Multi-Lingual Language Variety Identification using Conventional Deep Learning and Transfer Learning Approaches

Sameeah Noreen Hameed
School of Software, East China Jiaotong University, China
sami.hameed89@gmail.com

Muhammad Adnan Ashraf
Department of Computer Science, Northwestern Polytechnical University, China
adnan.ashraf@mail.nwpu.edu.cn

Qiao Ya-nan
School of Computer Science and Technology, Xi'an Jiaotong University, China
qiaoyanan@mail.xjtu.edu.cn

Abstract: *Language variety identification tends to identify lexical and semantic variations in different varieties of a single language. Language variety identification helps build the linguistic profile of an author from written text which can be used for cyber forensics and marketing purposes. Investigating previous efforts for language variety identification, we hardly find any study that experiments with transfer learning approaches and/or performs a thorough comparison of different deep learning approaches on a range of benchmark datasets. So, to bridge this gap, we propose transfer learning approaches for language variety identification tasks and perform an extensive comparison of them with deep learning approaches on multiple varieties of four widely spoken languages, i.e., Arabic, English, Portuguese, and Spanish. This research has treated this task as a binary classification problem (Portuguese) and multi-class classification problem (Arabic, English, and Spanish). We applied two transfer learning Bidirectional Encoder Representations from Transformers (BERT), Universal Language Model Fine-tuning (ULMFiT), three deep learning-Convolutional Neural Networks (CNN), Bidirectional Long Short Term Memory (Bi-LSTM), Gated Recurrent Units (GRU), and an ensemble approach for identifying different varieties. A thorough comparison between the approaches suggests that the transfer learning based ULMFiT model outperforms all other approaches and produces the best accuracy results for binary and multi-class language variety identification tasks.*

Keywords: *Language variety identification, deep learning, transfer learning, binary classification.*

Received July 25, 2021; accepted December 13, 2021
https://doi.org/10.34028/iajit/19/5/1

1. Introduction

There are currently more than 7000 languages in the world[1], which are often categorized as regional languages and global languages. Regional languages are limited to a specific area, whereas global languages are spoken worldwide [7]. These global languages are often the first language in many regions worldwide, such as English, Spanish, Arabic, etc., However, the dialects, vocabulary, and grammar used in these global languages differ according to their spoken regions. For example, in the United Kingdom, the term "film" is used, whereas in the United States, it is termed as a "movie" These variations are influenced by their regional languages, culture, and social norms, which impact the structure of language such as semantic, grammar and spelling idiosyncrasies [22]. These subtle variations in the languages are termed language variety. Language variety identifications a popular task in Natural Language Processing (NLP) due to its wide potential applications in computer forensics, marketing, and content recommendation [12].

Speakers of different regions have a different dialect and accent, which makes identification of region/country using dialect relatively easy. In contrast, it becomes challenging to identify language variety from written text [23]. The task becomes even more challenging when the text is limited and noisy, such as social media text. It is challenging to find cues related to grammar, morphology, lexis, and syntax that helps to identify which region/variety language belongs.

Previously, researchers have researched using traditional machine learning approaches and deep learning approaches for language variety identification tasks [5]. Zampieri and Gebre [23] used character n-gram and word n-grams to identify two Portuguese varieties: European and Brazilian. Similarly, Lee and Bosch [15] used text statistics, syntactic features, and word n-grams features to identify language variety on Netherlandic and the Flemish variants of Dutch. Sierra *et al*. [20] developed deep learning based CNN model for language variety identification tasks in English, Spanish, Portuguese, and Arabic. Transfer learning approaches are current state-of-the-art for many NLP tasks. However, we hardly observe any effort towards language variety identification tasks using transfer learning.

