ISSN 2278-3091
Volume 9, No.1, January – February 2020
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse69912020.pdf
https://doi.org/10.30534/ijatcse/2020/69912020

Towards a system for predicting the category of educational and vocational guidance questions using bidirectional encoder representations of transformers (BERT)

Omar Zahour[1], El Habib Benlahmar[2], Ahmed Eddaoui[3], Oumaima Hourrane [4]
[1]Laboratory of Information Technology and Modeling, Hassan II University Of Casablanca, Faculty of Sciences Ben M'SIK,Morocco,orzahour@gmail.com
[2]Laboratory of Information Technology and Modeling,Hassan II University Of Casablanca , Faculty of Sciences Ben M'SIK,Morocco, h.benlahmer@gmail.com
[3]Laboratory of Information Technology and Modeling,Hassan II University Of Casablanca , Faculty of Sciences Ben M'SIK,Morocco, ahmed_eddaoui@yahoo.fr
[4]Laboratory of Information Technology and Modeling, Hassan II University Of Casablanca, Faculty of Sciences Ben M'SIK,Morocco,oumaima.hourrane@gmail.com

ABSTRACT

Educational and vocational guidance is a particularly important issue today, as it strongly determines the chances of successful professional integration into the increasingly difficult labor market. Families have understood this well since they have been interested, in the educational orientation of their child. In this sense, we have set up a system for classifying questions of educational and professional orientation in based on Holland's test using the BERT method. Text classification, particularly the classification of questions is a basic task in natural language processing which is a very profound concept in the field of artificial intelligence. As the most abundant data in the world today is in the form of texts, having a powerful word processing system is essential and is more than just a necessity. Recently, Transformers models such as Bidirectional encoder representations of transformers or BERT are a very popular NLP models known for producing remarkable results compared to other methods in a wide variety of NLP tasks. In this article, we demonstrate how to implement a multi-class classification using BERT. In particular, we explain the classification of questions concerning the field of educational and vocational guidance following the RIASEC typology of Holland. Our model allows us to obtain the category of each input question. In our case, we define four classes (Activity, Occupations, Abilities, and Personality) for the set of questions, which constitute our data set. The results of this approach demonstrate that our model achieves competitive performance.

Key words: Academic and vocational guidance; Text classification, Automatic natural language processing, BERT model, Holland RIASEC typology

1. INTRODUCTION

After a long-lasting study, the American researcher John Holland revealed that all the working people belong to one of the six types of workers. He named them as follows: "Realistic" (R), "Investigative" (I), "Artistic" (A), "Social", «Enterprising " (E), "Conventional". According to Holland - and numerous research have confirmed it - the profession or the trade chosen by a person is a form of expression of his personality and is, therefore, related to the type to which he belongs. His or her aptitudes, certain personality traits and interests would determine whether a worker belongs to any of the six types. So again, according to Holland, people of the same type would be attracted to the same kind of work. Why? Because these people are shaped by their personalities and by the fact that they pursue similar objectives and present the same physical or psychological dispositions towards their work. All people in a given job can be divided into six professional types.

The typology of a person is established by measuring his degree of affinity with each of the six types, to place these in order of importance, from the type, which corresponds best to him to that which corresponds least to him. For most people, it is especially the first two or three types of their personal classification, which determine their way of being and acting, both in their personal and professional life. For example, a person whose dominant type is "Investigator" and who has affinities with the "Realistic" type, it will be said to have an "IR" profile. To further characterize the typology of this person, it is possible to consider the third type to which he most closely resembles and to say, in case it is the "Social" type, that this person has an "IRS" profile.

Types can combine in all kinds of ways and it is the sort of their combination that determines personality.

In this article, our goal is to make an automatic classification of educational and vocational guidance questions according to the RIASEC typology of Holland using the BERT method in order to automatically generate the category of a new orientation question educational and professional, so that we can determine the three dominant types of a person's personality in order to orient them to the suitable trade for them according to the list of trades and professions.

Text classification is one of the fundamental tasks of natural language processing (NLP) to assign text to different categories. Text classification applications include sentiment analysis [1], textual similarities and plagiarism detection [2] [3] [4]; question classification [5] and subject classification [6]. Today, deep learning approaches have become the norm in categorizing texts, such as convolutional neural networks (CNN) [7], recurrent neural networks (RNN) [8] or certain more complex methods.

