Available online at https://www.asjp.cerist.dz/en/PresentationRevue/134

RIST

Information Processing at the Digital Age Journal

CERIST Natural Language Processing Challenge

March 29th, 2023.

Compact CNN-Based Architecture for Text Classification and Sentiment Analysis

Zoubir TALAI*, Nada KHERICI

University of Badji Mokhtar, Annaba, 23000, Algeria

Abstract

In the last decade, social media and internet involvement in people's life raised new challenges that modern AI needs to deal with. Textual data is generated every time an article is published or an online post is shared or even a simple comment is made. Among these challenges, we find text classification which is used to identify the general meaning of a set of words using AI methods. This paper presents our participation to the CERIST Natural Language Processing Challenge, where we proposed a simple yet effective convolutional neural network architecture that can be used for text classification and sentiment analysis. We tested our proposition on 5 different tweets datasets, Hate Speech, Fake News, Arabic Covid Sentiment, Arabic Sentiment, and English Sentiment, and obtained respectively 99,85%, 99,86%, 99,58%, 97,97%, 95,65% accuracy on the training subset and 98,43%, 94,74%, 87,53%, 54,90%, 60,62% accuracy on the validation subset.

Keywords: Text Classification; Sentiment Analysis; Deep Learning; Convolutional Neural Network.

* Corresponding author.
E-mail address: talai_zoubir@yahoo.com.

1. Introduction

With the explosion of the internet, the volume of the collected textual data from social media and other similar resources is significantly increasing. Tweets and comments can be short, and casual, with a lot of abbreviations, slang, etc. Extracting valuable insights from these comments can be challenging and time-consuming.

Natural Language Processing (NLP) includes methods that are used to extract knowledge from the textual data written by human beings. In addition to data volume, complexity is raised due to the unstructured nature of the text.

Text classification is one of the main tasks in natural language processing. It is the process of automatically classifying a textual document with the most relevant predefined labels.

Due to the lack of free Arabic datasets, Arabic text classification is more difficult. Therefore, authors in, Elnagar et al., 2020, introduced new rich datasets SANAD and NADiA for Arabic text classification tasks. They used and compared several deep learning (DL) models. Their best-achieved accuracy on the SANAD corpus was 96.94% by an attention-GRU model. As for NADiA, the attention-GRU achieved the highest overall accuracy of 88.68%.

Other datasets were investigated in, Ababneh, 2022, where authors trained and tested different models for text classification. They experimented with SANAD, Khaleej, Arabiya, Akhbarona, KALIMAT, Waten2004, and Khaleej2004 datasets. They tried different models and achieved 82% accuracy for the best one.

Aiming to improve classification accuracy, two different textual representations were explored, in Alzanin et al., 2022, word embedding using Word2Vec and stemmed text with term frequency-inverse document frequency (tf-idf). The authors tested three different classifiers on a collected and manually annotated dataset. Their result tops the current state-of-the-art score using a deep learning approach (RNN-GRU).

In AlAjlan, 2021, authors developed a method using a Convolutional Neural Network (CNN) that aims to detect any threat in images or Arabic comments which were shared. The collected data employing the Instagram API was manually labeled. The results showed that the accuracy of the developed model reached 99% for comment classification.

Authors in, Al-Hassan and Al-Dossari, 2021, aimed to identify and classify Arabic tweets using SVM compared to 4 deep learning models. Their dataset of 11 K tweets was collected and manually labeled. The results show that all 4 deep learning models outperform the SVM model in detecting hateful tweets.

For more details on advances done in text classification using deep learning techniques, great work was done by a group of researchers in Lavanya and Sasikala, 2021, Kowsari et al., 2019, Li et al., 2020.

In this context, this study aims to propose a single CNN architecture for both text classification and sentiment analysis that can be used on different datasets. The proposed architecture will be used on 5 different datasets that contain Arabic and English tweets.

2. Dataset

To validate our proposed architecture, we choose to work with 5 different datasets that contain tweets in Arabic and English languages. These datasets were provided by the organizers of the CERIST NLP challenge Hadj Ameur and Aliane, 2021.