[1] https://www.ethnologue.com/guides/how-many-languages

This research is an effort towards automatic language variety identification tasks using a range of deep learning and transfer learning approaches. We focus on four different languages including Arabic (Egypt, Gulf, Levantine, and Maghrebi), English (Australia, Canada, Great Britain, Ireland, New Zealand, and United States, Portuguese (Portugal and Brazil), and Spanish (Argentina, Chile, Columbia, Mexico, Spain, Peru, and Venezuela). We used PAN 2017, Rangel *et al*. [18] Profiling Competition corpora for this research. The language varieties are distributed into two classification problems, i.e., binary (Portuguese) and multiclass (Arabic, English, and Spanish). We applied four deep learning based approaches: CNN, Bi-LSTM, GRU, and an ensemble model, whereas Google's BERT and FastAi's ULMFiT were used as transfer learning approaches. Further, a thorough comparison of deep learning and transfer learning models for language variety identification tasks has been performed.

This paper is organized as follows: First, we present the related work in section 2. Next, in section 3, we describe the methods for language variety identification. Subsequently, the dataset used in this research is discussed in section 4. We detail the experiments carried out and the results in section 5. Finally, we give conclusions and future works in section 6.

2. Related Work

Language variety identification is a sub-task of the language identification task. Language Identification is a task to determine the language from written text automatically. Language variety identification is a relatively complex task than language identification as it calls for finding variation in the same language rather than two different languages.

PAN competition on author profiling task is one of the primary contributors of language variety identification task. Language variety identification, together with gender identification, was part of the PAN 2017 Author Profiling Competition [18]. The focus of the task was the small and noisy data of Twitter tweets [17]. It focused on four languages along with the multiple sub-varieties. The languages include English, Arabic, Portuguese, and Spanish. Most of the participants opted machine learning based approaches, including word embedding, tf-idf, n-grams, and stylistic patterns. Few participants also experimented with deep learning techniques, including CNN and RNN. The overall best result was obtained using a combination of character n-grams (with n between 3 and 5) and tf-idf word n-grams (with n between 1 and 2) [1]. The language variety best accuracy results on different languages were Arabic 0.8313, English 0.8988 Portuguese 0.9813 Spanish 0.9621, and Average 0.9184.

Discriminating between Similar Languages (DSL) shared tasks in 2017 held as a part of for Similar Languages, Varieties, and Dialects (VarDial) [24]. Eleven teams participated for the task that involved determining the language or language variety for written news extracts. It featured 14 different languages, including multiple verities. These languages include Bosnian, Croatian, Serbian; Malay and Indonesian; Persian and Dari; Canadian and Hexagonal French; Brazilian and European Portuguese; Argentinian, Peninsular, and Peruvian Spanish. The top team used word and character n-grams for feature selection with Support Vector Machine (SVM) classifier. They achieved the best result of F1 score=0.927 [2]. Most of the top-performing teams used a similar approach which falls under traditional machine learning approaches.

In the 2018 VarDial Evaluation Campaign [25], two tasks were conducted. Task one discriminated between Dutch and Flemish in Subtitles (DFS), and task two was Indo-Aryan Language Identification (ILI). DFS was composed of a dataset consisting of over 50,000 subtitle phrases of a movie. The primary purpose was to identify text is written in the Netherlandic or the Flemish variant of the Dutch language. This was to step towards developing and comparing language variety classification models using subtitles and thereby analyzing the proximity of the language varieties in a new way. The best result was obtained from the Tubingen-Oslo [5] team that used one system based on a linear SVM classifier and another based on RNN with the result of F1 score=0.66. Similarly, the 2019 edition of VarDial [26] focused on discriminating between Mainland and Taiwan variations of Mandarin Chinese, in addition to other tasks. The participant proposed traditional machine learning, deep learning, and ensemble approaches. The best results for traditional Chinese were obtained using the character n-gram approach, whereas for the simplified approach ensemble model produced the best results [26]. VarDial 2020 Evaluation campaign [9] did not focus specifically on language variety identification tasks.

