A Deep Learning Approach for the Romanized Tunisian Dialect Identification

Jihene Younes[1], Hadhemi Achour[1], Emna Souissi[2], and Ahmed Ferchichi[1]
[1]Université de Tunis, ISGT, Tunisia
[2]Université de Tunis, ENSIT, Tunisia

Abstract: *Language identification is an important task in natural language processing that consists of determining the language of a given text. It has increasingly picked the interest of researchers for the past few years, especially for code-switching informal textual content. This paper, focuses on the identification of the Romanized user-generated Tunisian dialect on the social web. Segmented and annotated a corpus extracted from social media and propose a deep learning approach for the identification task. A Bidirectional Long Short-Term Memory neural network with Conditional Random Fields decoding (BLSTM-CRF) had been used. For word embeddings, a combination of word-character BLSTM vector representation and Fast Text embeddings that takes into consideration character n-gram features. The overall accuracy obtained is 98.65%.*

Keywords: *Tunisian dialect, language identification, deep learning, BLSTM, CRF, natural language processing.*

Received August 25, 2019; accepted April 28, 2020
https://doi.org/10.34028/iajit/17/6/12

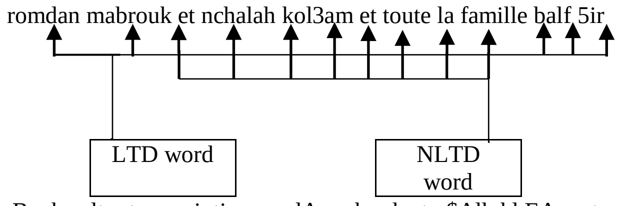
1. Introduction

The language situation in Tunisia is characterized by a huge diversity. Besides the linguistic phenomena resulting of the speaking practices and regional dialects, we witness the presence of foreign languages like French and English. This diversity is mainly due to historical, cultural and political reasons [14].

Modern Standard Arabic (MSA) is the official language in Tunisia. It's the language taught in schools and used in administration, press and state media. The Tunisian Dialect (TD) on the other hand remains the predominant oral language. It is naturally spoken by a large part of the Tunisian population. In the social web, users resort to Tunisian dialect in their written exchanges. They often include language loans, especially from French and English. Textual productions on Tunisian social networks are indeed strongly characterized by the code switching and the multilingualism phenomena.

The diversity that Tunisia witnessed especially politically and historically, favoured the emergence of multilingualism. Indeed, Tunisians use more than one language in their social exchanges (Modern Standard Arabic, Tunisian dialect, French, English, etc.,). Code switching[1], on the other hand, refers to the act of changing between two or more languages in the same discourse according to the Cambridge dictionary[2]. This phenomenon was widely discussed by linguists since it may occur in both oral and written conversations [57].

Both multilingualism and code switching make the language processing a non-trivial task, especially for the Tunisian dialect in its written form. The TD used on the social web is fundamentally a spoken language. It lacks available language resources namely corpora, lexica and dictionaries to allow its automatic processing. To create such resources, we need to harvest a large amount of textual content from the social web. Multilingualism and code switching make this task more challenging. The textual content written by Tunisians on the social web includes messages from different languages. The same message may itself include words from several languages. These phenomena occur essentially with the Tunisian dialect written with the Latin script (LTD), which we also refer to as Romanized Tunisian dialect. Non-LTD (NLTD) refers, in this paper, to the written content that doesn't belong to the Latin Tunisian dialect. Figure 1 shows an example.


romdan mabrouk et nchalah kol3am et toute la famille balf 5ir
LTD word
NLTD word

Buckwalter transcription: rmdAn mbrwk et n$Alh kl EAm et toute la famille b Alf xyr
Arabic transliteration: رمضان مبروك et نشاله كل عام et toute la famille ب الف خير
Translation: Ramadan Mubarak and I wish good health for all your family.

Figure 1. Example of an LTD message.

[1]Also called language alternation

[2]https://dictionary.cambridge.org/dictionary/english/code-switching

The message shown in Figure 1 illustrates the alternation between LTD words and words from other languages NLTD in the same message. Dealing with such language productions is difficult since we cannot effortlessly identify the point of switching between languages, neither the language of each word. In addition, practitioners of the Romanized TD on social media don't rely on defined syntactic or orthographic rules in their writings.

Another difficulty that we may encounter when dealing with the language identification task, is the ambiguity at a word level. In fact, the same word can belong to more than one language and bear different meanings. In the example of Figure 1, the word "la" is a French word that means "the" in this context. In other contexts, it can be considered as an LTD word that means "no" ("لا").

All these factors led us to tackle, in this paper, the word-level identification of the Tunisian dialect written on the social web using the Latin script. The automatic LTD detection, not only allows constructing sizable adequate language resources for the TD language processing, but can also be a crucial step for any NLP application treating the Tunisian dialect on the social web, such as sentiment analysis, topic detection, machine translation, etc., In order to automatically recognize the LTD words within multilingual and code-switched textual productions, we propose a deep learning-based solution. In fact, while approaches to NLP tasks based on deep learning show that they reach state-of-the-art results [8, 44, 49], they remain still little explored when dealing with the Arabic dialects and with the Tunisian dialect in particular [60]. In this work, we propose to use a word-character Bidirectional Long Short-Term Memory model with Conditional Random Fields decoding (BLSTM-CRF), using word embeddings that take into account character n-gram features.

The remainder of this paper is organized as follows: section 2 is devoted to a literature review on the language identification of code-switching texts. We motivate our work in section 3, then in section 4, we describe the datasets we used for the Latin TD identification task, and present a brief study of the code-switching and multilingualism phenomena in our corpus. Section 5 presents the proposed deep learning model and finally, section 6 is dedicated to experiments and results.

2. Related Work

The language Identification (LI) task was extensively studied by NLP researchers. Despite the numerous works on LI, dating back several years, we focus in this paper solely on the recent ones. The works we present are rather interested in the texts incorporating the code-switching phenomenon, dealing with informal short texts. Our study allowed us to distinguish three main types of approaches for the LI task: language modelling, machine learning and deep learning-based algorithms.

2.1. Language Modelling Approaches

A considerable amount of studies on the LI task for code-switched texts relied on language models. The resort to this approach was based on the assumption that each language is characterized by a specific character behaviour and a distinct phonology and morphology [2].

Some works on LI dealt with MSA and Arabic Dialects (AD). This was the case of Zaidan and Callison-Burch [61] who proposed an approach using a smoothed n-gram model. Elfardy *et al*. [18] presented an identification system for the code-switched MSA-Egyptian words that relies on an MSA morphological analyser. To identify the Romanized form of the Arabic dialects (Arabizi), Eskander *et al*. [19] resorted to a language model for name tagging, to which they introduced a set of features.

