International Journal of Electrical and Computer Engineering (IJECE)
Vol. 11, No. 2, April 2021, pp. 1627~1633
ISSN: 2088-8708, DOI: 10.11591/ijece.v11i2.pp1627-1633

Arabic tweeps dialect prediction based on machine learning approach

Khaled Alrifai, Ghaida Rebdawi, Nada Ghneim
Department of Informatics, Higher Institute for Applied Sciences and Technology, Syria

Article Info

Article history:

Received Nov 22, 2019
Revised Aug 14, 2020
Accepted Nov 4, 2020

Keywords:

Arabic dialects detection
Author profiling
Machine learning
Social media analysis
Text mining

ABSTRACT

In this paper, we present our approach for profiling Arabic authors on Twitter, based on their tweets. We consider here the dialect of an Arabic author as an important trait to be predicted. For this purpose, many indicators, feature vectors and machine learning-based classifiers were implemented. The results of these classifiers were compared to find out the best dialect prediction model. The best dialect prediction model was obtained using random forest classifier with full forms and their stems as feature vector.

Corresponding Author:

Khaled Alrifai
Department of Informatics
Higher Institute for Applied Sciences and Technology
Barzeh, Damascus, Syria
Email: khaled.alrifai@hiast.edu.sy

1. INTRODUCTION

Author profiling on social media is a method of analysing the author writings on social media in order to uncover different traits of the author (e.g. gender and age) based on stylistic or content-based features. This method aims at taking advantage of a huge volume of data generated by a huge number of authors, in order to classify them into predefined classes based on their traits. Author profiling has many useful applications in the domain of social media analysis, such as in marketing and advertising, as well as in the forensic and security areas [1].

With the birth and rise of social media, internet users in the Arab world were quick to embrace the new technology, and utilize all what social media has to offer to connect, communicate, and share information with others using Arabic language [2]. Arabic language that used in social media has two forms: the first, is the modern standard Arabic (MSA), which is widely used in formal situations like formal speeches, government and official contents; the second, is known as dialectal Arabic (DA) which is the informal private language, predominantly found as spoken vernaculars with no written standards. Dialects differ in morphologies, grammatical cases, vocabularies and verb conjugations [3]. These differences call for dialect-specific processing and modeling when building Arabic automatic analysis systems [4].

The natural language processing (NLP) community has aggregated dialectal Arabic into four regional language groups: Egyptian, Maghreban, Gulf, and Levantine dialects, in addition to modern standard Arabic (MSA), the Arabic formal language. An objective comparison of the varieties of Arabic dialects could potentially lead to the conclusion that Arabic dialects are historically related, but are mutually unintelligible dialects [5]. Author profiling is a classification problem that could be solved using various approaches. These approaches are based on the selected features extracted from the author's writings, and the classifier used in the development of the prediction model. A lot of researches in the field of author profiling previewed a comparison study between multiple features [6] and classifiers [7] to select the best combination of them for predicting a specific trait.

Features used in dialect identification problems are content-based and style-based features [8]. In content-based features, character *n*-grams and word *n*-grams are widely used. Kheng *et al.* in [9] used word *n*-grams with values between 1 and 3 for *n*. Markov *et al.* in [10] combined character *n*-grams and word *n*-grams with values of 3-4 for typed characters, 3-7 for untyped characters and 2-3 for words, respectively. In untyped characters, *n*-gram types are ignored (e.g. 'the' as a whole word is no different from 'the' in the middle of a word), but in typed characters, *n*-grams of different types are distinguished (e.g. *n*-grams may be suffixes, punctuations, words etc.). Similarly, Ciobanu *et al.* in [11] combined character and word *n*-grams with values of *n* of 1-6 and 1-2 respectively. In [12, 13], tf-idf *n*-grams were combined with word embedding, and with 2-gram characters in the beginning and ending, respectively.

Many features selection criteria have been used: gain ratio [14], bag-of-words [15, 16], the 100 most discriminant words per class from a list of 500 topic words [17], latent semantic analysis LSA [9], and specific lists of words for dialect [18]. In style-based features, character flooding (i.e. lengthened words) and emoticons or/and laugher expressions [15, 18] were commonly used. Markov *et al.* in [10] also combined domain names that are used in links, with different kinds of *n*-grams. Arcial *et al.* in [15] combined emotional features such as: emotions, appraisal, admiration, positive/negative emoticons, and positive/ negative words. Martinc *et al.* in [18] also used emojis and sentiment words.