We divided each dataset into two subsets: the first one is composed of 75% of the tweets that will be used for training the models and the remaining 25% will be dedicated to validation. Table 1 shows a summary of the used datasets.

Table 1. Datasets summary

Name	Size (Nb. Tweets)	Labels
Hate Speech	8662	Hateful / Not Hateful
Fake News	8661	Yes / No / Maybe / Can't Decide
Arabic Covid sentiment	4128	Neutral / Positive / Negative
Arabic sentiment	3343	Neutral / Positive / Negative
English sentiment	16173	Neutral / Positive / Negative

3. Proposed Architecture

To solve text classification efficiently, we saw that CNN convey the most for this task. Word embedding and context can be learned alongside the training of the neural network. We simply add an embedding layer as input for the proposed CNN. This technique is based on the Word2Vec algorithm and represents each word of the vocabulary with a specific dense vector for which a chosen size is fixed and called embedding dimension, Mikolov et al., 2013. When dealing with large datasets, a limited number of the most important words are considered for the training process.

The next layer consists of a 1D convolution where text features are extracted using multiple fixed-size kernels. The results of this operation are passed by a global max-pooling layer that performs a size reduction ensuring that the most critical information is extracted from the convolution results, it is also used to flattens the data so it can be fed to the dense layer. At this stage, it is a traditional neural network with fully connected layers. Figure 1 illustrates the used architecture.

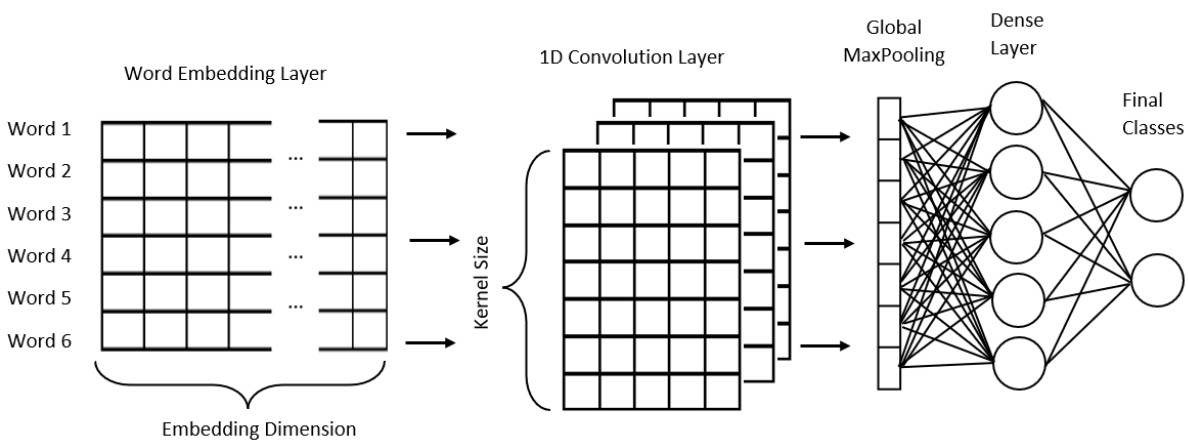


Figure 1. Used architecture for Text Classification

4. Results and Discussion

The proposed architecture was used to train a model for every dataset. For each training, data was passed in batches of 128 for 50 epochs, and the learning process was early stopped if no progress was made. Figure 2. Shows the evolution of the training process on each dataset.

After several experimentations, we noticed that the network achieved the best results using only the first 10000 words in each dataset, we also found that the optimal embedding dimension is 64.

For the convolution filters, a kernel of size 7 was used to extract text features. The role of these filters is to learn the context meaning between words. We used 32 filters which is enough for the used dataset. The dropout layer helps the learning process by randomly deactivating 30% of neurons during the training phase to avoid overfitting the training set.