As a contribution to the first shared task on language identification on codeswitched data [55], Shreshta [53] focused on the language identification of code-switched Spanish-English and Nepali-English texts. He resorted to an incremental n-gram approach.

Regarding the second shared task on language identification on code-switched data [39], several researchers relied on language modelling approaches such as Chanda *et al*. [10] who focused on the language identification of code-switching English-Spanish tweets. They used an algorithm that generates the word's n-gram and checks its presence in the dictionaries. Shirvani *et al*. [42] focused on the Swahili-English code-switched texts and introduced several features including character n-grams. This approach was adopted by Piergallini *et al*. [43] for code-switched English-Spanishwith 17 new features including Part Of Speech (POS) tags.

In other contexts, Jhamtani *et al*. [29] focused on the word-level identification of code-switched Hindi-English texts. They used a model that relies on several features like character n-grams the neighbouring words' POS tags. They also experimented with several classifiers namely Decision Trees, Support Vector Machines (SVM) and Random Forests. Nguyen and Cornips [40] were interested in identifying code-switched Dutch-Limburgish tweets from a province in the Netherlands using words probabilities. Guellil and Azouaou [23] focused on the language identification of Algerian dialect. Their approach was based on an Algerian lexicon and on improved Levenstein distance.

2.2. Machine Learning Approaches

With the proliferation of textual content, especially on social media, researches on the language identification task have become more oriented to machine learning based approaches. This was the case of Giwa and Davel [20] who resorted to the Naïve Bayes (NB) and SVM to deal with the LI of the code-switched south African words. Sadat *et al*. [46] focused on the identification of 18 Arabic dialects using a character-based n-gram Markov Model and NB.

Several works have been carried out as a contribution to the first shared task on language identification on code switched data [55], such as that of Bar and Dershowitz [4] who focused on the identification of the code-switched English-Spanish tweets using SVM with different features. Barman *et al*. [5] focused on identifying the code-switched words in Bengali, English and Hindi textual contents on social media using SVM and CRF. To identify Nepali-English and Spanish-English code-switched tweets, Barman *et al*. [6] proposed a classification approach using K-Nearest Neighbors (KNN) and SVM. Chittaranjan *et al*. [13] were interested in the LI for the code-switched English-Spanish, English-Nepali, English-Mandarin and MSA-AD texts using CRF with several features. The same languages were identified by King *et al*. [31] with an extension of a Markov Model. In the same context, Lin *et al*. [34] used a baseline CRF model with labelled data and CRF auto encoder with word embeddings and word lists as features with unlabeled data. Papalexakis *et al*. [41] used NB with a set of features to predict code-switching in an online Turkish-Dutch forum for immigrant community in the Netherlands.

The second shared task LI in code-Switched data [39] included several contributions that relied on machine learning models as well. Chanda *et al*. [11] focused on identifying code switched Bengali-English language productions. Part of their experiments were performed using machine learning algorithms namely J48, KNN and Random Forest and different features. Shrestha [52] worked on the identification of code-switched English-Spanish and MSA-AD language productions using CRF with a set of features. Sikdar and Gambäk [54] resorted to CRF as well and focused on the LI of code-switched English-Spanish texts. In the same context, Xia [56] worked on the LI of English-Spanish code-switched tweets using CRF.

The identification of other languages was tackled by several researchers such as Samih and Maier [48] who focused on the MSA-Moroccan code-switched texts using CRF. Schulz and Keller [51] worked on the LI of code-switching in Latin-Middle English textual contents with CRF. Al-Badrashiny and Diab [2] proposed a CRF based approach as well for the LI of English-Spanish, Nepali-English, English-Hindi, Arabizi-English, Arabic-Engari, MSA-Egyptian, Levantine-MSA and Gulf-MSA. Dongen [15] worked on the LI of code-switched Dutch-English language productions in social media using SVM, Decision Trees and CRF. Rijhwani *et al*. [45] performed the LI task on a variety of languages namely Dutch, English, French, German, Portuguese, Spanish and Turkish. They developed an Hidden Markov Model (HMM)-based Generalized Word-level Language Detection system. Aridhi *et al*. [3] worked on the LI of Romanized TD using N-Gram Cumulative Frequency Addition [1] and SVM. Salameh *et al*. [47], focused on the LI of several AD and MSA. They resorted to the Multinomial Naïve Bayes (MNB) and trained, for each dialect, a 5-gram character level language model using KenLM [26]. Lichouri *et al*. [33] focused on the LI for AD and Algerian dialects. They used Linear SVM, Bernoulli Naïve Bayes (BNB) and MNB.

2.3. Deep Learning Approaches

Deep learning is a subset of machine learning that has caught a great interest from researchers in recent years. However, we have identified relatively few works regarding the LI task compared to machine learning and language modelling approaches. Chang and Lin [12] focused on detecting the language of code-switching twitter corpus for the English-Spanish, English-Nepali, Mandarin-English and MSA-Egyptian languages as a contribution to the first shared task on language identification on codeswitched data [55]. The LI was based on the Elman-type and the Jordan-type Recurrent Neural Networks (RNN) with the optional inclusion of pre-trained Word2Vec and character n-gram features.

In the second shared task on language identification in code-switched data [39], Jaech *et al*. [28] focused on the LI of code-switched English-Spanish and MSA-AD tweets. The LI system includes a Convolutional Neural Network (CNN) component that provides words embeddings and a Bidirectional Long Short-Term Memory (BLSTM) component for labelling. Samih *et al*. [49] focused on identifying code-switched English-Spanish and MSA-Egyptian dialect textual contents. They resorted to the LSTM neural network with a CRF layer and used a template including various features.

To identify code-switched Hindi-English and Spanish-English textual content, Mave *et al*. [36] resorted to BLSTM, word-character LSTM and CRF with different features (n-grams, POS tags, etc.,). Mager *et al*. [35] proposed a model based on segmental RNN to identify Spanish-Wixarika and German-Turkish code-switching content.

Regarding MSA and Arabic dialects identification, Elaraby and Abdul-Mageed [17] used three machine learning classifiers: Logistic Regression, Multinomial Naïve Bayes and SVM and resorted to six deep learning models as well namely CNN, LSTM, Contextual LSTM (CLSTM), BLSTM, Bidirectional Gated Recurrent Units (BiGRU) and BLSTM with attention mechanisms.

Only one work, to the best of our knowledge, focused on the Tunisian Dialect in particular (TD), using a deep learning approach, namely that of Sayadi *et al*. [50]. The authors resorted to LSTM RNN to identify TD-MSA and a set of 5 AD-MSA. They used the election data set for the TD language and the multidialect parallel corpus of arabic [9] for the AD languages.