Concerning classification algorithms, most researchers used traditional machine learning algorithms such as logistic regression [12, 18, 19], SVMs [9-11, 16, 20-22], and distance-based methods [14, 15, 17]. Some researchers employed deep learning techniques for this purpose. For example, Kodiyan *et al.* in [23] applied recurrent neural networks (RNN), whereas Schaetti in [13] and Sierra *et al.* in [24] used convolutional neural networks (CNN). Finally, Salvador *et al.* in [25] applied deep averaging networks.

Dialect of Arabic tweeps (Twitter users) is the trait under study of author profiling in this paper. Accordingly, the required task is to develop a model that can predict the dialect of a tweep based on his/her Arabic tweets. In the rest of this paper, we present our methodology that includes: the characteristics of training and testing data, the features used for the developed model, and a step-by-step approach to build the prediction model in section 2. In section 3, a brief discussion of the results is addressed. At the end, insights for the future and a short summary are presented.

2. RESEARCH METHOD

In this section, we describe the dataset used in this work, and the features developed for the prediction model. The proposed model is explained in detail hereafter, including: data pre-processing, features extraction, features filtering and the algorithms with their evaluation criteria.

2.1. Dataset

In our research, we used training dataset from PAN conference 2017 [8]. One of PAN 2017 tasks was about Arabic tweeps profiling according to their dialects. This data consists of 240,000 Arabic tweets written by 2,400 authors (100 tweets for each author). Authors were tagged with their dialects. Dialects were divided into four classes: Levantine, Gulf, Egyptian and Maghreban. Authors were categorized into 4 classes of 600 authors each. As a testing dataset, PAN 2017 prepared also a dataset consists of 160,000 Arabic tweets written by 1600 authors (100 tweets for each author), divided equally into the four classes described in the training data.

2.2. Studied features

We implemented several experiments using different feature vectors. We categorized these features into:

a. Content-based features:
   - Uni-gram, bi-gram and tri-gram of words
   - Stems of words
   - Lemmas of words
   - Words part of speech tags (POS), i.e. NOUN_MS_PRON and V_PRON
   - Character *n*-gram, where *n* ranges from 2 to 7.

b. Style-based features:
   - Links to websites ("http")

- Hashtags to active public trends ("#")
- Mentions to other authors ("@")
- Lengthened words, i.e. the intentional repetition of a character in a word to emphasize and to exaggerate in describing something like laughing "ههههههههه", magnification "وااااااااااااااااو", indignation "لااااااااا", etc.
- Average tweets length, i.e. the average number of words in an author tweets
- Tweets punctuation marks, i.e. the summation of punctuation marks used in an author tweets.

2.3. Our model

In our attempt to find out the best prediction model, we prepared the dataset and extracted the features. These features have been filtered to reduce the size of feature vectors. Depending on reduced feature vectors, we implemented several experiments that differed from each other in feature vector and the algorithm used for training. The resulting models were compared using specific evaluation criteria to select the best one.

2.3.1. Data pre-processing

Before starting feature extraction stage, we concatenated all the 100 tweets for each author into one long text. This long text was tokenized using Farasa tokenizer [26]. All extracted tokens have been grouped and weighted with their frequency in the dataset (all the tokens from all authors).

2.3.2. Features extraction

After the tokenization step, lemmas and stems were extracted from the calculated tokens using Farasa toolbox. Tokens were used also to extract character 2-7 grams. In all content-based features, the calculated value for each feature was the frequency of use in the dataset. This step produced a huge size of feature vector that should be reduced.

Style-based features were also calculated and extracted for each author. We considered a word is lengthened if it included a character repeated three times at least. In case of average tweets length and tweets punctuation marks, the value of these features was the calculated count itself. In case of hashtags, mentions, links and lengthened words, the value of these features were the normalized usage ratio. These features were counted then normalized into the interval [0,100].

2.3.3. Features filtering

The number of elements of each content-based feature vector was very huge, which made the training process very hard and time-consuming. We applied the following steps to reduce the feature vector size:

- Eliminating features with a value less than five (we have four classes). The probability that these items contribute in the classification is low
- Discarding all elements with Information Gain IG equals to zero.

