TD labels and we assign the label ’other’ to all the other comments. The resulted corpus contains 95k annotated comments with 3 classes. Table 1 describes the re-partition of comments according to used classes. The length of each comment varies between 1 and 350. Figure 2 shows the variation of the number of words in each comments in the proposed corpus.

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

Table 1: TADID corpus statistics.

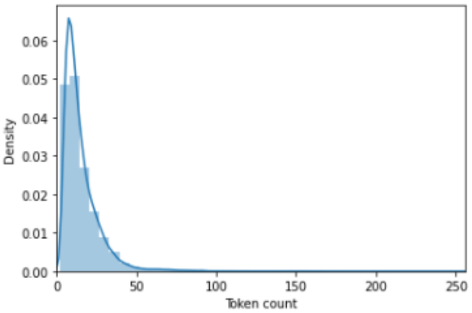


Figure 2: Tokens variation in the sentence.

4.2. Data pre-processing

In order to improve the quality of the corpus and consequently of the learning model, we propose the following pre-processing in our corpus:

- The removal of non-Arabic words to delete vocabularies in other language,
- The removal of numbers and all possible punctuation marks,
- The removal of tags that can appear in social network comments or in tweets of the NADI corpus,
- The removal of special characters and emoticons,
- The deletion of diacritics from written word in Arabic language,
- The normalization of few Arabic characters to unify them into only form.

[4] Translation in English: The child does not go to school
[5] https://sites.google.com/view/madar-shared-task/home
[6] https://sites.google.com/site/torjmanepnr/6-corpus

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

Table 2: Evaluating classifiers on the dataset using Word feature.

4.3. TADID model

Several methods have been proposed to implement ADID systems using different approaches. Recently transfer learning has shown a great efficiency compared to Deep learning approaches in context of dialect identification. Classical approaches also show a good performance to identify AD. For this, we propose to apply these two approaches on the constructed corpus. Thus, we divide it into 3 sets: 80% for the train, 10% for the development (Dev) and 10% for the test set.

Traditional approaches: We train 3 classifiers using Word features: the Naive Bayes (NB), support vector machine (SVM) and Multi-layer Perceptron (MLP) classifier. As it is shown in Table 2, the best score is obtained by **N**aive **B**ayes **C**lassifier (NBC). We try to reproduce the training of this classifier using Word and character n-gram features. We also use TermFrequency-Inverse Document Frequency (Tf-Idf) scores learned on extracted n-grams character ranging from 1 to 3-grams. Table 3 reports the results of NBC model on the development set and test set. The best identification accuracy score is produced using NBC with uni-gram word level and 3-gram character level features.

	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**

Table 3: NBC evaluation using n-gram features.

Transfer learning approaches: The trend that shows high accuracy for ADID domain is actually transfer learning. In this context, BERT and GPT-2 are the most popular models based on transformers. In order to identify TD, we opt for using this pre-trained language models for Arabic based on Bidirectional Encoder Representation Transformers (BERT) architecture . BERT has two original models: BERT-base, which has 12 encoder layers, 768 for hidden size and 12 multi-head attention heads; and BERT-large, which has 24 encoder layers, 1024 for hidden size and 16 multi-head attention heads. They were created on a large English corpus, and then, several other BERT models dedicated to the specific language were developed. In this paper, we adapt to TD a BERT identification model pretrained on arabic data. For this, we compare different available models namely; bert-large-arabic (Safaya et al., 2020), bert-base-arabic (Safaya et al., 2020), albert-base-arabic (Safaya, 2020) and Multi-dialect-Arabic-BERT, in order to select the closest model to the TD. A brief comparison of these two models on test set is described in Table 4. Through this comparison, we test the albert-base-arabic model (Safaya, 2020) which is an evolutionary model of BERT, it uses reduced number of parameters compared to BERT. The evaluation of ALbert model on the test set is presented in Table 4. As shown in the Table, bert-base-arabic model gives a higher score than the ALBERT model despite the two models being driven mainly by MSA data and by AD. Then, we evaluate the Multi-dialect-Arabic-BERT (Talafha et al., 2020b) model which is intended for country level ADID. According to the results of tested models with Bert family, Multi-dialect-Arabic-BERT pre-trained model show a good almost human identification precision. For this, we chose this Transformer model for building our TADID system. We use the following configuration for this system: 3 epochs, a batch size of 10, maximum sentence length fixed on calculated maximum tokens number in the sentences and AdamW optimizer with lr = 2e-5.