3. Motivation

The previously presented works demonstrate the increasing interest of the NLP community to the language identification task, especially with the proliferation of informal textual content use on social media. We notice that despite the good performance of the language modelling approaches, researchers have been resorting in the last few years to machine learning and deep learning models.

These approaches have proven to be extremely performant notably when we dispose of large amounts of language data. We notice, however, that the exploration of deep learning algorithms is still in progress especially for the TD language.

In fact, we counted only one work using deep learning methods [50] that concerned the identification of TD written in the Arabic alphabet. As for its Latin form, only one work was performed [3] using N-Gram Cumulative Frequency Addition and an SVM model. We should note that efforts were invested in the Tunisian dialect NLP using Neural Networks such as speech recognition [25]. Indeed, LI is still a task that hasn't been fully explored for the Tunisian dialect. In the present work, we focus on the identification of the Latin form of the TD. First, since this form of writing is predominant on the Tunisian social web [58] and second, because deep learning approaches haven't been experimented yet on LTD.

4. Used Data

4.1. Corpus Annotation

The LTD corpus we used in our LI experiments is extracted from that of Younes *et al*. [58]. The initial corpus is composed of Romanized messages collected from Tunisian social web pages. We used 13,656 messages of this corpus including 167,337 words. Our segmentation process was performed automatically at first, considering the space character as a delimiter. We subsequently checked the segmentation manually to correct cases of errors such as attached words, emoticons or symbols. The words were afterwards annotated according to 5 categories:

- LTD: all words belonging to the Tunisian dialect language according to its context within the whole message to which it belongs
- NLTD: foreign languages (English, French, etc.,)
- PUNCT: punctuation marks (Example: ./ ./ ? etc.,)
- SYMB: symbols that Tunisians use in the internet, other than punctuation (Example: + / - / < / > etc.,)
- EMO: emoticons namely any representation of facial expression using symbols and punctuation marks (Example: ☺ / ☹ etc.,)

Figure 2 illustrates an example of annotation on a portion of our corpus.

Message 1:	Ouii ritha chérie !
Buckwalter:	Ouii rythA chérie !
Translation:	Yes I saw it sweety !
Message 2:	b journée Rabi m3ak :* ++
Buckwalter:	b journéer by mEAk :* ++
Translation:	good day may god be with you (kissing emoticon) see you

⬇

Segmentation into words and annotation

⬇

ouii	NLTD
ritha	LTD
chérie	NLTD
!	PUNCT
b	NLTD
journée	NLTD
Rabi	LTD
m3ak	LTD
:*	EMO
++	SYMB

Figure 2. Corpus annotation process.

The corpus is subsequently divided into train (80% of the corpus) and test (20% of the corpus) sets randomly. Table 1 shows the details.

Table 1. Statistics on the corpus.

Corpus	#Msg	#Words					
		LTD	NLTD	PUNCT	SYMB	EMO	total
Global	13,656	76,416	79,646	8402	629	2244	167,337
Train	10,882	60,611	65,787	6358	510	1792	135,058
Test	2774	15,805	13,859	2044	119	452	32,279

4.2. Multilingualism and Code Switching

Before proceeding to the identification task, we propose to assess the multilingualism and code-switching behaviour in our corpus. Therefore, we compute 2 metrics:

- Multilingual Index (M-Index): computes the inequality of the language tags distribution in a corpus of at least two languages [7]. It is calculated as follows:

$$M - Index = \frac{1-\sum p_j^2}{(k-1)\sum p_j^2} \tag{1}$$

p_j is the total number of words in the language j over the total number of words in the corpus and k represents the total number of languages in the corpus. The value of M-Index can range between 0 and 1. A value of 1 means that the corpus has equal number of words from each language where a value of 0 means that the corpus includes only one language.

- Integration Index (I-Index): computes the approximate probability that any given token in the corpus is a switch point [24]. The integration index is calculated as follows:

$$I - Index = \frac{1}{n-1} \sum_{1 \leq i = j-1 \leq n-1} S(l_i l_j) \quad (2)$$

Where l_i represents the words tagged by language and i ranges from 1 to the corpus size n-1. $S(l_i l_j)$ can be either 0 or 1. A value of 1 means that $l_i \neq l_j$ and a value of 0 indicates that $l_i = l_j$.

For our experiments, we consider only 2 languages: LTD if the word belongs to the Tunisian dialect or NLTD if the word belongs to a foreign language. Table 2 shows the results of the code-switching metrics.

Table 2. Code-switchingmetrics.

Language pairs	M-Index	I-Index
LTD - NLTD	0.98	0.16

Based on the results shown in Table 2, we note that our corpus is indeed multilingual (M-index > 0). The amount of NLTD is approximately equal to that of LTD since the M-Index is close to 1. On the other hand, the probability of switching between LTD and NLTD in our corpus is equal to 0.16.

4.3. Language Ambiguity

Besides the code-switching phenomenon that characterizes the LTD texts written on the social web, there is another aspect of this language that is worth mentioning since it further complicates the LI task, that is the language ambiguity. This phenomenon rises when 2 or more words are written in the same way but bear different meanings and belong to different languages. Indeed, as we mentioned in the previous sections of this paper, practitioners of the LTD language on the social web don’t follow any guidelines in their writings namely syntactic rules or orthographic conventions. This evenly applies to the LTD and the foreign languages.

On one hand, LTD is a language that is fundamentally oral and lacks consequently spelling rules for its written form. On the other hand, foreign languages, like French and English, can be written by Tunisians either correctly following their adequate spelling rules or in an informal form which we chose to refer to as “SMS form”. This denomination amounts to the fact that this language appeared with the Short Message Service in mobile phones (texts) in the year 2000 [21]. With SMS texts, users began to personalize their writings by omitting vowels, replacing characters by digits, using acronyms, etc.,We give examples of this phenomenon in Table 3.

Table 3. Examples of words in SMS form.

Language	Word	SMS form
French	Demain (tomorrow)	2m1
English	See you	CU
French	Quand (when)	Kan

The use of the SMS form of foreign languages by Tunisians enhances the ambiguity cases. Examples are shown in Table 4.

Table 4. Examples of language ambiguity in LTD.

Word	Languages	Meanings
Ki	LTD	When
	French-SMS	Who (correct transcription: “qui”)
Bled	LTD	Country
	English	Lost blood
machin	LTD	We’re going
	French	Thingy (colloquial)

The examples given in Table 4 prove that we are not only dealing with simple code switching in the LI task, but also with a high language ambiguity rate. We tried to measure this ambiguity in our corpus using 5 lexicons, presented by Younes and Souissi [59]. The lexicons correspond to LTD, French, French-SMS, English and English-SMS languages, which are the most used languages by Tunisians based on our observations of the writings on the social web. Table 5 shows the details of the used lexicons.