2.3.4. Model training

In our experiments, we trained different models using Weka toolbox. The features mentioned in 3.2. have been used separately or jointly to create various feature vectors to be tested in our experiments. According to the classifiers, we used initially support vector machine (SVM) in order to find out the best feature vector through several experiments. Using the resulting best feature vector, we trained other classifiers, such as: sequential minimal optimization (SMO), random forest (RF) and naïve bayes (NB) as training algorithms to be compared with SVM results.

2.3.5. Evaluation of models

For the evaluation, we used both training and testing dataset to find out the best model. In the training phase, we used F-measure (F1) over 10-folds cross-validation (F1Train), and in the testing phase, we calculated F1 over the testing dataset (F1Test).

3. RESULTS AND DISCUSSIONS

In this section, we present our experiments for dialect prediction. Initially, we used SVM with polynomial classifier to uncover the best feature vector, then we tried other classifiers to find out the best prediction model. The used features abbreviated here as: CNGram for character *n*-gram; UniGram, BiGram and TriGram for word uni-gram, bi-gram and tri-gram respectively, Stem for stems, Lemma for lemmas, POS for part of speech tags, LEN for lengthened words ratio, ATL for average tweets length, TPM for tweets punctuation marks and LHM for links, hashtags and mentions usability ratios.

3.1. Features vector comparison using SVM

As we mentioned above, several feature vectors have been used to train a number of models in multiple experiments. We used here SVM classifier with polynomial kernel as a training algorithm, and calculated the evaluation criteria for comparison. Figure 1 shows the results.