Zaghouani and Charfi [21] collected Twitter tweets for Arabic language variety identification. They collected data from 11 regions and 16 countries and termed the dataset Arap-Tweets. They collected the dataset based on the words that differentiate these dialects and the location of the users. Similarly, Franco-Salvador *et al*. [8] performed the classification of language varieties on the different versions of Spanish. They used word embeddings to classify Spanish varieties from blog data automatically. Similarly, Rangel *et al*. [19] also used the Spanish blogs for the language variety identification task. They experimented successfully with the low dimensional model based on text statistics.

To summarize, the related work shows that language variety identification is a widely researched topic in the NLP. The current state-of-the-art approaches are built upon the traditional machine learning based N-gram approach. There have been only a few deep learning based efforts, with limited success. Also, the literature lacks a thorough evaluation of deep learning approaches for language variety identification.

3. Methods for Language Variety Identification

Methods for language variety identification are divided into three main parts:

1. Deep learning.
2. Ensemble.
3. Transfer learning methods.

3.1. Deep Learning Methods

Three deep learning models have been used, which include CNN, Bi-LSTM, and GRU:

- **Convolutional Neural Network**

Convolutional Neural Networks (CNN) [14] has been applied for the different author profiling tasks [11], including language variety identification [19]. CNN consists of three main layers, i.e., an input layer, a convolutional layer with pooling, and an output layer. The data is provided to the Input layer, which forwards the data to the Convolutional Layer. Convolutional Layer makes use of a fixed-sized sliding window (kernel size) to extract the convolutions of fixed length and apply filters over all the sentences sequentially on the input data. It is the concatenated embedding vectors in the given window size and a weight vector (filters). The activation function is then applied to the concatenated vector. The pooling activity is then utilized to consolidate the feature from various convolution windows into a vector. We apply Max-pooling in our experiments which captures the high values and discards the rest of the values from resultant vectors of convolutions. A fully Connected Layer followed by Softmax Layer is then applied to produce the probabilities of each language variety as an output.

- **Bi-directional Long Short Term Memory**

Long Short Term Memory (LSTM) [16] is a widely used variant of Recurrent Neural Network (RNN). LSTM retains the previous information in the memory cells and uses it to process the next instruction cyclically. It has four NN layers (three gates and one cell state), At first, the forget gate chooses information dependent on verifiable historical data that is to be held or disposed of. Input gate distinguishes the information to be updated for computing a new candidate vector. The candidate vector converges with the output gate esteems to refresh the new cell state. Lastly, the filtered vector information to be outputted is sent through the output gate.

Bidirectional LSTM [10] is an augmentation to the basic LSTM. Bi-LSTMs train two LSTMs rather than one LSTM on the information succession. The first on the information succession as-is in the forward direction and the other in the opposite backward direction. This bi-directional nature helps capture a better understanding of language, which can be typically very helpful in language variety identification.

- **Gated Recurrent Units**

Gated Recurrent Unit (GRU) [4] is also a variation of RNN to tackle the vanishing gradient problem. GRU utilizes two gates instead of the three gates and cell state in LSTM. These gates are called update gate and reset gate. Fundamentally, these are two vectors that choose what data ought to be passed as output. The exceptional thing about them is that they can keep important data from quite a while in the present in memory, without forgetting it through time or expelling data that is immaterial to the expectation. That helps in memorizing the complex language structure for efficient feature extraction and classification.

3.2. Ensemble Method

The ensemble method tries to improve the prediction capability by consolidating the results of more than one classifier. There are various flavors of ensemble strategies, and they significantly rely upon the dataset [3]. In ensemble methods, results of the individual classifiers are disregarded, and these results are combined with multiple other classifiers to improve the overall results of a specific task.

For our investigations, we used CNN, Bi-LSTM, and GRU model as our base classifiers and then applied the majority voting technique as our ensemble method. The majority voting technique records the predicted classes of all the base classifiers. After recording all the outputs, the model picks the class with the most votes/predictions as the final output.