	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**

Table 4: Transformer classifier evaluation using BERT family.

BERT TADID model: since it shows the highest identification score in the previous experiments, we built a TADID model using the Multi-dialect-Arabic-BERT. Indeed, in addition to the BERT model layers, we add another layer to combine a hidden single layer neural network classifier with the last layer of the BERT model. Each layer of the BERT contains a list of token integration and produces the same amount of integration with the same size hidden on the output. The output of the token's final model layer is used as a feature of the sequence to load our classifier. Using the added layer, the Accuracy score increased from 88.82 to 93.18% on test data. Figure 3 presents the output confusion matrix of our TADID system. As shown in confusion matrix, all Tunisian sentences are well predicted.

5. Resources to build TD-MSA Translation model

In the following, we first present the intended corpus for NMT of TD-MSA, and second, we describe our

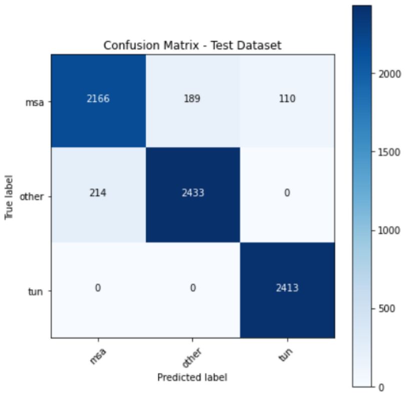


Figure 3: Output confusion matrix of TDID system.

proposed NMT model based on Transformer architectures and its results.

5.1. TD-MSA parallel corpus

5.1.1. Social media corpus

In order to build a TD-MSA parallel corpus, we deployed the TADID model on a dataset of untagged comments scraped from YouTube and Facebook to filter them and predict Tunisian comments. The system allows to predict 2k comments in TD among 6.5k downloaded public comments. Then, since we are native speakers, we manually translated these sentences into MSA. Table 5 presents statistics of the created comments in the source and the target language. In the TD side, the vocabulary is 10.7k words whereas it is 9k words in the MSA side.

Social medi corpus	#Lines	#Tokens	#Vocabularies
TD comments	2k	22k	10.7
MSA comments	2k	23.6k	9k

Table 5: Social media corpus statistics.

5.1.2. Corpus distribution

Given the lack of parallel corpus for TD data, the translated comments will be freely distributed for the research community [7]. This will enable researchers to use the TD-MSA corpus and, to push forward the research of machine translation of AD and other researches in NLP. The parallel corpus will be a folder that consists of two files. The first contains the MSA comments and the second contains the corresponding alignment of TD comments.

5.1.3. Substitution augmentation method for TD corpus

NMT models have shown impressive results on large size of a bi-text corpus, that's why we propose an augmentation method at the word level using the multi-dialect-bert-base-arabic language model (Talafha et al., 2020b). In fact, we generate different TD sentences from the original TD sentences of the corpus without losing meaning of sentence by substiting words with their synonyms. The chosen words for substitution are selected randomly. This augmentation method makes it possible to obtain 68k parallel sentences. Table 6 shows more details on the corpus statistic. We preprocess this corpus using the same pre-processing steps applied in TADID corpus.

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

Table 6: The statistics of the created TD-MSA corpus for NMT model.