The method using deep learning for text classification requires entering text in a deep network to obtain a representation of the text. Then, it requires entering the text representation in the softmax function to obtain the probability of each category. CNN-based models [7] [9] [10] can obtain text representations with local information. RNN-based models [11] [12] can obtain text representations containing long-term information. Therefore, some methods must be modelled by combining the advantages of CNN and RNN, such as C-LSTM [13], CNN-LSTM [14] and DRNN [15].

On this basis, some models use attention mechanisms to allow the model to focus on the key information contained in the text. For example, the HAN model [16] adopts a hierarchical attention mechanism to divide the text into two levels of sentences and words and uses bidirectional RNN as an encoder. DCCNN [17] first uses a multilayer CNN to capture representations of different characteristics of the n-gram, then uses the attention mechanism to obtain representations by selecting the most important characteristic. MEAN [18] tries to alleviate the problem by integrating three types of linguistic knowledge of feelings into the deep neural network via attention mechanisms. DiSAN [19] is a new attention mechanism in which the attention between the elements of the input sequences is directional and multidimensional. Some models also use the attention mechanism as the primary means. For example, Bi-BloSAN [20] offers the Block self-attention mechanism as an encoder for text and uses the network of doors to extract functionality. All these models use the attention mechanism to select a more important characteristic, which corresponds more to the mode of observation of people than traditional max carpooling and pooling of averages.

In addition, prior training on linguistic models has proven to be effective in learning universal language representations by using large amounts of unlabelled data. Elmo [21], GPT [22], ULMFiT [23] and BERT [24] are among the most remarkable examples. These are neural network language models formed from textual data using unsupervised objectives. For example, BERT is based on a bidirectional multi-layer transformer and is trained in plain text for the prediction of hidden words and the tasks of prediction of the following sentences. To apply a pre-trained model to specific tasks, we need to refine those using task-specific training data and design additional task-specific layers after the pre-training module. For example, to perform text categorization tasks, BERT adds a simple software layer after the pre-formed model and can be refined in this way to create advanced models for the text classification tasks of certain datasets.
BERT model works well in text classification tasks because of its language comprehension capabilities. To solve the problem of classification of educational and vocational guidance questions, we adopted BERT model to use this classification to automatically generate educational and vocational orientation questionnaires according to the four classes that are based on Holland's model and theory and its RIASEC typology. In this article, we propose BERT model for text classification. Our model can more effectively obtain the class of the proposed question.

This article is organized as follows. The first part is devoted to related studies. The second part is dedicated to the method used. The experiment and the results obtained are presented in the third and the last part respectively. Finally, we provide a conclusion according to the perspectives of this research.

2. RELATED STUDIES

2.1 The RIASEC Test

The RIASEC Test or "HOLLAND test", developed by psychologist John HOLLAND, is a theory on careers and vocational choices based on 6 types of people at work: "Realistic", "Investigative", "Artistic "," Social "," Enterprising t "," Conventional ".

According to HOLLAND, people of the same "type" often do the same kind of work, because they are related by their personality, because they pursue similar objectives, because they have the same physical dispositions and psychological about their work.

The typology of a person is established by measuring his degree of affinity with each of the six types to classify him in order of importance, from the most marked type to the least marked type.

For most people, it is mainly the first two or three types of their personal hierarchy that have a significant influence on their way of being and acting, both in their personal and professional lives.

Figure 1: Schema of RIASEC typology

Realistic: he is resourceful and pragmatic. The realist likes the concrete, the terrain.

Investigative: he is an intellectual, he is curious and analytical. The investigator likes to solve problems, needs to understand. He is comfortable with theoretical knowledge.

Artistic: he acts in relation to what, he feels, he has intuition. The artist likes to create, is non-conformist. He needs a different work environment and learns by experimentation.

Social: he is benevolent, warm. He enjoys relationships and needs to communicate, to teach, to help. The social likes to learn through teamwork and collaboration.

Enterprising: he is ambitious, convincing. He likes to lead, sell, motivate. He is comfortable in a competitive environment and learns on the job.

Conventional: it is orderly, meticulous. He is structured, procedural, loves office work. The conventional needs instructions and methods to learn.