3.3. Transfer Learning Methods

Transfer learning models extract information from one task and use it to predict a comparable task. It includes a pre-trained language model that is trained on an enormous amount of information and afterward fine-tuned to align the pre-trained model on the target task. It is further fine-tuned to perform the target task. The efficiency is significantly improved for the tasks that don't have a huge amount of data and/or require a computational capacity to train deep learning models. Two widely known Transfer learning models used in this research are

1. BERT
2. ULMFiT. These models are discussed in detail below:

- **BERT**

Bidirectional Encoder Representations from Transformers (BERT) [6] is based on the Transformer model. It uses Masked Language Model (MLM) and the next sentence prediction task to generate a Language Model (LM). LM intends to get a profound comprehension of language. The pre-trained LM is then used to fine-tune the BERT model for target task classification.

The BERT model uses multiple stacked encoders layers termed the transformer layer. Each encoder layer consists of numerous feed forward network layers and self-attention heads. The initial input token is provided with a classification token [CLS]. BERT accepts a succession of words as information in the form of embeddings that keeps streaming up the stack of layers. Each layer puts forth a concentrated effort consideration using the self-attention layer and forwards its outcomes through a feed-forward system. Afterward, it hands it off to the following encoder layer. Depending upon the classification task, the focus of each output vector changes. For the language variety identification task, the focus remains on the [CLS] token vector. This vector is then further passed to the single-layered feed forward neural network followed by the softmax layer that outputs the prediction results.

- **ULMFiT**

ULMFiT [13] performs text classification tasks by following three stages

1. General domain language model pre-training.
2. Target task language model fine-tuning.
3. Target task classifier fine-tuning.

General domain language modeling (next word prediction) can catch the general properties it fills in as a perfect source task for pre-preparing a system. The system is preprepared on Wikitext-103 [17]. The pre-trained language model is then fine-tuned for the target task. This progression improves the order model on small datasets. After the language model is trained and fine-tuned, ULMFiT requires the final stage of the target task classifier fine-tuning. This stage requires training the model for two linear blocks, an activation layer ReLU and finally, the softmax layer. The two primary operations performed in the linear layers are often termed Concat pooling and gradual unfreezing. To avoid losing the impact of concat pooling and maximizing the benefit of the finetuned language model, gradual unfreezing for fine-tuning the classifier is used. The last LSTM layer is first unfrozen, and the model is calibrated for one iteration. After that, the next lower layer is unfrozen. The process is repeated for all the layers until all the layers are fine-tuned and converge. Both a forward and backward language model are pre-trained. The classifier is then fine-tuned for both the language models independently. The average of the two classifier predictions is taken as the final output.

4. Data

For this research, we used PAN 2017 Author Profiling corpora. It consists of four languages, with each language having multiple varieties. These languages include English, Spanish, Arabic, and Portuguese. For Arabic, there were four varieties: Egypt, Gulf, Levantine, and Maghrebi. Six different varieties were used for English, including Australia, Canada, Great Britain, Ireland, New Zealand, and the United States. For Portuguese, two language varieties were included, i.e., Portugal and Brazil. For Spanish, seven different varieties considered were Argentina, Chile, Columbia, Mexico, Spain, Peru, and Venezuela.

The corpus consisted of 1000 authors per language variety with100 tweets per author. The complete corpus is divided into 60% and 40% ratios for each language variety for training and test datasets. Since only the Training corpus is publicly available, we have used only the training sets for our experimentation-ns. The training corpus is balanced for all varieties in a single language. Each variety consists of 600 authors with 100 tweets per author. The detail of each dataset used for experimentation is given in Table 1.

Table 1. Number of authors and tweets in training corpus.

Language	Authors	Tweets
Arabic	2400 Authors	24000 Tweets
English	3600 Authors	36000 Tweets
Portuguese	1200 Authors	12000 Tweets
Spanish	4200 Authors	42000 Tweets

5. Experiments and Results

The experiments are performed in two different settings.