Table 5. Used lexicons.

Lexicon		#Entries
LTD (Enriched)		27,712
NLTD	French	336,531
	French-SMS	770
	English	354,986
	English-SMS	950

We performed the ambiguity counting assuming that any word belonging to LTD and any other language at the same time is considered ambiguous. Therefore, we used our annotated corpus and the lexicons presented in Table 5 and computed the ambiguity rate considering the following cases:

- Number of words that are annotated as LTD and belong at the same time to one of the NLTD lexicons.
- Number of words that are annotated as NLTD and belong at the same time to the LTD lexicon.

We note, however, that the used lexicons do not entirely cover our corpus’s vocabulary. The results shown in Table 6 give us an approximate idea about the language ambiguity in our corpus.

Table 6. Ambiguity rates.

	Words with repetition	Words without repetition
#Words	167,337	36,493
#covered words	148,593	32,276
#coverage rate	88.80%	88.44%
#ambiguous words	30,585	1346
Ambiguity rate	20.58%	4.17%

As shown in Table 6, 20.58% of the covered words in the corpus are ambiguous when we counted the repeated words. On the other hand, 4.17% of the covered vocabulary words (without repetition) are ambiguous.

5. Proposed model

Our corpus is composed of messages. To each word of these messages is assigned a well-defined tag. Accordingly, we chose to model the LTD identification as a sequence labelling task which consists in assigning a tag to each word indicating whether it belongs to LTD or not. We thus propose an approach based on deep learning, namely a BLSTM-CRF model whose components will be described in the following section.

5.1. Model Components

- *Word embeddings*: Throughout our experiments, we aimed to explore a word vector representation that catches the morphological structure of the words. This kind of information is not taken into account by traditional word embeddings, like Wor2Vec[3] where the words are processed like atomic entities. Therefore, we resorted to FastText[4], an open source library for text representations that was developed by Facebook [30]. We used the skip-gram model which predicts the context given the word. We chose this model since it is known to perform well with small amounts of training data and rare words [38]. Therefore, given an LTD word (w$_i$), the skip-gram model predicts the surrounding context words. Regarding the n-gram features, the parameters of the model have been adjusted to take into account character n-grams from 2 to 6. The structure of our word vector representation model is illustrated in Figure 3

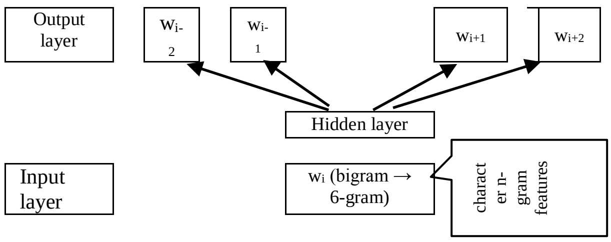


Figure 3. Fast Text word embeddings using a skip-gram model.

- *BLSTM-CRF*: Bidirectional long short-term memory was first introduced by Hochreiter and Schmidhuber [27]. It is a variant of RNN, capable of capturing long-range dependencies on sequential data. LSTM unit, is basically composed of 4 layers, interacting and controlling the information to forget or to pass on to the next time step [37]. Figure 4 illustrates the interaction between the layers in an LSTM unit.

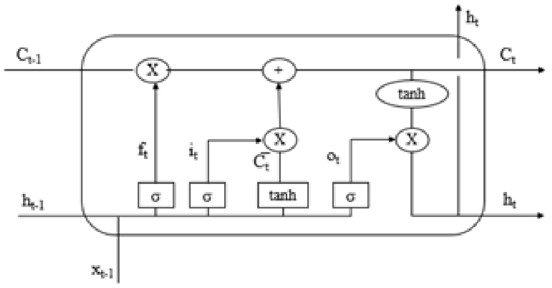


Figure 4. LSTM unit.

The layers shown in Figure 4 can be summarized as follows, where f_t, i_t, o_t and C_t represent the forget, input and output gates respectively and δ is the sigmoid function [22]:

$$f_t = \delta(W_f[h_{t-1}, x_t] + b_f) \quad (3)$$

$$\widetilde{C_t} = tanh(W_c[h_{t-1}, x_t] + b_c) \quad (4)$$

$$i_t = \delta(W_i[h_{t-1}, x_t] + b_i) \quad (5)$$

$$C_t = f_t C_{t-1} + i_t \tilde{C}_t \quad (6)$$

$$o_t = \delta(W_o[h_{t-1} x_t] + b_o) \quad (7)$$

$$h_t = o_t\, tanh(C_t) \quad (8)$$

LSTM's hidden state captures only past information. It is however more constructive for the LTD identification task to benefit from both past and future contexts. This issue is solved by presenting each sequence forwards and backwards to two separate hidden states, which is the basic idea for the bidirectional LSTM [16].

As each word is associated to a vector capturing information about its context in the message, a vector of scores is computed in order to make the final prediction. In the decoding step, we can normalize the scores into a probability that the LTD word is categorized as belonging to a well-defined language. The tagging decision in this method is local and based solely on the context of the word in the message but doesn't take into account the adjacent tagging decisions. Therefore, we propose to add a CRF output layer to our model, in order to consider the correlations between tags and choose the best sequence of labels for a given input LTD message [37].

- *Word-char embeddings*: For each word, we aim to generate a vector containing features extracted from the character level. Therefore, we resort to a character level BLSTM [32]. The model is run on the character embeddings and the final states are subsequently concatenated to generate the word vector [32]. Figure 5 illustrates the word-char vector representation for the LTD word "dar" (meaning: house).

[3] https://code.google.com/archive/p/word2vec/

[4] https://fasttext.cc/

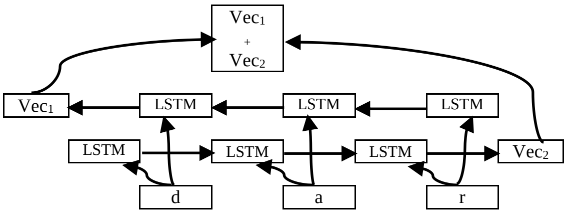


Figure 5. Word-char embeddings model.

The resulting vector captures the morphology of the word since it took into account each character separately.

5.2. LTD Identification Model

The input of our identification model is twofold word embeddings:

1. Word embeddings with Fasttext: captures character n-gram features from 2 to 6.
2. Word-char embeddings with BLSTM: captures character embeddings.