The global max pooling layer is used to reduce feature dimensions and feed the output to the fully connected layer. This last one is kept simple and small, we used only 10 neurons for the hidden layer and the last layer was modified according to every dataset. Table 2 shows the summary of the used model.

Table 2. Proposed architecture

	Layer	Labels
1	Embedding (features=10000, embedding dimension=64)	/
2	1D Convolution (filters= 32, kernel size=7)	Relu
3	Dropout (0.3)	/
4	Global Max Pooling	/
5	Dense (10)	Relu
6	Dense(N)	Softmax

On the Hate Speech dataset, the network learns quickly and the loss drops as well as the accuracy goes up, which means that the model generalizes well. The same behavior is observed when training a model with the Fake News dataset.

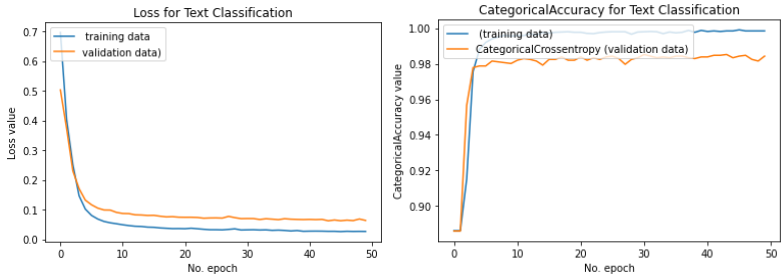
In the case of the Arabic Covid Sentiment dataset, the model seems to struggle when it reaches epoch N°5, the accuracy no longer increases neither the loss decreases. In this situation, the model may be overfitting the training data and need to be stopped as soon as it stops evolving. The same issue is visible when training with the last two datasets, Arabic and English Sentiment Analysis. The training curve indicates an overfitting problem.

Table 3 shows a brief comparison between the results obtained using the same architecture defined in Table 2 with the 5 datasets presented earlier.

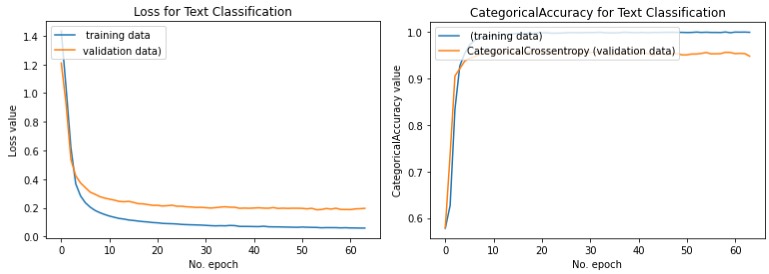
Table 3. Obtained results

Dataset	Accuracy achieved on training dataset	Accuracy achieved on validation dataset
Hate Speech	99.85%	98.43%
Fake News	99.86%	94.74%
Arabic Covid sentiment	99.58%	87.53%
Arabic sentiment	97.97%	54.90%
English sentiment	95.65%	60.62%

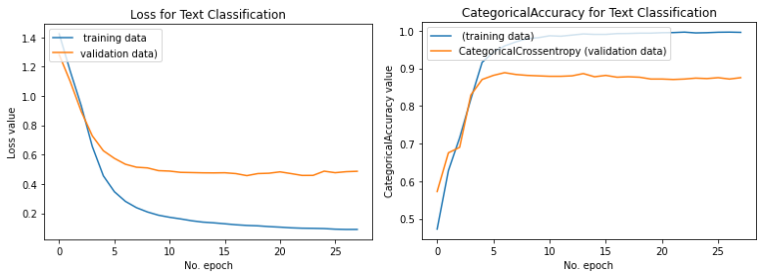
(A)



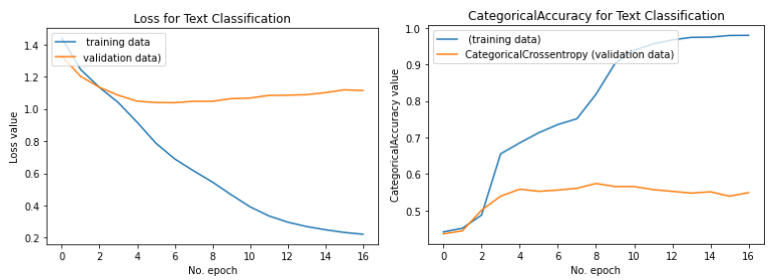
(B)