1. Deep learning models (including ensemble)
2. Transfer learning models.

5.1. Experimental Setup

All the experiments were performed using Python 3.5 framework with Keras deep learning library and Tensorflow as backend except BERT that was performed using PyTorch. The experiments were performed on the four different language varieties sub-corpora, i.e., Arabic, English, Portuguese, and Spanish (see section 4).

Each language dataset was divided into three parts

1. Training.
2. Validation.
3. Test. 90% of each dataset was used for training purposes, and 10% of the data was used to test the trained models. 10% of the training data was utilized for validation purposes. The reason for selecting train-test split instead of the K-fold was primarily based on PAN 2017 language variety

competition, where train-test split was used for evaluation. So, to effectively compare our results with the PAN 2017 competition results, we also used the train-test split approach.

All the parameters were found through early experimentations and finding the optimum parameters. The CNN-based optimum parameters are given in Table 2. For the experiments using Bi-LSTM and GRU, a single-layered network was utilized with the memory units/ neurons set to 64, recurrent dropout was set 0.3. Batch size and Epochs were the same as in Table 2.

Table 2. CNN optimum parameters.

Parameter	Value
Layers	3
Filters numbers	128
Kernel size	3
Activation function	ReLU
Pooling type	Max
DropOut	0.5
Batch Size	256
Epochs	50

Regarding the Transfer Learning models, For BERT we used the BERT-Multilingual that supports 104 languages including all the four languages used in this research. On the other hand, ULMFiT only has one pre-trained language model available, which is of the English Language. So, we used the pre-trained model for only English Language. For other languages, we generated a custom language model from scratch.

5.2. Results Using Deep Learning Models

Four deep learning models, CNN, Bi-LSTM, GRU, and ensemble method, were used for experimentation. Table 3 shows the results obtained using deep learning models and the ensemble approach. The results show that the best result (Accuracy=97.50%) was obtained for the Portuguese language using the CNN model.

Table 3. Results using deep learning and ensemble approach.

Language	Problem Type	Model	Accuracy
Arabic	Multi-class	GRU	67.50
		Bi-LSTM	59.17
		CNN	69.58
		Ensemble	67.50
English	Multi-class	GRU	71.94
		Bi-LSTM	37.70
		CNN	69.17
		Ensemble	61.94
Portuguese	Binary	GRU	92.50
		Bi-LSTM	95.83
		CNN	**97.50**
		Ensemble	96.67
Spanish	Multi-class	GRU	87.14
		Bi-LSTM	40.48
		CNN	81.12
		Ensemble	76.67

The results show that the CNN model produces the best results of 69.58% for the Arabic language. The GRU model and the ensemble method produce slightly lesser accuracy test results of 67.50%, whereas the Bi-LSTM model produced the least 59.17%. This suggests that CNN was better able to identify the discriminating features among the different varieties of the Arabic language. All the results obtained were able to surpass the base Most Common Category (MCC) score.

For the English language, the results depict that the GRU model produces the best result (Accuracy 71.94%). The CNN model has lesser accuracy results of 69.17% compared to the GRU. Interestingly, the Bi-LSTM model results were very low compared to the other models. This shows that the bidirectional nature of the LSTM model doesn't help to find accurate discriminating features for the English Language varieties. All the results obtained were again able to surpass the base MCC score.

Portuguese language variety identification was treated as a binary classification problem as it has only two classes, i.e., Portugal and Brazil. It produced higher results than any other language. The best result (Accuracy=97.50%), as discussed above, was obtained using the CNN model. The ensemble method produced a slightly lower accuracy score of 96.67%. All the other models also had very encouraging results.