These representations are subsequently concatenated to get a vector representing each word considering its context within the whole LTD message. The BLSTM model learns, thus, the fixed dimensional representation from the embedding layers. Finally, we add a CRF layer at the output to take into account the neighbouring tagging decisions. Figure 6 summarizes the components of our LTD identification model.

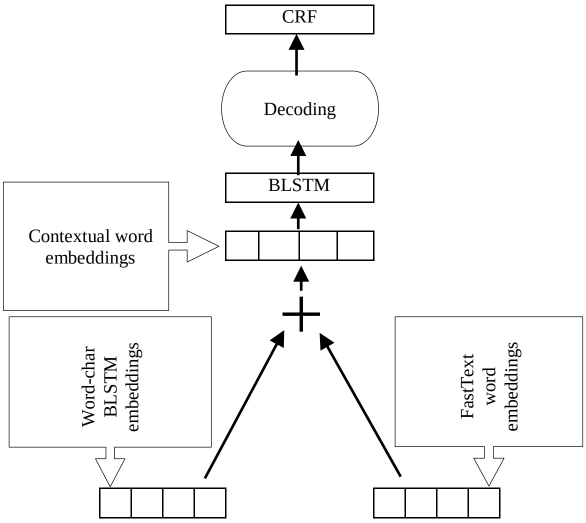


Figure 6. Proposed LI model.

As shown in Figure 6, the model's input takes each word's fasttext and word-char vector representations then concatenates them into a contextual word embedding vector that considers the word's morphology. This resulting vector is learnt by the BLSTM model and the decoding layer is provided by CRF to ensure the consideration of the neighbouring tagging decisions.

6. Experiments and Results

The hyperparameters of our model were tuned as follows[5]:

- Number of hidden layers: 400
- Number of epochs: 40
- Batch size: 20
- Dropout rate: 0.5
- Decay rate: 0.9

We performed several experiments using other combinations of the proposed model which we detail in Table 7.

Table 7. Tested models and their components.

Model	**Components**			
	Word2Vec	FastText	Char-word	CRF layer
BLSTM-wv	√			
BLSTM-cwv	√		√	
BLSTM-ft		√		
BLSTM-cft		√	√	
BLSTM-CRF-wv	√			√
BLSTM-CRF-cwv	√		√	√
BLSTM-CRF-ft		√		√
BLSTM-CRF-cft		√	√	√

We experimented using a CRF model as well to which we introduced 5 word-level features namely: size, suffixes and prefixes, word containing digits, word containing consecutive 3 consonants and Word2Vec embeddings. The results in terms of overall accuracy are presented in Table 8.

Table 8. Overall identification results.

Model	**Overall accuracy**
CRF	95.44%
BLSTM-wv	97.56%
BLSTM-cwv	98.56%
BLSTM-ft	98.24%
BLSTM-cft	98.60%
BLSTM-CRF-wv	97.59%
BLSTM-CRF-cwv	98.49%
BLSTM-CRF-ft	98.25%
BLSTM-CRF-cft	**98.65%**

Based on the results shown in Table 8, we notice that the best accuracy is obtained when we use the combination of BLSTM-CRF with FastText and word-character embeddings. We detail the results given by this model in terms of precision, recall and F1-score for each tag in Table 9.

Table 9. Detailed results of the proposed model.

Accuracy	**Tags**	**Precision**	**Recall**	**F1-score**
98.65%	LTD	98.77%	98.55%	98.66%
	NLTD	98.32%	98.59%	98.45%
	EMO	99.56%	99.56%	99.56%
	PUNCT	100%	100%	100%
	SYMB	96.33%	94.71%	95.51%

[5] We note that we used the NumPy and Tensorflow libraries.

The LI identification model we proposed for the LTD language yielded encouraging results on a test set of 2774 messages including 32,279 words. We reached an overall accuracy of 98.65%.

We believe that the good performance of the model is due to the consideration of the words' morphological structures as well as their contexts within the messages in which they appear.

Comparing our work to that of Aridhi *et al*. [3], in which the same corpus was used, we notice that the approach we followed outperforms that of Aridhi's *et al*. [3]. The authors in [3] focused on the word-level identification of Romanized TD and took into account the morphological structure of the words by adopting n-gram cumulative frequency addition [1] and SVM. However, the context of the messages to which the words belong was not considered. The best results were obtained with a 3-gram SVM classifier using a combination of 3 features: N-gram frequency, N-gram location and digit presence in N-gram. The accuracy of the model reached 93.57% [3]. In our work, we tried to take advantage of the word contexts and the neighboring tagging decisions and we combined 2 vector representations of the words. The captured context allowed us to reach a good overall performance of 98.65%.

The error rate is low when it comes to the emoticons, symbols and punctuation as the ambiguity between them is practically unremarkable. However, cases of errors occurred in the LTD/NLTD pairs. Based on the F1-score of the NLTD tag, we notice that most of the errors result of tagging an LTD word as a foreign word. We summarize and categorize the observed errors as follows:

1. LTD words with foreign roots: we notice several cases of erroneous identification with the words having non-LTD roots and LTD affixes. We show some examples in Table 10.

Table 10. Examples of LTD words with foreign roots.

LTD word	Meaning	Root	Root language	LTD affixes
Encadreurek	Your supervisor	encadreur	French	Ek
Partagili	Share to me	partag	French	Ili
Maconnectitech	I haven't connected	connect	French / English	ma / itech

2. LTD words with foreign origins: this case is different from the previous one since the foreign word's morphology is fully altered to match a Tunisian dialect conjugation (verb, plural, etc.,) Table 11 shows some examples.

Table 11. Examples of LTD words with foreign origins.

LTD word	meaning	source word	source language
Nrivzou	We revise	réviser / revise	French / English
Fransis	French people	Français	French
Mvaryes	That caught a virus	Virus	French / English
Triguel	She adjusts	Régler	French

3. LTD conjunctions attached to foreign words: this kind of errors occur when an LTD coordination conjunction or a determiner is attached to a non-LTD word. Some examples are given in Table 12.