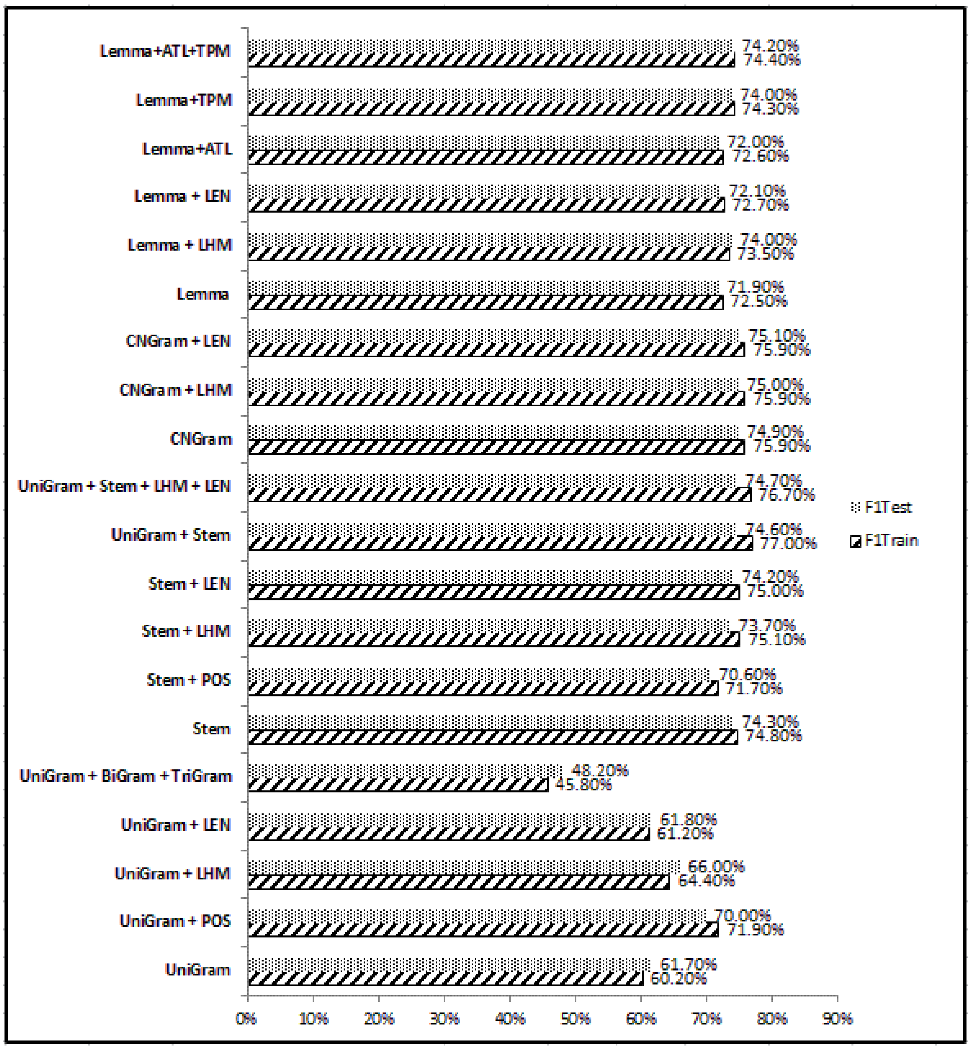

Figure 1. Comparison between feature vectors using polynomial-SVM

At the first step, we trained our model using the UniGram alone (F1Train = 60.2%), then when we added some features, we noticed that adding *POS* was the most effective (71.7%). Then, we tried the concatenation of UniGram, BiGram and TriGram of words, the result was relatively poor (45.8%). Using Stem instead of the UniGram increased the accuracy (74.8%). Moreover, using Stem concatenated with the UniGram, LEN and LHM ratios produced better accuracy (77%, 75% and 75.1% respectively). Using CNGram as main feature vector led to good accuracy (75.9%). Using Lemma with other features let to good accuracy, especially with ATL and TPM (74.4%). At the end, we noticed that using Stem with UniGram produced the best F1Train (77%), and F1Test (74.6%).

From the results above, we can notice that content-based features played an effective role in the dialect prediction model because different Arabic dialects use different words to reflect the same meaning, for example, the concept of "much" is represented using "كتير" in Levantine, "أوي" in Egyptian, "بزاف" in Maghreban and "وايد" in Gulf dialects. We can notice also that best results were obtained using Stem as a content-based feature (compared to Fullform and Lemma). This can be explained by the fact that several words related to the same origin are represented by one stem, for example, the words "الجامعات", "جامعاتهم" and "الجامعة" have the same stem "جامع". Consequently, by using stems, we make a trade-off between the number of features and the origins of the words.

Regarding style-based features, we can notice that ATL has a good effect on the results. It seems that some dialects allow expressing the idea using reduced number of words. Lengthening words is a commonly used practice in social media. It seems that the use of lengthened words differs from one dialect to another, thus, LEN enhanced the prediction model. Using character n-gram combines the best features of uni-gram of words and uni-gram of stems with all related prefixes and suffixes in one feature vector. Therefore, using CNGram enables taking advantage of both words and stems used at the same time without duplication.

3.2. Classifiers comparison

Here, we used the best features vector discovered in section 4.1 to train new models using different classifiers. We compared SVM, NB, RF and SMO classifiers. The results are presented in Figure 2. The best classifier was random forest (RF). It produced the best F1Train (80.6%) and F1Test (78.2%). It is worth mentioning that the appropriate choice of the classifier is considered a major step of any machine learning problem. The configuration of the classifier itself plays a crucial role also. In our research, when we used SVM, we noticed that the kernel of SVM is a very important parameter which should be selected accurately. We tried the linear, polynomial and exponential kernel. The polynomial kernel gave the best result (F1Train was 69.7% for linear, 77% for polynomial, 70.6% for exponential).