Spanish is the most widely spoken language in the world. The results for Spanish shows that the GRU model was able to produce the best accuracy result of 87.14%. The CNN model produced the lesser accuracy results of 81.12% compared to the GRU. The ensemble method produces an accuracy of 76.67%, whereas the Bi-LSTM model produced was again very low (40.48%). This suggests that GRU was better able to identify the discriminating features among the different varieties of Spanish.

5.3. Results using Transfer Learning Models

Table 4 shows that the highest accuracy score of 98.03% was obtained on Portuguese language varieties using the ULMFiT model. Considering other languages, the highest results for Arabic (Accuracy= 75%), for English (Accuracy=78.89%), and for Spanish (Accuracy=90.71%) were obtained using the ULMFiT model.

Table 4. Results using a transfer learning approach.

Language	Problem Type	Model	Accuracy
Arabic	Multi-class	ULMFiT	75.00
		BERT	68.06
English	Multi-class	ULMFiT	78.89
		BERT	73.46
Portuguese	Binary	ULMFiT	**98.03**
		BERT	96.67
Spanish	Multi-class	ULMFiT	90.71
		BERT	81.12

Generally, the ULMFiT model results were towards the higher side than the BERT model. The BERT model results were also very encouraging, and it showed similar patterns as the ULMFiT model. The highest accuracy result of 96.67% using the BERT model was achieved for the Portuguese language varieties. For the Arabic language varieties identification task, 68.06% was the accuracy score. For English, 73.46% was the accuracy result. 81.12% accuracy score was achieved using the BERT model for the Spanish language variety identification task.

The best result didn't outperform the PAN 2017 competition results (Arabic-83.13%, English-90.04%, Portuguese–98.03%, and Spanish-90.71%) [17] As the experiments were performed on only the training set. It reduces the training size as well the testing data is different than the PAN 2017. However, these results are on a similar pattern, and the result using ULMFiT on the Portuguese language is theoretically similar to PAN 2017 best. Additionally, these results give a clear guideline to the researchers seeking advice on which deep learning and transfer learning models are best for the language variety identification task.

Both deep Learning and Transfer Learning produced very encouraging results for the task of automated language variety identification. Figure 1 compares the results obtained using deep learning models and the transfer learning models. The results suggest that both the different techniques produced a similar pattern of results for all the language varieties. Our analyses bolster various critical disclosures for the reasonable execution of neural network-based model both in the case of binary class and multiclass classification on the small to medium level.

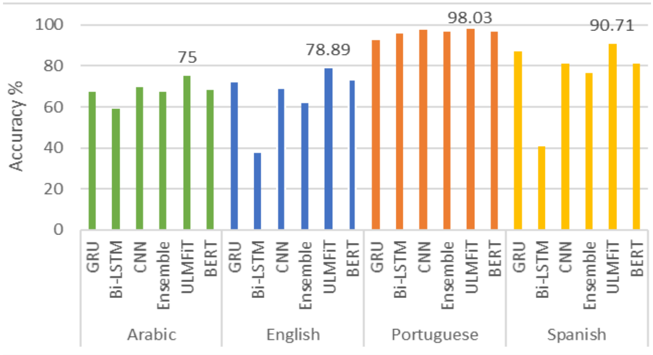

Figure 1. Comparison of experimental results.

The deep learning and transfer learning models have a significantly greater efficiency while treating binary classification. The results are slightly on the lower side when treating multi classification problems. The deep learning models were able to predict better when a substantially large amount of data was used (Spanish); the number of classes in these cases doesn't have any impact. Spanish language having seven different language varieties and English having six language varieties were able to produce better accuracy results than the Arabic language having the four language varieties. It also highlights that it is vital to semantically comprehend the entire task. Further, transfer learning models perform better even when the dataset size is not enormous. The pre-trained language models trained to huge dataset helps effectively in learning the semantics of the language, which can help differentiate even the closely related languages and their varieties'. Also, transfer learning models performed better than the deep learning models when the dataset is small to medium. This also upholds the fact that deep learning models require huge datasets for better performance.