2.2 Deep Neural Network

Recently, deep neural networks have achieved good results in natural language processing. Recurrent neural networks (RNN), including short-term long-term memory (LSTM) and recurrent gated units (GRU), are ideal for processing word sequences. Several variants are also proposed, such as Tree-LSTM [11] and TG-LSTM [12]. CNN is also one of the most popular deep neural networks. VDCNN [9] is trying to build a deeper CNN for text classification. [10] Adopted several filters with different window sizes to extract convoluted characteristics at several scales for text classification. DCNN [25] uses a dynamic k-max pool mechanism. DPCNN [26] aims to deepen CNNs without considerably increasing computation costs. [27] Presents a new weight initialization method to improve CNNs for text classification. LK-MTL [28] is a multitasking convolutional neural network with the Leaky unit, which has memory and a forgetting mechanism to filter the flow of functionality between tasks. Unlike the above methods, char-CNN [7] is a character-level model that encodes the characters of the input text. [29],[30],[31].

Naturally, some methods attempt to combine the CNN and RNN models to obtain the advantages of both. C-LSTM [13] first uses CNN to capture local text information, and then uses the LSTM network to encode each output from the convolution kernel to capture global information. CNN-RNN [14] also uses a similar structure, but the difference between the two models is that the CNN layer is connected differently to the RNN layer. DRNN [15] uses the RNN unit to replace the convolution kernel. It uses CNN structure and RNN coding.

2.3 Pre-training Model

Recently, as in the case of computer vision research, the pre-training model has obtained very good results in several natural language processing tasks. They generally learn universal language representations by utilizing large quantities of untagged data and adopt additional task-specific layers after the preform module for different tasks. Elmo [21] is devoted to the extraction of contextual entities from a language model. It allows you to take stock of several benchmarks in NLP, in particular, the answer to the question [32], sentiment analysis [1] and named entity recognition [33]. GPT [22] and ULMFiT [23] pre-train a model architecture on an LM objective before adjusting this same model for a supervised downstream task such as text classification. In addition, we quote the BERT method [24].

3. THE PROPOSED METHOD

The BERT algorithm is built on revolutionary techniques such as the models and transformers seq2seq (sequence to sequence). The seq2seq model is a network that converts a given sequence of words into a different sequence and can link words that seem more important. The LSTM network is a good example of the seq2seq model. The transformer architecture is also responsible for transforming one sequence into another, but without depending on recurring networks such as LSTM or GRU.
In our classification model, we used BERT architecture [24], which is based on a bidirectional multi-layer transformer and is trained in plain text for the prediction of masked words and the prediction of the following sentence.

Figure one shows the structure of BERT when it performs text categorization tasks.

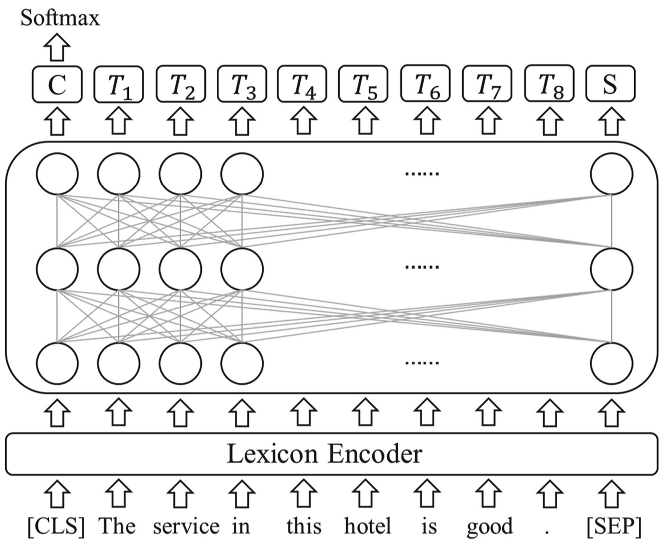

Figure 2: Schema of BERT model

In the entry, [CLS] and [SEP] are the start and end marks of a sentence. The lexicon encoder generates the sum of token incorporations, segmentation incorporations and position incorporations. Thanks to the multilayer self-attention mechanism (transformer encoder) in the box, the corresponding output value will be assigned to each input token. Among them is the representation of the whole text because it gets the information from all the words. Finally, we enter C in the Softmax layer to obtain the classification results. C pays attention to the importance of each word in the text and each word is equal and independent. However, he does not pay attention to the information contained in certain fragments or sentences of the text.

4. EXPEREMINT AND RESULTS

4.1 Dataset and Features

Our dataset was collected from the RIASEC test based on Holland's theory [35], [36], [37], [28], [32], [33]. [34] It contains two columns namely:

Question: It contains questions and statements that measure either the occupations or the activities or abilities or the personality of the users.