Table 12. Examples of foreign words with LTD conjunctions.

word	Meaning	LTD conjunction	NLTD word	NLTD language
elfauteuil	the armchair	el (the)	fauteuil (armchair)	French
estade	the stadium	e (the)	stade (stadium)	French
lgoogle	for google	l (for / to)	google (search engine)	English
wpartie	and the section	w (and)	partie (section)	French

4. Oddly written LTD words: these words are written in LTD but with a particular orthography, rarely seen in everyday communication. They are incorrectly considered by the LI model as non-LTD words. We distinguish among them:
   - Words written with English language phonemes: most Tunisians follow French language conventions when they write in the Latin script. For example, they refer to the character succession "ou" to transcribe the long vowel "و" in Arabic. The succession "ou" is mostly used in the French language as the phoneme /u: /. This same phoneme is referred to, in English for example, as the succession "oo". Therefore, using English phonemes in the corpus resulted in erroneous tags.
   - Words written using French orthographic particularities: some used characters led to error cases are known to be used mostly in the French language such as the succession "gu" to designate the phoneme /g/ or the character "ç" that is particularly used, among others, in the French language.
   - Use of accented letters: other cases of errors were observed when the LTD word includes accented characters. It's unlikely that users resort to this level of precision in their written exchanges in social media as they usually opt for the simplest and the quickest transcription.
   - Use of the single quote ('): rare cases of errors were noticed when some users resort to the single quote character in the middle of the LTD word to refer to the Arabic diacritization sign "Soukoun" (ـْ), when it's more likely to find it in other languages such as French or English.

Table 13 shows examples of oddly written LTD words.

Table13. Examples of oddly written LTD words.

LTD word	Meaning
Sout**'**ha	her voice
Mon**gu**ela	a watch
Matl**oo**b	Demanded
Sm**é**	Sky

Based on the error analysis, we notice that part of the observed inaccuracies is related to the rarity of words. We can remedy this issue by enlarging the size of the corpus to cover a maximum of vocabulary. A larger corpus will certainly incorporate different kinds and styles of written LTD texts and consequently help reduce the errors related to oddly written words.

Regarding the errors related to conjunctions and determiners that are occasionally attached to non-LTD words, they amount to our segmentation method which relied on the space character. Such errors might be avoided if we adopt another segmentation method where we take morphological information into consideration such as affixes and stop words.

7. Conclusions

We presented in this paper an approach for the language identification of Tunisian dialect written in the social web using the Latin script. We resorted to a deep learning method based on Bidirectional Long Short-Term Memory with CRF decoding (BLSTM-CRF) using word-char BLSTM embeddings combined with FastText embeddings. The latter takes into consideration character n-gram features and captures each word's morphology.

We used a corpus composed of messages from Tunisian social web pages. The corpus was subsequently segmented and annotated according to 5 categories mainly Latin Tunisian dialect, foreign words, punctuation, emoticons and symbols. We performed a counting on the corpus to assess the code-switching behaviour and get the measure of the multilingualism. We found a multilingual index of 0.98 and an integration index of 0.16.

The identification results were quite good with an overall accuracy of 98.65%. Cases of observed errors were related to the rarity of the words, others were due to the adopted segmentation method. The inaccuracy cases can be reduced by enlarging the corpus size to include a maximum of vocabulary and by improving the segmentation process.

We aim, in future studies, to further explore the identification approach and use it to generate sizable LTD corpora rich in dialectal content. We intend to create lexica and dictionaries, for this form of TD as well, that will allow researchers to lead exhaustive studies and develop adequate NLP tools.

Moreover, we plan to focus in future work on the Tunisian dialect written in the Arabic script. This form of writing presents a significant challenge giving the high language ambiguity between Arabic TD and MSA.

References

[1] Ahmed B., Cha S., and Tappert C., "Language Identification from Text Using N-gram Based Cumulative Frequency Addition," *in Proceedings of Student/Faculty Research Day*, USA, pp. 1-8, 2004.

[2] Al-Badrashiny M. and Diab M., "LILI: A Simple Language Independent Approach for Language Identification," *in Proceedings of COLING 26th International Conference on Computational Linguistics: Technical Papers*, Osaka, pp. 1211-1219, 2016.

[3] Aridhi C., Achour H., Souissi E., and Younes J., "Word-Level Identification of Romanized Tunisian Dialect," *in Proceedings of International Conference on Applications of Natural Language to Information Systems*, Liège, pp. 170-175, 2017.

[4] Bar K. and Dershowitz N., "The Tel Aviv University System for the Code-Switching Workshop Shared Task," *in Proceedings of 1st Workshop on Computational Approaches to Code Switching*, Doha, pp. 139-143, 2014.

[5] Barman U., Das A., Wagner J., and Foster J., "Code Mixing: A Challenge for Language Identification in the Language of Social Media," *in Proceedings of 1st Workshop on Computational Approaches to Code Switching*, Doha, pp. 13-23, 2014.

[6] Barman U., Wagner J., Chrupala G., and Foster J., "DCU-UVT: Word-Level Language Classification with Code-Mixed Data," *in Proceedings of 1st Workshop on Computational Approaches to Code Switching*, Doha, pp. 127-132, 2014.

[7] Barnett R., Codò E., Eppler E., Forcadell M., Gardner-Chloros P., Hout R., Moyer M., Torras M., Turell M., and Sebba M., "The LIDES Coding Manual: A document for Preparing and Analyzing Language Interaction Data Version," *International Journal of Bilingualism*, vol. 4, no. 2, pp. 131-271, 2000.

[8] Bartz C., Herold T., Yang H., and Meinel C., "Language Identification Using Deep Convolutional Recurrent Neural Networks," *in Proceedings of International Conference on Neural Information Processing*, Guangzhou, pp. 880-889, 2017.

[9] Bouamor H., Habash N., and Oflazer K., "A Multidialectal Parallel Corpus of Arabic," *in Proceedings of 9th International Conference on Language Resources and Evaluation*, Reykjavik, pp. 1240-1245, 2014.

[10] Chanda A., Das D., and Mazumdar C., "Columbia-Jadavpur submission for EMNLP 2016 Code-Switching Workshop Shared Task: System Description," *in Proceedings of the 2nd Workshop on Computational Approaches to Code Switching*, Austin, pp.112-115, 2016.

[11] Chanda A., Das D., and Mazumdar C., "Unraveling the English-Bengali Code Mixing

Phenomenon," *in Proceedings of 2nd Workshop on Computational Approaches to Code Switching*, Austin, pp. 80-89, 2016.

[12] Chang J. and Lin C., "Recurrent-neural-network for Language Detection on Twitter Code-Switching Corpus," *arXiv preprint, arXiv:1412.4314*, pp. 1-9, 2014.

[13] Chittaranjan G., Vyas Y., Bali K., and Choudhury M., "Word-level Language Identification using CRF: Code-switching Shared Task Report of MSR India System," *in Proceedings of 1st Workshop on Computational Approaches to Code Switching*, Doha, pp. 73-79, 2014.

[14] Daoud M., "The Language Situation in Tunisia," *Current Issues in Language Planning*, vol. 2, no. 1, pp. 1-52, 2001.