6. Conclusions

This research attempts to perform an automatic language variety identification task on a set of different languages and their multiple varieties. Compared to the prior effort, transfer learning (ULMFiT and BERT) based approaches have been introduced along with a broader set of deep learning approaches (CNN, Bi-LSTM, GRU, and an ensemble model) on language varieties of Arabic, English, Portuguese, and Spanish. Our thorough experimentations concluded that the transfer learning based ULMFiT model outperforms all other models on both binary and multi-class classification. Among the deep learning models, CNN and GRU models were the ones that performed better. The ensemble model was not able to improve efficiency. The highest accuracy result was obtained on the binary classification task of the Portuguese language using the ULMFiT model. In the future, we would like to enhance the language variety identification to other widely spoken languages and investigate the task using other transfer learning models.

References

[1] Basile A., Gareth D., Maria M., Josine R., Hessel H., and Nissim M., "Is there Life beyond N-Grams? A Simple SVM-Based Author Profiling System," *in Working Notes of CLEF*, Ireland, pp. 1-7, 2017.

[2] Bestgen Y., "Improving the Character Ngram Model for The DSL Task With BM25 Weighting and Less Frequently Used Feature Sets," *in Proceedings of the 4th Workshop on NLP for Similar Languages, Varieties and Dialects*, Valencia, pp. 115-123, 2017.

[3] Chan S., Honari M., Benetti B., Lakhani A., and Fyshe A., "Ensemble Methods for Native Language Identification," *in Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, Copenhagen, pp. 217-223, 2017.

[4] Chung J., Gulcehre C., Cho K., and Bengio Y., "Empirical Evaluation of Gated Recurrent

Neural Networks on Sequence Modeling," *in NIPS*, 2014.

[5] Çoltekin C., Rama T., and Blaschke V., "Tübingen-Oslo Team at the VarDial 2018 Evaluation Campaign: An Analysis of N-Gram Features in Language Variety Identification," *in Proceedings of the 5th Workshop on NLP for Similar Languages, Varieties and Dialects,* Santa Fe, pp. 55-65, 2018.

[6] Devlin J., Chang M., Lee K., and Toutanova K., "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *in Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*, United Stated of America-USA, pp. 4171-4186, 2019.

[7] Dunn J., "Mapping Languages: The Corpus of Global Language Use," *Language Resources and Evaluation*, vol. 54, no. 4, pp. 999-1018, 2020.

[8] Franco-Salvador M., Rangel F., Rosso p., Taulé M., and Antònia M., "Language Variety Identification Using Distributed Representations of Words and Documents," *in Proceedings of the Conference on Cross-Language Evaluation Forum For European Languages*, Toulouse, pp. 28-40, 2015.

[9] Gaman M., Dirk H., Ionescu R., Jauhiainen H., JauhiainenT., Lindén K., Ljubešić N., Partanen N., Purschke C., Scherrer Y., and Zampieri M., "A report on the VarDial Evaluation Campaign," *in Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects,* Barcelona, pp. 1-14, 2020.

[10] Graves A. and Schmidhuber J., "Framewise Phoneme Classification with Bidirectional LSTM Networks," *in Proceedings IEEE International Joint Conference on Neural Networks*, Montreal, pp. 2047-2052, 2005.

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

[12] Issa H., Issa S., and Shah W., "A Novel Method for Gender and Age Detection Based on EEG Brain Signals," *The International Arab Journal of Information Technology*, vol. 18, no. 5, pp. 704-710, 2021.

[13] Jeremy H. and Ruder S., "Universal Language Model Fine-Tuning for Text Classification," *in Proceedings of 56th Annual Meeting of the Association for Computational Linguistics,* Australia, pp. 328-339, 2018.

[14] Kalchbrenner N., Grefenstette E., and Blunsom P., "A Convolutional Neural Network for Modelling Sentences," *in Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, Baltimore, pp. 655-665, 2014.