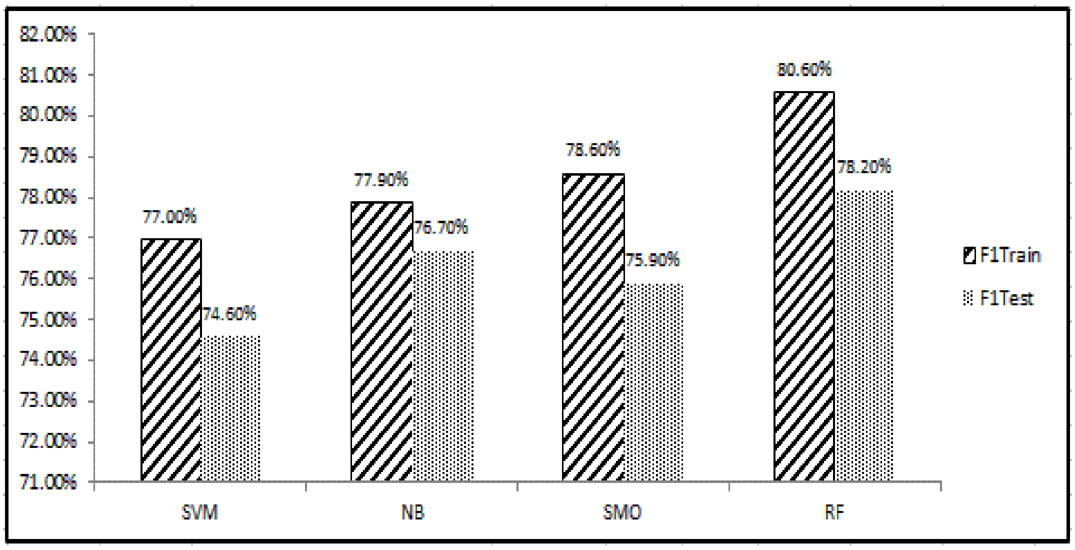

Figure 2. Classifiers comparison

4. CONCLUSION

In this research, we presented our work in author profiling of Arabic tweeps concerning dialect trait. We trained several models using various features and classifiers to find out the best model for predicting author dialect. We found that using RF classifier with full forms and their stems as a feature vector led to the best model with F1Train (80.6%) and F1Test (78.2%). It will be worth investigating using lemmatizer for Arabic vernaculars instead of the currently used lemmatizer which is made for modern standard Arabic (MSA). Moreover, we intend to study the effect of using deep learning algorithms for Arabic dialects classification in case of availability a huge dataset collected from Arabic writers.

REFERENCES

[1] M. B. op Vollenbroek, et al., "GronUP: Groningen User Profiling. Notebook for PAN at CLEF 2016," *Conference and Labs of the Evaluation Forum, Évora, Portugal, CEUR Workshop Proceedings*, 2016, pp. 846-857.

[2] TNS, "Arab Social Media Report," *Arab Social Media Influencers Summit*, 2015.

[3] M. A. Ali, "Artificial intelligence and natural language processing: the Arabic corpora in online translation software," *International Journal of Advanced and Applied Sciences*, vol. 3, no. 9, pp. 59-66, 2016.

[4] F. Huang, “Improved Arabic Dialect Classification with Social Media Data,” *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2015, pp. 2118-2126.

[5] A. Ali, et al., “Automatic Dialect Detection in Arabic Broadcast Speech,” *arXiv: 1509.06928*, 2015.

[6] E. Weren, et al., “Examining multiple features for author profiling,” *Journal of Information and Data Management*, vol. 5, no. 3, pp. 266-279, 2014.

[7] K. Alrifai, Ghaida Rebdawi, and Nada Ghneim, “Comparison Of Machine Learning Approaches In Arabic Tweeps Gender Prediction,” *International Journal of Scientific & Technology Research*, vol. 8, no. 11, pp. 2892-2895, 2019.

[8] F. Rangel, et al., “Overview of the 5th Author Profiling Task at PAN 2017: Gender and Language Variety Identification in Twitter,” *Working notes papers of the CLEF*, pp. 1613-0073, 2017.

[9] G. Kheng, et al., “INSA LYON and UNI PASSAU’s participation at PAN@CLEF’17: Author Profiling task,” *Conference and Labs of the Evaluation Forum, Dublin, Ireland, CEUR Workshop Proceedings*, 2017.

[10] I. Markov, et al., “Language- and Subtask-Dependent Feature Selection and Classifier Parameter Tuning for Author Profiling,” *Conference and Labs of the Evaluation Forum, Dublin, Ireland, CEUR Workshop Proceedings*, 2017.

[11] A. M. Ciobanu, et al., “Including Dialects and Language Varieties in Author Profiling,” *Conference and Labs of the Evaluation Forum, Dublin, Ireland, CEUR Workshop Proceedings*, 2017.

[12] A. Poulston, et al., “Using TF-IDF n-gram and Word Embedding Cluster Ensembles for Author Profiling,” *Conference and Labs of the Evaluation Forum, Dublin, Ireland, CEUR Workshop Proceedings*, 2017.

[13] N. Schaetti, “UniNE at CLEF 2017: TF-IDF and Deep-Learning for Author Profiling,” *Conference and Labs of the Evaluation Forum, Dublin, Ireland, CEUR Workshop Proceedings*, 2017.

[14] M. Kocher and J. Savoy, “UniNE at CLEF 2017: Author Profiling Reasoning,” *Conference and Labs of the Evaluation Forum, Dublin, Ireland, CEUR Workshop Proceedings*, 2017.

[15] Y. Adame-Arcia, et al., “Author Profiling, instance-based Similarity Classification,” *Conference and Labs of the Evaluation Forum, Dublin, Ireland, CEUR Workshop Proceedings*, 2017.

[16] E. S. Tellez, et al., “Gender and language-variety identification with MicroTC,” *Conference and Labs of the Evaluation Forum, Dublin, Ireland, CEUR Workshop Proceedings*, 2017.