5.2. TD-MSA NMT model

We use a self-attention-based transformer model which has the ability to press at different states of the input sequence to estimate a representation of that sequence. It is based on a self-attention mechanism to detect the latent space representations of input and output without the need of Recurrent/convolution neural networks of the sequence. It is based on an Transformer-encoder, Transformer-decoder and a linear layer. The input sentences are passed through the N encoder layers which produces an output signal for each word/token in the sequence. The decoder received the output of the encoder and its input (self-attention) to predict the next word. The output of the decoder has a linear layer input and its output is returned. To train this model, we divide the preprocessed corpus, intended for the NMT task, into 90% for the train and 10% for the test set. To configure the input for training model, data is encoded into token ID sequences using the tokenization of the multi-dialect-bert-base-arabic model. We use small hyperparameters for its configuration since we don't have a large corpus: num_layers= 6, d_model= 128, dff= 512, num_heads= 8 and a dropout_rate= 0.1 The Adam optimizer is used with the parameters beta1 = 0.9, beta2 = 0.98 and epsilon = 10e-9. We test this model on several epoch number values in order to optimize the model. For each epoch, we tested different values of batch size (514, 64, 32). We retain the value 512 as it leads to the best result in each epoch. The best BLEU score reached 20.88%, it is obtained with 30 epochs.

We notice, in translated sentences, a large number of rare words that the model cannot translate. Rare words outside the vocabulary are presented by the symbol [UNK]. To solve this problem and improve the translation quality, we investigate the created transformer on the level of subword units in order to encode these rare words. For this, we segment the train, development and test set, in both languages, into subword units using Byte Pair Encoding (BPE). Indeed, BPE (Gage, 1994) is an easy data compression method that iteratively replaces the more frequent pair of bytes in a sequence

[7] https://github.com/sk-cmd/Parallel-data-TD-MSA

with a one unused byte. We reconfigured our created transformer model with all vocabulary which contain all needed subwords to represent the source and target training data. We use the subword_nmt library to segment words into subwords (BPE) according to their frequency in the corpus. This technique i.e. BPE improves the BLEU score up to 22.76%. Table 7 shows the obtained BLEU scores on the test set using two different configuration of transformer model. As it is shown in Table 7, learning model with the vocabulary of sub-words surpasses that trained by the sequence of words. The technique of sub-words was previously reported to be helpful for low-resource machine translation in NMT (Richburg et al., 2020). Therefore, we consider the NMT model that the one configured with the vocabulary of sub-words.

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

Table 7: BLEU scores of Transformer model for each configuration.

6. MAGES: Modern standard Arabic texts GEnration tool from Social media

In order to implement the developed models, we develop a tool that cascades the two models as follows: Given a corpus taken from social networks, the identification model makes it possible to identify the MSA and the TD texts and it attributes the tag other for the other dialects. Then, it translates the TD comments to the MSA and leaves the comments written in MSA intact. Comments with the tag other are not processed for the moment. They will be treated in future works. Figure 4 describes the pipeline that we have included in the MAGES tool. The different models of this system have been evaluated on a corpus containing 1406 comments: 500 parallel sentences TD-MSA used in (Kchaou et al., 2020) and 406 comments in other languages.

Evaluation of the identification module: From test corpus, the system generates 444 sentences in MSA among 500 MSA comments, i.e. an accuracy of 93%. Indeed, it has correctly identified 410 TD comments.

Evaluation of the translation module: The tagged comments with the MSA class are passed to the output of the system, and we project the NMT model on the identified TD comments. The translation module achieves a BLEU score of 20.63 (while it achieved 22.72 on test set in its configuration). This can be explained by the lack of social networks data in the test data because the used translation model seems to perform well, its BLEU score largely surpassed that of the proposed statistical model in (Kchaou et al., 2020) when it was tested on the same test set of (Kchaou et al., 2020), it increased from 15 to 22.72.