(C)



(D)



(E)

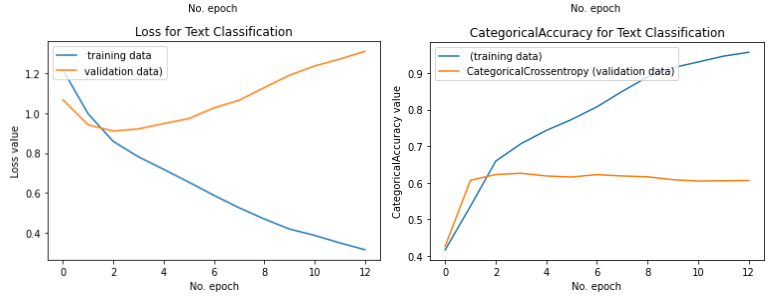


Figure 2. Loss and Accuracy for validation Subset: A: Hate Speech Dataset, B: Fake News Dataset, C: Covid sentiment Dataset, D: Arabic sentiment Dataset, E: English sentiment Dataset

The above results imply that the proposed architecture is indeed effective for text classification and sentiment analysis tasks. The poor performance that was noticed with Arabic and English sentiment datasets can be improved either by expanding these datasets or fine-tuning the used parameters.

5. Conclusion

Text classification and sentiment analysis are challenging tasks that modern AI deals with every time a text is published or an online post is shared or even a simple comment is made. Rising to such challenges needs methods and techniques that can do the job efficiently and quickly as possible. The proposed CNN architecture proved to be accurate and efficient as needed. It achieved excellent results on training datasets as well as on validation datasets for all of the Hate Speech, Fake News and Covid Sentiment datasets. The struggle that the models face training with Arabic and English sentiment datasets indicates that more records must be added to achieve better results.

References

Ababneh, A. H., 2022. Investigating the relevance of Arabic text classification datasets based on supervised learning. Journal of Electronic Science and Technology, 20(2), 100160.

AlAjlan, S. A., & Saudagar, A. K. J., 2021. Machine learning approach for threat detection on social media posts containing Arabic text. Evolutionary Intelligence, 14(2), 811-822.

Al-Hassan, A., & Al-Dossari, H., 2022. Detection of hate speech in Arabic tweets using deep learning. Multimedia systems, 28(6), 1963-1974.

Alzanin, S. M., Azmi, A. M., & Aboalsamh, H. A., 2022. Short text classification for Arabic social media tweets. Journal of King Saud University-Computer and Information Sciences, 34(9), 6595-6604.

Elnagar, A., Al-Debsi, R., & Einea, O., 2020. Arabic text classification using deep learning models. Information Processing & Management, 57(1), 102121.

Hadj ameur M.S., Aliane H., 2021. AraCOVID19-MFH: Arabic COVID-19 multi-label fake news & hate speech detection dataset. Procedia Computer Science 189, 232-241.

Kowsari, K., Jafari Meimandi, K., Heidarysafa, M., Mendu, S., Barnes, L., & Brown, D., 2019. Text classification algorithms: A survey. Information, 10(4), 150.

Lavanya, P. M., & Sasikala, E., 2021. Deep learning techniques on text classification using Natural language processing (NLP) in social healthcare network: A comprehensive survey. In 2021 3rd international conference on signal processing and communication (ICPSC) (pp. 603-609). IEEE.

Li, Q., Peng, H., Li, J., Xia, C., Yang, R., Sun, L., Yu, P. S., & He, L., 2020. A survey on text classification: From shallow to deep learning. arXiv preprint arXiv:2008.00364.

Mikolov, T., Chen, K., Corrado, G., & Dean, J., 2013. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.