[17] J. A. Khan, “Author Profile Prediction Using Trend and Word Frequency Based Analysis in Text,” *Conference and Labs of the Evaluation Forum, Dublin, Ireland, CEUR Workshop Proceedings*, 2017.

[18] M. Martinc, et al., “PAN 2017: Author Profiling-Gender and Language Variety Prediction,” *Conference and Labs of the Evaluation Forum, Dublin, Ireland, CEUR Workshop Proceedings*, 2017.

[19] L. Akhtyamova, et al., “Twitter Author Profiling UsingWord Embeddings and Logistic Regression,” *Conference and Labs of the Evaluation Forum, Dublin, Ireland, CEUR Workshop Proceedings*, 2017.

[20] R. R. Oliveira and R. F. de O. Neto, “Using character n-grams and style features for gender and language variety classification,” *Conference and Labs of the Evaluation Forum, Dublin, Ireland, CEUR Workshop Proceedings*, 2017.

[21] A. Ogaltsov and A. Romanov, “Language Variety and Gender Classification for Author Profiling in PAN 2017,” *Conference and Labs of the Evaluation Forum, Dublin, Ireland, CEUR Workshop Proceedings*, 2017.

[22] A. Basile, et al., “N-GrAM: New Groningen Author-profiling Model,” *arXiv:1707.03764,* 2017.

[23] D. Kodiyan, et al., “Author Profiling with Bidirectional RNNs using Attention with GRUs,” *Conference and Labs of the Evaluation Forum, Dublin, Ireland, CEUR Workshop Proceedings*, 2017.

[24] S. Sierra, et al., “Convolutional Neural Networks for Author Profiling,” *Conference and Labs of the Evaluation Forum, Dublin, Ireland, CEUR Workshop Proceedings*, 2017.

[25] M. Franco-Salvador, et al., “Subword-based Deep Averaging Networks for Author Profiling in Social Media,” *Conference and Labs of the Evaluation Forum, Dublin, Ireland, CEUR Workshop Proceedings*, 2017.

[26] A. Abdelali, et al., “Farasa: A Fast and Furious Segmenter for Arabic,” *Qatar Computing Research Institute, Hamad Bin Khalifa University, Doha, Qatar. Proceedings of NAACL-HLT 2016 (Demonstrations)*, San Diego, California, Association for Computational Linguistics, 2016, pp. 11-16.

BIOGRAPHIES OF AUTHORS

Ghaida Rebdawi Ph.D. in Software Engineering from INSA de Lyon, France. Research Director, Deputy Director for Educational Affairs, and Professor of Software Engineering at HIAST, Damascus Syria. Current researches include the use of Business Process Modeling to Manage change Requirements in Agile Software Development, Author profiling from Arabic Social media using Machine Learning and NLP techniques, and the development of an Ontology in Arabic. Co-author of many e-Content in Software Engineering for Syrian Virtual University (SVU). Co-author of many Books in Software Engineering in Arabic language, and in the translation of IT Books and in the production of professional dictionaries from English to Arabic.

Nada Ghneim Ph.D. in Language Sciences (Speech Communication) from the "Institut de la Communication Parlée"- Stendhal (Grenoble III) University, France, 1997, and a Postgraduate Degree (DEA) in Artificial Intelligence (Image, Robotics, Vision), from the National High School of Computer Science and Applied Mathematics in Grenoble (ENSIMAG), France, 1993. Nowadays, I'm an Assistant Professor at the Faculty of Informatics & Communication Engineering, at the Arab International University (AIU), Damascus, Syria. I'm also a Researcher/Lecturer, at Higher Institute for Applied Sciences and Technology (HIAST), and at the Information Technology Engineering Faculty (Damascus University). I'm a member of the Syrian Computer Society, and I have many publications in Speech and Natural Language Processing domain, such as Arabic Text-to-Speech, Sentiment Analysis, Morphological and Syntactic Analysis, Dictionary and Ontology Building.

Khaled Alrifai Ph.D. candidate in Higher Institute for Applied Sciences and Technology HIAST, Damascus, Syria. I hold a master degree from HIAST entitled: Information and decision support system, and license degree in information technology engineering from Damascus university specialized in artificial intelligence. Currently, I'm interested in Arabic data analysis researches and in all related AI techniques. I depend on NLP and data mining to be used in case of Arabic language to propose beneficial tools for business and academic purposes.