Categories: we have four classes (labels) of categories; Activity (0), Occupations (1), Abilities (2), Personality (3)

4.2 Experiment Steps

The experimental steps are explained as follow. At first, to implement BERT or use it for inference, certain conditions must be met. BERT expects data in a specific format and data sets are generally structured to have the following four characteristics:

guide: unique identifier representing an observation.
text_a: The text that we have to classify in given categories.
text_b: It is used when we form a model to understand the relationship between sentences and it does not apply to classification problems.

label: it is made up of labels or classes or categories to which a given text belongs.

In our dataset, we have text_a and label. Let's convert this to the format required by BERT. The next step is to create the objects for each of the above features for all of the records in our dataset using the InputExample class provided in the BERT library.

We now have an appropriate format for our BERT model and we can begin data pre-processing.
We will do the following:
• Normalize the text by converting all characters from spaces to spaces and putting the alphabets according to the type of model used (Cased or uncased).
• Tokenize the text or divide the sentence into words and separate all the punctuation characters from the text.
• Add CLS and SEP tokens to distinguish the beginning and the end of a sentence.
• Break words into WordPieces based on similarity (i.e. "call" -> ["call", "## ing"])
• Match the words of the text to indexes using the vocabulary specific to BERT, which is saved, in the vocab.txt file of BERT.

All of the above operations are effortlessly managed by BERT's tokenization package.
In the output, we have an original sentence from the training set. Then the sentence tokens are printed. Entry IDs are token IDs, each ID representing a unique token. The input masks make it possible to distinguish the tokens from the filling elements. In the experiment, the 0 represents the filling elements. The specified sequence length determines the filling. If the token length is less than the specified sequence length, the tokenized will do the filling to respect the length of the sequence. Segment IDs are used to distinguish different sentences. We only have one segment of text, so all the IDs of the segment are the identical. If two sentences are to be processed, each word in the first sentence will be hidden at 0 and each word in the second sentence will be hidden at 1.
When the entry is ready, we can now load the BERT model, initiate it with the required parameters and metrics.

4.3 The Results

We used the BERT algorithm to classify the category of academic and professional guidance questions according to Holland's RIASEC typology. The figures below show the results obtained by this model.

We obtained the value 0.955 as an accuracy value; it is a very good value of accuracy.

```
INFO:tensorflow:Starting evaluation at 2019-12-30T21:17:44Z
INFO:tensorflow:Graph was finalized.
INFO:tensorflow:Graph was finalized.
INFO:tensorflow:Restoring parameters from /GD/My Drive/Colab Notebooks/BERT/bert_orientation_category/model.ckpt-149
INFO:tensorflow:Restoring parameters from /GD/My Drive/Colab Notebooks/BERT/bert_orientation_category/model.ckpt-149
INFO:tensorflow:Running local_init_op.
INFO:tensorflow:Running local_init_op.
INFO:tensorflow:Done running local_init_op.
INFO:tensorflow:Done running local_init_op.
INFO:tensorflow:Finished evaluation at 2019-12-30-21:19:21
INFO:tensorflow:Finished evaluation at 2019-12-30-21:19:21
INFO:tensorflow:Saving dict for global step 149: eval_accuracy = 0.95555556, false_negatives = 2.0, false_positives = 0.0,
INFO:tensorflow:Saving dict for global step 149: eval_accuracy = 0.95555556, false_negatives = 2.0, false_positives = 0.0,
INFO:tensorflow:Saving 'checkpoint_path' summary for global step 149: /GD/My Drive/Colab Notebooks/BERT/bert_orientation_c
INFO:tensorflow:Saving 'checkpoint_path' summary for global step 149: /GD/My Drive/Colab Notebooks/BERT/bert_orientation_c
{'eval_accuracy': 0.95555556,
 'false_negatives': 2.0,
 'false_positives': 0.0,
 'global_step': 149,
 'loss': 0.2783026,
 'true_negatives': 14.0,
 'true_positives': 29.0}
```

Figure 3: The evaluation indicator of our model

```
tests
```

```
[('parler beaucoup',
  array([-0.23553367, -1.5852612 , -6.159671  , -5.8626537 ], dtype=float32),
  0,
  'Activite'),
 ('Ma capacité à expliquer les choses clairement est :',
  array([-7.064723e+00, -6.829784e+00, -3.227147e-03, -6.656129e+00],
        dtype=float32),
  2,
  'Aptitudes'),
 ('J'aime la précision dans tout ce que je fais.',
  array([-6.8608012e+00, -6.8656878e+00, -6.8063369e+00, -3.2029063e-03],
        dtype=float32),
  3,
  'Personnalite'),
 (' Orthophoniste  orthopédagogue (correction des troubles de l'apprentissage) ',
  array([-6.4367089e+00, -4.3249642e-03, -6.5254297e+00, -6.6859336e+00],
        dtype=float32),
  1,
  'Occupations')]
```