Sentiment analysis of dialect content in social networks: The main objective of the developed system is to facilitate the creation of parallel corpus on one hand and on the other hand to allow the application of MSA linguistic resources such as sentiment analysis. In this work, we highlight the effect of the proposed pipeline on sentiment analysis on dialect textual content in social networks. For this, we manually annotate the test set of this study by 3 classes: neutral, positive and negative. Table 8 describes the distribution of the data according to used classes. We use the collection of pre-trained models CAMeLBERT (Inoue et al., 2021). We studied, on the annotated test set, the difference between two opinion analysis models: CAMeLBERT-AD which is intended for SA for dialect texts (1) and CAMeLBERT-MSA model which is designed for the SA of MSA texts (2). The first (1) achieved an F-mesure of 33.92% on the input data and the second achieved an F-mesure of 49.10% on the output data. In order to show the efficiency of each model, we also test the SA CAMeLBERT-MSA model on the system input in TD, and conversely the CAMeLBERT-AD model on the text in MSA. The results show, whatever the used model, the F-mesure on the system output in MSA is more efficient than the F-mesure on the system input in TD. Table 9 presents the F-mesure score of each model. The results show that the approach of standardization of dialect content is better than that of independent treatment of Arabic dialects.

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

Table 8: Distribution of sentiment classes in test corpus

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

Table 9: Evaluation of CamelBert model on the test corpus.

7. Conclusion

In this paper, we proposed a pipeline to standardise the written TD texts in social networks in order to facilitate the computational analysis of poorly endowed languages. For this, different resources have been developed. We proposed at first an identification model for TD and MSA from a corpus scraped from social

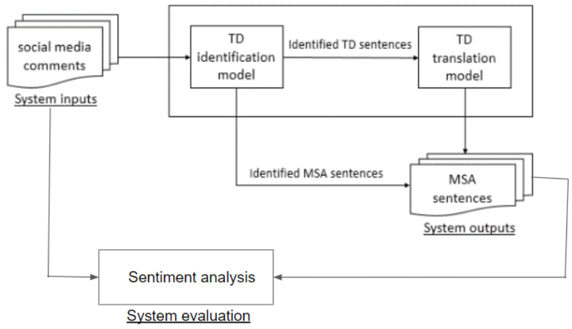


Figure 4: Steps for translating social network comments to the MSA.

networks containing a mixture of Arabic dialects and MSA. Secondly, we built a parallel corpus containing 64k parallel sentences in which 2k were manually built. On this corpus, we learned a neural translation model. The cascade of these two models gave rise to an MSA text generation tool. We have shown the benefits of the proposed pipeline through an application of a sentiment analysis model. For future work, we plan to introduce the written comments in Arabizi: Arabic dialect written in Latin script. We would like also to exploit other advanced pre-training methods, in order to translate TD into a foreign language like English or french for example. Another perspective would be to investigate the effectiveness of the proposed techniques on other Arabic dialects.

8. Bibliographical References

Al-Ibrahim, R. and Duwairi, R. M. (2020). Neural machine translation from jordanian dialect to modern standard arabic. In *2020 11th International Conference on Information and Communication Systems (ICICS)*, pages 173–178.

Aliwy, A., Taher, H., and AboAltaheen, Z. (2020). Arabic dialects identification for all Arabic countries. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 302–307, Barcelona, Spain (Online), December. Association for Computational Linguistics.

AlKhamissi, B., Gabr, M., ElNokrashy, M. N., and Essam, K. (2021). Adapting MARBERT for improved arabic dialect identification: Submission to the NADI 2021 shared task. *CoRR*, abs/2103.01065.

Almansor, E. H. and Al-Ani, A. (2017). Translating dialectal Arabic as low resource language using word embedding. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 52–57, Varna, Bulgaria, September. INCOMA Ltd.