[15] Dongen N., Analysis and Prediction of Dutch-English Code-switching in Dutch Social Media Messages, Master's Thesis, Universiteit van Amsterdam, 2017.

[16] Dyer C., Ballesteros M., Ling W., Matthews A., and Smith N., "Transition-Based Dependency Parsing with Stack Long Short-Term Memory," *in Proceedings of 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, Beijing, pp. 334-343, 2015.

[17] Elaraby M. and Abdul-Mageed M., "Deep Models for Arabic Dialect Identification on Benchmarked Data," *in Proceedings of 5th Workshop on NLP for Similar Languages, Varieties and Dialects*, Santa Fe, pp. 263-274, 2018.

[18] Elfardy H., Al-Badrashiny M., and Diab M., "AIDA: Identifying Code Switching in Informal Arabic Text," *in Proceedings of 1st Workshop on Computational Approaches to Code Switching*, Doha, pp. 94-101, 2014.

[19] Eskander R., Al-Badrashiny M., Habash N., and Rambow O., "Foreign Words and the Automatic Processing of Arabic Social Media Text Written in Roman Script," *in Proceedings of 1st Workshop on Computational Approaches to Code Switching*, Doha, pp. 1-12, 2014.

[20] Giwa O. and Davel M., "N-Gram based Language Identification of Individual Words," *in Proceedings of Conference: Pattern Recognition Association of South Africa*, Johannesburg, pp. 1-22, 2013.

[21] Goumi A., Volckaert-Legrier O., Bert-Erboul A., and Bernicot J., "SMS Length and Function: A Comparative Study of 13-to 18-Year-Old Girls and Boys," *European Review of Applied Psychology*, vol. 61, no. 4, pp. 175-184, 2011.

[22] Graves A., Mohamed A., and Hinton G., "Speech Recognition with Deep Recurrent Neural Networks," *in Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vancouver, pp. 6645-6649, 2013.

[23] Guellil I. and Azouaou F., "Arabic Dialect Identification with an Unsupervised Learning (based on a lexicon) Application case: ALGERIAN Dialect," *in Proceedings of IEEE International Conference on Computational Science and Engineering, IEEE International Conference on Embedded and Ubiquitous Computing, and International Symposium on Distributed Computing and Applications to Business, Engineering and Science*, AnYang, pp. 724-731, 2016.

[24] Guzman G., Serigos J., Bullock B., and Toribio A., "Simple Tools for Exploring Variation in Codeswitching for Linguists," *in Proceedings of 2nd Workshop on Computational Approaches to Code Switching*, Austin, pp. 12-20, 2016.

[25] Hassine M., Boussaid L., and Hassani M., "Tunisian Dialect Recognition Based on Hybrid Techniques," *The International Arab Journal of Information Technology*, vol. 15, no. 1, pp. 58-65, 2018.

[26] Heafield K., "KenLM: Faster and Smaller Language Model Queries," *in Proceedings of 6th Workshop on Statistical Machine Translation*, Edinburgh, pp. 187-197, 2011.

[27] Hochreiter S. and Schmidhuber J., "Long Short-Term Memory," *Neural Computation Archive*, vol. 9, no. 8, pp.1735-1780, 1997.

[28] Jaech A., Mulcaire G., Hathi S., Ostendorf M., and Smith N., "A Neural Model for Language Identification in Code-Switched Tweets," *in Proceedings of 2nd Workshop on Computational Approaches to Code Switching*, Austin, pp. 60-64, 2016.

[29] Jhamtani H., Kumar B., and Raychoudhury V., "Word-level Language Identification in Bi-lingual Code-switched Texts," *in Proceedings of 28th Pacific Asia Conference on Language, Information and Computation*, Phuket, pp. 348-357, 2014.

[30] Joulin A., Grave E., Bojanowski P., Douze M., Jégou H., and Mikolov T., "FastText.zip: Compressing Text Classification Models," *CoRR, abs/1612.03651*, 2016.

[31] King L., Baucom E., Gilmanov T., Kübler S., Whyatt D., Maier W., and Rodrigues P., "The IUCL+ System: Word-Level Language Identification via Extended Markov Models," *in Proceedings of 1st Workshop on Computational Approaches to Code Switching*, Doha, pp. 102-106, 2014.

[32] Lample G., Ballesteros M., Subramanian S., Kawakami K., and Dyer C., "Neural Architectures for Named Entity Recognition," *in Proceedings of the Conference of the North*

American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego, pp. 260-270, 2016.

[33] Lichouri M., Abbasa M., Freihatc A., and Megtoufa D., "Word-Level vs Sentence-Level Language Identification: Application to Algerian and Arabic Dialects," *Procedia Computer Science*, vol. 142, pp. 246-253, 2018.

[34] Lin C., Ammar W., Levin L., and Dyer C., "The CMU Submission for the Shared Task on Language Identification in Code-Switched Data," *in Proceedings of 1st Workshop on Computational Approaches to Code Switching*, Doha, pp. 80-86, 2014.

[35] Mager M., Çetinoğlu Ö., and Kann K., "Subword-Level Language Identification for Intra-Word Code-Switching," *Ground AI*, vol. 1, 2019.

[36] Mave D., Maharjan S., and Solorio T., "Language Identification and Analysis of Code-Switched Social Media Text," *in Proceedings of 3rd Workshop on Computational Approaches to Code-Switching*, Melbourne, pp. 51-61, 2018.

[37] Ma X. and Hovy E., "End-To-End Sequence Labeling Via Bi-Directional LSTM-CNNs-CRF," *in Proceedings of 54th Annual Meeting of the Association for Computational Linguistics*, Berlin, pp. 1064-1074, 2016.

[38] Mikolov T., Sutskever I., Chen K., Corrado G., and Dean J., "Distributed Representations of Words and Phrases and their Compositionality," *in Proceedings of 26th International Conference on Neural Information Processing Systems 2*, Lake Tahoe, pp. 3111-3119, 2013.

[39] Molina G., Rey-Villamizar N., Solorio T., AlGhamdi F., Ghoneim M., Hawwari A., and Diab M., "Overview for the Second Shared Task on Language Identification in Code-Switched Data," *in Proceedings of 2nd Workshop on Computational Approaches to Code Switching*, Austin, pp. 40-49, 2016.

[40] Nguyen D. and Cornips L., "Automatic Detection of Intra-Word Code-Switching," *in Proceedings of 14th Annual SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, Berlin, pp. 82-86, 2016.

[41] Papalexakis E., Nguyen D., and Dogruöz A., "Predicting Code-Switching in Multilingual Communication for Immigrant Communities," *in Proceedings of 1st Workshop on Computational Approaches to Code Switching*, Doha, pp. 42-50, 2014.