Figure 4: The test result of our classification model

```
predictions[0]
```

```
('Planter  entretenir des arbres  des arbustes  des fleurs ou cultiver le sol.',
 array([-3.6433050e-03, -6.5696979e+00, -7.0253077e+00, -6.6110625e+00],
       dtype=float32),
 0,
 'Activite')
```

Figure 5: Prediction of question categories

5. CONCLUSION

In this article, we propose a new classification model called BERT. The purpose is to obtain the classification of educational and vocational guidance questions based on the Holland typology. Besides, we obtained good results at the level of the classification test of our model. To improve our model well, we recommend working on another model, which combines the BERT method with CNN neural networks. In our previous work, we have already done an automatic classification of these questions using neural networks [37][38][39];

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

REFERENCES

[1] Maas, A.L., et al.:" **Learning word vectors for sentiment analysis**. "In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, vol. 1. Association for Computational Linguistics (2011).

[2] Hourrane, Oumaima, et al. "**Using Deep Learning Word Embeddings for Citations Similarity in Academic Papers.**" International Conference on Big Data, Cloud and Applications. Springer, Cham, 2018. https://doi.org/10.1007/978-3-319-96292-4_15

[3] Hourrane, Oumaima, and El Habib Benlahmar. "**Survey of plagiarism detection approaches and big data techniques related to plagiarism candidate retrieval**." Proceedings of the 2nd International Conference on Big Data, Cloud and Applications. ACM, 2017.

[4] Oumaima Hourrane and El Habib Benlahmer, "**Rich Style Embedding for Intrinsic Plagiarism Detection**" International Journal of Advanced Computer Science and Applications(IJACSA), 10(11), 2019. http://dx.doi.org/10.14569/IJACSA.2019.0101185

[5] Zhang, D., Lee, W.S.: '**Question classification using support vector machines**.' In: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM (2003).

[6] Wang, S., Manning, C.D.: '**Baselines and bigrams: simple, good sentiment and topic classification**.' In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers, vol. 2. Association for Computational Linguistics.

[7] Zhang, X., Zhao, J., LeCun, Y.: '**Character-level convolutional networks for text classification.**' In: Advances in Neural Information Processing Systems (2015).

[8] Chung, J., et al. '**Empirical evaluation of gated recurrent neural networks on sequence modeling**.' arXiv preprint arXiv:1412.3555 (2014).

[9] Conneau, A., et al." **Very deep convolutional networks for text classification**." arXiv preprint arXiv:1606.01781 (2016). https://doi.org/10.18653/v1/E17-1104

[10] Kim, Y."**Convolutional neural networks for sentence classification.** ",arXiv preprint arXiv:1408.5882 (2014).

[11] Tai, K.S., Socher, R., Manning, C.D."Improved semantic representations from tree-structured long short-term memory networks.", arXiv preprint arXiv:1503.00075 (2015).

[12] Huang, M., Qian, Q., Zhu, X." **Encoding syntactic knowledge in neural networks for sentiment classification.** ",ACM Trans. Inf. Syst. (TOIS) 35(3), 26 (2017).

[13] Zhou, C., et al." **A-C-LSTM neural network for text classification**.", arXiv preprint arXiv:1511.08630 (2015).

[14] Xiao, Y., Cho, K." **Efficient character-level document classification by combining convolution and recurrent layers**.", arXiv preprint arXiv:1602.00367 (2016).

[15] Wang, B." **Disconnected recurrent neural networks for text categorization**.", In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Long Papers, vol. 1 (2018).

[16] Yang, Z., et al."**Hierarchical attention networks for document classification.** ", In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (2016). https://doi.org/10.18653/v1/N16-1174

[17] Wang, S., Huang, M., Deng, Z."**Densely connected CNN with multi-scale feature attention for text classification.** ", In IJCAI (2018).

[18] Lei, Z., et al."**A multi-sentiment-resource enhanced attention network for sentiment classification**. ",arXiv preprint arXiv:1807.04990 (2018). https://doi.org/10.18653/v1/P18-2120

[19] Shen, T., et al." **DiSAN: directional self-attention network for RNN/CNN-free language understanding**. ", In Thirty-Second AAAI Conference on Artificial Intelligence (2018).