[15] Lee C. and Bosch A., "Exploring Lexical and Syntactic Features for Language Variety Identification," *in Proceedings of the 4th Workshop on NLP for Similar Languages, Varieties and Dialects*, Valencia, pp. 190-199, 2017.

[16] Merity S., Keskar S., and Socher R., "Regularizing and Optimizing LSTM Language Models," *in Proceedings of International Conference on Learning Representations*, pp. 1-8, 2018.

[17] Miura Y., Taniguchi T., Taniguchi M., Misawa S., and Ohkuma T., "Using Social Networks to Improve Language Variety Identification with Neural Networks," *in Proceedings of the 8th International Joint Conference on Natural Language Processing*, Taipei, pp. 263-270, 2017.

[18] Rangel F., Rosso P., Potthast M., and Stein B., "Overview of the 5th Author Profiling Task at pan 2017: Gender and Language Variety Identification in Twitter," in CLEF, Ireland, pp. 1-26, 2017.

[19] Rangel F., Franco-SalvadorM., and Rosso P. "A Low Dimensionality Representation for Language Variety Identification," *in Proceedings of International Conference on Intelligent Text Processing and Computational Linguistics*, Konya, pp. 156-169, 2016.

[20] Sierra S., Montes-y-Gómez M., Solorio T., and González F., "Convolutional Neural Networks for Author Profiling," in CLEF, Ireland, pp. 1-7, 2017.

[21] Zaghouani W., and Charfi A., "Arap-Tweet: A Large Multi-Dialect Twitter Corpus for Gender, Age and Language Variety Identification," *in Proceedings of the 11th International Conference on Language Resources and Evaluation*, Miyazaki, pp 1-18, 2018.

[22] Zampieri M., Nakov P., and Scherrer Y., "Natural Language Processing for Similar Languages, Varieties, and Dialects: Survey," *Natural Language Engineering*, vol. 26, no. 6, pp. 595-612, 2020.

[23] Zampieri M. and Gebre B., "Automatic identification of language varieties: The case of Portuguese," *in Proceedings of 11th Conference on Natural Language Processing*, Vienna, pp. 233-237, 2012.

[24] Zampieri M., Malmasi S., Ljubešić N., Nakov P., Ali A., Tiedemann J., Scherrer Y., and Aepli N., "Findings of the VarDial Evaluation Campaign 2017," *in Proceedings of the 4th Workshop on NLP for Similar Languages, Varieties and Dialects*, Valencia, pp 1-15, 2017.

[25] Zampieri M., Malmasi S., Nakov P., Ali A., Shon S., and et al., "Language Identification and Morphosyntactic Tagging: The Second VarDial Evaluation Campaign, *in Proceedings of the 5th Workshop on NLP for Similar Languages,*

Varieties and Dialects, Santa Fe, pp. 1-17, 2018.

[26] Zampieri M., Malmasi S., Scherrer Y., Samardžić T., Tyers F., and et al. "A Report on The Third Vardial Evaluation Campaign," *in Proceedings of the 6th Workshop on NLP for Similar Languages, Varieties and Dialects*, Ann Arbor, pp 1-16, 2019.

Sameeah Noreen Hameed is instructor at the East China Jiaotong University, China. She received her Master's degree in Computer Science from Xi'an Jiaotong University (XJTU), China. Her current research interests include NLP, Information retrieval, and GIS.

Muhammad Adnan Ashraf is a Ph.D. scholar at Northwestern Polytechnical University, China. He is also working as a lecturer in Computer Science at COMSATS University Islamabad, Pakistan. His current research is in NLP and information retrieval.

Qiao Ya-nan is an associate professor at XJTU. He received his Ph.D. Degree in Computer Science from Xian Jiaotong University, China. His current research is in block chain, cloud computing, information retrieval, text mining and social network analysis. His research has been financially supported by National key R&D Program of China, National Natural Science Foundation of China, etc.