Baniata, L. H., Park, S., and Park, S.-B. (2018). A neural machine translation model for arabic dialects that utilizes multitask learning (mtl). In *Computational Intelligence and Neuroscience.*, page 10.

Beltagy, A., Abdelrahman, W., and ElSherief, O. (2020). Arabic dialect identification using bert-based domain adaptation. *CoRR*, abs/2011.06977.

Bouamor, H., Habash, N., and Oflazer, K. (2014). A multidialectal parallel corpus of Arabic. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1240–1245, Reykjavik, Iceland, May. European Language Resources Association (ELRA).

Boujelbane, R., Ellouze Khemekhem, M., and Belguith, L. H. (2013). Mapping rules for building a Tunisian dialect lexicon and generating corpora. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 419–428, Nagoya, Japan, October. Asian Federation of Natural Language Processing.

Gage, P. (1994). A new algorithm for data compression. *The C Users Journal archive*, 12:23–38.

Hamdi, A., Boujelbane, R., Habash, N., and Nasr, A. (2013). The Effects of Factorizing Root and Pattern Mapping in Bidirectional Tunisian - Standard Arabic Machine Translation. In *MT Summit 2013*, page pas d'édition papier, France, September.

Inoue, G., Alhafni, B., Baimukan, N., Bouamor, H., and Habash, N. (2021). The interplay of variant, size, and task type in Arabic pre-trained language models. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, Kyiv, Ukraine (Online), April. Association for Computational Linguistics.

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), April. Association for Computational Linguistics.

Karima, M., Salima, H., S, J., M, A., and Kamel, S. (2015). Machine translation experiments on padic: A parallel arabic dialect corpus. In *Proceedings of 29th Paclic Asia Conference on Language, Information and Computation.*, pages 26–34.

Kchaou, S., Bougares, F., and Hadrich-Belguith, L. (2019). LIUM-MIRACL participation in the MADAR Arabic dialect identification shared task. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 219–223, Florence, Italy, August. Association for Computational Linguistics.

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), December. Association for Computational Linguistics.

Nayel, H., Hassan, A., Sobhi, M., and El-Sawy, A. (2021). Machine learning-based approach for Arabic dialect identification. In *Proceedings of the*

Sixth Arabic Natural Language Processing Workshop, pages 287–290, Kyiv, Ukraine (Virtual), April. Association for Computational Linguistics.

Richburg, A., Eskander, R., Muresan, S., and Carpuat, M. (2020). An evaluation of subword segmentation strategies for neural machine translation of morphologically rich languages. In *Proceedings of the The Fourth Widening Natural Language Processing Workshop*, pages 151–155, Seattle, USA, July. Association for Computational Linguistics.

Safaya, A., Abdullatif, M., and Yuret, D. (2020). KUISAIL at SemEval-2020 task 12: BERT-CNN for offensive speech identification in social media. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 2054–2059, Barcelona (online), December. International Committee for Computational Linguistics.

Safaya, A. (2020). Arabic-albert, August.

Talafha, B., Farhan, W., Altakrouri, A., and Al-Natsheh, H. (2019). Mawdoo3 AI at MADAR shared task: Arabic tweet dialect identification. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 239–243, Florence, Italy, August. Association for Computational Linguistics.

Talafha, B., Ali, M., Za'ter, M. E., Seelawi, H., Tuffaha, I., Samir, M., Farhan, W., and Al-Natsheh, H. T. (2020a). Multi-dialect arabic BERT for country-level dialect identification. *CoRR*, abs/2007.05612.

Talafha, B., Ali, M., Za'ter, M. E., Seelawi, H., Tuffaha, I., Samir, M., Farhan, W., and Al-Natsheh, H. T. (2020b). Multi-dialect arabic bert for country-level dialect identification.

Zribi, I., Ellouze, M., Belguith, L., and Blache, P. (2017). Morphological disambiguation of tunisian dialect. *J. King Saud Univ. Comput. Inf. Sci.*, 29:147–155.