[20] Shen, T., et al."**Bi-directional block self-attention for fast and memory-efficient sequence modeling.** ",arXiv preprint arXiv:1804.00857 (2018).

[21] Peters, M.E., et al."**Deep contextualized word representations**.", arXiv preprint arXiv:1802.05365 (2018).

[22] Radford, A., et al." **Improving language understanding by generative pre-training**", (2018). https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/langu ageunsupervised/languageunderstandingpaper.pdf.

[23] Howard, J., Ruder, S." **Universal language model fine-tuning for text classification**.", arXiv preprint arXiv:1801.06146 (2018).

[24] Devlin, J., et al." **Bert: pre-training of deep bidirectional transformers for language understanding**.", arXiv preprint arXiv:1810.04805 (2018).

[25] Kalchbrenner, N., Grefenstette, E., Blunsom, P."**A convolutional neural network for modelling sentences**.", arXiv preprint arXiv:1404.2188 (2014).

[26] Johnson, R., Zhang, T."**Deep pyramid convolutional neural networks for text categorization**.", In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Long Papers, vol. 1 (2017).
https://doi.org/10.18653/v1/P17-1052

[27] Li, S., et al." **Initializing convolutional filters with semantic features for text classification**.", In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (2017).

[28] Xiao, L., et al."**Learning what to share: leaky multi-task network for text classification**.", In: Proceedings of the 27th International Conference on Computational Linguistics (2018).

[29] Adil Alharthi,Nouf Alzahrani,Ikram Moualla" **Convolutional Neural Network based on Transfer Learning for Medical Forms Classification.",** In: International Journal of Advanced Trends in Computer Science and Engineering, Volume 8, No.6, November – December 2019, Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse115862019.pdf
https://doi.org/10.30534/ijatcse/2019/115862019

[30] Mohammed Y.Alzahrani,Ahmed H.Alahmadi" **Breast Cancer Image Classification Using the Convolution Neural Network.",** In: International Journal of Advanced Trends in Computer Science and Engineering, Volume 8, No.6, November – December 2019, Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse120862019.pdf
https://doi.org/10.30534/ijatcse/2019/120862019

[31] Ali Mohammad Alqudah,Hiam Alquraan,Isam Abu Qasmieh,Amin Alqudah,Wafaa Al-Sharu" **Brain Tumor Classification Using Deep Learning Technique - A Comparison between Cropped, Uncropped, and Segmented Lesion Images with Different Sizes**.", In: International Journal of Advanced Trends in Computer Science and Engineering, Volume 8, No.6, November – December 2019, Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse155862019.pdf
https://doi.org/10.30534/ijatcse/2019/155862019

[32] Rajpurkar, P., et al." **SQuAD: 100,000+ questions for machine comprehension of text.** ",arXiv preprint arXiv:1606.05250 (2016).

[33] Sang, E.F., De Meulder, F." **Introduction to the CoNLL-2003 shared task: languageindependent named entity recognition**.", arXiv preprint cs/0306050 (2003).

[34] Omar Zahour, El Habib Benlahmar, Ahmed Eddaoui " **E-orientation : entre prescription des théories et prise de décision** " Conference TIM'16,2016.

[35] Omar Zahour, El Habib Benlahmar,Ahmed Eddaoui " **E-Orientation: Between prescription of theories and decision-making** " Conference SITA'16, October 2016.

[36] Omar Zahour, El Habib Benlahmar, Ahmed Eddaoui "**E-orientation : Vers une modélisation des facteurs d'orientation scolaire** " Conference TIM'18, 2018.

[37] Omar Zahour, El Habib Benlahmar, Ahmed Eddaoui, Oumaima Hourrane "**Automatic Classification of Academic and Vocational Guidance Questions using Multiclass Neural Network** '' January 2019, International Journal of Advanced Computer Science and Applications 10(10); DOI: 10.14569/IJACSA.2019.0101072

[38] Hourrane, Oumaima, and El Habib Benlahmar. "**Survey of plagiarism detection approaches and big data techniques related to plagiarism candidate retrieval**." Proceedings of the 2nd International Conference on Big Data, Cloud and Applications. ACM, 2017.

[39] Hourrane, Oumaima, et al. "**Using Deep Learning Word Embeddings for Citations Similarity in Academic Papers."** International Conference on Big Data, Cloud and Applications. Springer, Cham, 2018.
https://doi.org/10.1007/978-3-319-96292-4_15