[42] Shirvani R., Piergallini M., Gautam G., and Chouikha M., "The Howard University System Submission for the Shared Task in Language Identification in Spanish-English Codeswitching," *in Proceedings of 2nd Workshop on Computational Approaches to Code Switching*, Austin, pp. 116-120, 2016.

[43] Piergallini M., Shirvani R., Gautam G., and Chouikha M., "Word-Level Language Identification and Predicting Codeswitching Points in Swahili-English Language Data," *in Proceedings of 2nd Workshop on Computational Approaches to Code Switching*, Austin, pp. 21-29, 2016.

[44] Revay S. and Teschke M., "Multiclass Language Identification using Deep Learning on Spectral Images of Audio Signals," *arXiv:1905.04348v1*, 2019.

[45] Rijhwani S., Sequeira R., Choudhury M., Bali K., and Maddila C., "Estimating Code-Switching on Twitter with a Novel Generalized Word-Level Language Detection Technique," *in Proceedings of 55th Annual Meeting of the Association for Computational Linguistics*, Vancouver, pp. 1971-1982, 2017.

[46] Sadat F., Kazemi F., and Farzindar A., "Automatic Identification of Arabic Language Varieties and Dialects in Social Media," *in Proceedings of 2nd Workshop on Natural Language Processing for Social Media*, Dublin, pp. 22-27, 2014.

[47] Salameh M., Bouamor H., and Habash N., "Fine-Grained Arabic Dialect Identification," *in Proceedings of 27th International Conference Computational Linguistics*, Santa Fe, pp. 1332-1344, 2018.

[48] Samih Y. and Maier W., "Detecting Code-Switching in Moroccan Arabic Social Media," *in Proceedings of 4th International Workshop on Natural Language Processing for Social Media SocialNLP*, New York, 2016.

[49] Samih Y., Maharjan S., Attia M., Kallmeyer L., and Solorio T., "Multilingual Codeswitching Identification via LSTM Recurrent Neural Networks," *in Proceedings of 2nd Workshop on Computational Approaches to Code Switching*, Austin, pp. 50-59, 2016.

[50] Sayadi K., Hamidi M., Bui M., Liwicki M., and Fischer A., "Character-Level Dialect Identification in Arabic Using Long Short-Term Memory," *in Proceedings of International Conference on Computational Linguistics and Intelligent Text Processing*, Budapest, pp. 324-337, 2017.

[51] Schulz S. and Keller M., "Code-Switching Ubique Est - Language Identification and Part-of-Speech Tagging for Historical Mixed Text," *in Proceedings of 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, Berlin, pp. 43-51, 2016.

[52] Shrestha P., "Codeswitching Detection via Lexical Features using Conditional Random

Fields," *in Proceedings of 2nd Workshop on Computational Approaches to Code Switching*, Austin, pp. 121-126, 2016.

[53] Shrestha P., "Incremental N-gram Approach for Language Identification in Code-Switched Text," *in Proceedings of 1st Workshop on Computational Approaches to Code Switching*, Doha, pp. 13-138, 2014.

[54] Sikdar U. and Gambäck B., "Language Identification in Code-Switched Text Using Conditional Random Fields and Babelnet," *in Proceedings of 2nd Workshop on Computational Approaches to Code Switching*, Austin, pp. 127-131, 2016.

[55] Solorio T., Blair E., Maharjan S., Bethard S., Diab M., Gohneim M., Hawwari A., AlGhamdi F., Hirschberg J., Chang A., and Fung P., "Overview for the First Shared Task on Language Identification in Code-Switched Data," *in Proceedings of 1st Workshop on Computational Approaches to Code Switching*, Doha, pp. 62-72, 2014.

[56] Xia M., "Codeswitching Language Identification Using Subword Information Enriched Word Vectors," *in Proceedings of 2nd Workshop on Computational Approaches to Code Switching*, Austin, pp. 132-136, 2016.

[57] Yankova D. and Vassileva I., "Functions and Mechanisms of Code-Switching," *Bulgarian Canadians, Étudescanadiennes/Canadian Studies*, vol. 74, pp. 103-121, 2013.

[58] Younes J., Achour H., and Souissi E., "Constructing Linguistic Resources for the Tunisian Dialect Using Textual User-Generated Contents on The Social Web," *in Proceedings of International Conference on Web Engineering*, Rotterdam, pp. 3-14, 2015.

[59] Younes J. and Souissi E., "A Quantitative View of Tunisian Dialect Electronic Writing," *in Proceedings of 5th International Conference on Arabic Language Processing CITALA*, Oujda, pp. 63-72, 2014.

[60] Younes J., Souissi E., Achour H., and Ferchichi A., "Un Etat De L'art Du Traitement Automatique Du Dialecte Tunisien," *Traitement Automatique des Langues*, vol. 59, no. 3, pp. 93-117, 2018.

[61] Zaidan O. and Callison-Burch C., "Arabic Dialect Identification," *Computational Linguistics*, vol. 40, no. 1, pp. 171-202, 2014.

Jihene Younes is PhD student at the ISGT, University of Tunis, Tunisia. She received her Master's in Computer Science from the ENSIT, University of Tunis, Tunisia. Her current research interests include the automatic processing of the Tunisian dialect.

Hadhemi Achour is Assistant Professor, teaching Computer Science at the ISGT, University of Tunis, Tunisia. She received her PhD in Computer Science at the University of Paris 7 in France. Her doctoral research was conducted at the France's National Scientific Research Centre (CNRS). Her main research interests are related to Text Mining, Natural Language Processing and their applications, including Arabic and Tunisian dialect language processing. She participated in several European projects and in ALECSO coordinated studies and research projects.

Emna Souissi is Assistant Professor and teaching Computer Science at the ENSIT, University of Tunis, Tunisia. She holds a PhD in Computer Science from the University of Paris 7, France. Her research interests are mainly related to the field of natural language processing and its applications, with a focus on the Arabic NLP. Her PhD research was conducted within the CNRS. In this context, she has participated in several European and Canadian projects. She is currently conducting research on the treatment of Arabic dialects and mainly Tunisian.

Ahmed Ferchichi has been a professor of computer science since 1980. He is a PhD in computer science from Joseph-Fourrier University of Grenoble. His research interests include teaching programming and software engineering, modeling training curricula and educational systems, achieving sustainable development goals by the use of information technology and artificial intelligence, promoting information technology culture in Arabic. He taught at the University of Tunis from 1980 to 2011, where he directed the academic affairs of the ISGT during the period 2000-2003. Since 2012, he teaches at the University of Jendouba. In 2018, he was member of the national commission for the supervision of computer science study programs.