Ministry of Higher Education and Research
Higher School of Computer Science 08 May 1945 - Sidi Bel Abbes

Second Year Second Cycle - Artificial Intelligence and Data Science

LAB 02 & 03 : FEATURE EXTRACTION & WORD EMBEDDING

Presented By : FELLAH Abdelnour.

Date: April 26, 2024

1 INTRODUCTION

In this report we will explore and discuss the results of different feature extraction and word embedding techniques on the task of text classification more specifically,**author recognition** task on the **spooky dataset**,we will as well discuss our attempts to improve the obtained results.

2 LAB 02 : FEATURE EXTRACTION & WORD EMBEDDING

In this section of the report we're going to go through all what we tried in lab 02 and discuss the results we've got.

2.1 TEXT PREPROCESSING

For the preprossessing of the dataset we applies the following normalization steps :

- Removing repetitive characters and misspelled words.
- Normalizing unicode characters.
- handle special entries (emails,html tags and urls).
- Captilization : the text was transformed to lower case.
- Removing punctuations.
- Stop words removal.
- Stemming.

2.2 TOKENIZATION & VECTORIZATION TECHNIQUES

We tried all the combinations of the following tokenization techniques :

- Space based.
- Rule based.
- Word piece.

and the following vectorization methods :

- Bag of words.
- Tf-Idf.
- Binary Bag of words.

2.3 RESULTS OF DIFFERENT TOKENIZATION & VECTORIZATION METHODS

To compare the different preprocessing techniques,we trained a Multi layer perceptron with one hidden layer of size 8 and a **relu** activation function using Adam optimizer with learning rate equals to 0.01 and weight decay equals to 0.5, the results of the validation set are presented in the following table :

Tokenization	Vectorization	Accuracy	F1 score	Precision	Recall
Space Based	Bag of words	0.794	0.792	0.798	0.788
Space Based	Tf-Idf	**0.807**	**0.805**	**0.811**	**0.801**
Space Based	Binary Bag of words	0.791	0.789	0.796	0.786
Word Piece	Bag of words	0.769	0.768	0.774	0.764
Word Piece	Tf-Idf	0.777	0.775	0.788	0.768
Word Piece	Binary Bag of words	0.775	0.773	0.778	0.770
Rule based	Bag of words	0.791	0.789	0.799	0.783
Rule based	Tf-Idf	0.799	0.798	0.800	0.798
Rule based	Binary Bag of words	0.789	0.786	0.798	0.780

Table 1: The results of applying different tokenization and vectorization techniques

We notice that Tf-Idf with space based tokenization gave the best results across all the metrics.

2.4 RESULTS OF WORD EMBEDDING TECHNIQUES

In this section we will presents the results of four different word embedding techniques that were used to calculate a sentence embedding for each document in the dataset then this embedding were used to train an MLP classifier with a one hidden layer of size 32 and relu activation function

Method	Accuracy	F1 score	Precision	Recall
Continuous Bag Of Words	0.536	0.531	0.532	0.530
Skip n-grams	**0.694**	**0.694**	**0.694**	**0.694**
Glove	0.403	0.191	0.134	0.333
Fast Text	0.595	0.584	0.600	0.581

Table 2: The results of using different word embedding methods

The results are generally very poor,but skip–gram gave better results than the rest of the word embedding methods.

3 LAB 03 : IMPROVING THE RESULTS

To improve the results we kept the same preprocessing steps and used space based tokenization,however instead of relaying on ready-to-use embedding vectors or some feature extraction technique we used an embedding layer to learn the representation of the words in our vocabulary as we're training the model, also the a dropout layer was added after the mean layer,and its value was tuned manually.

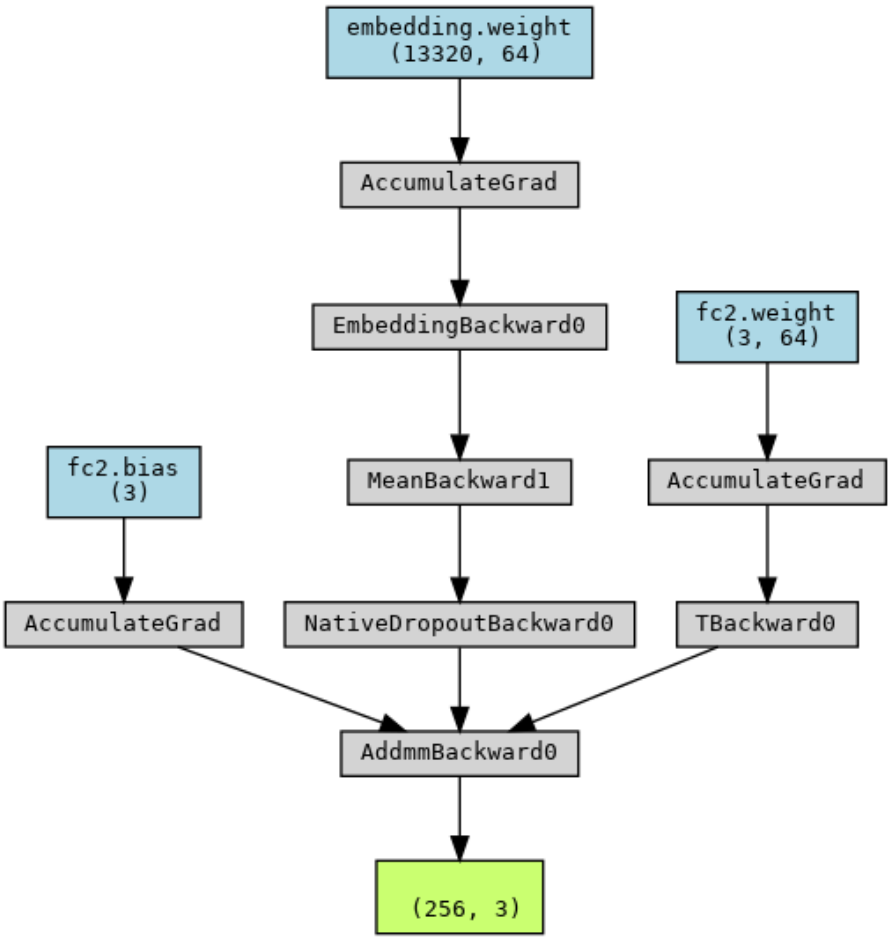


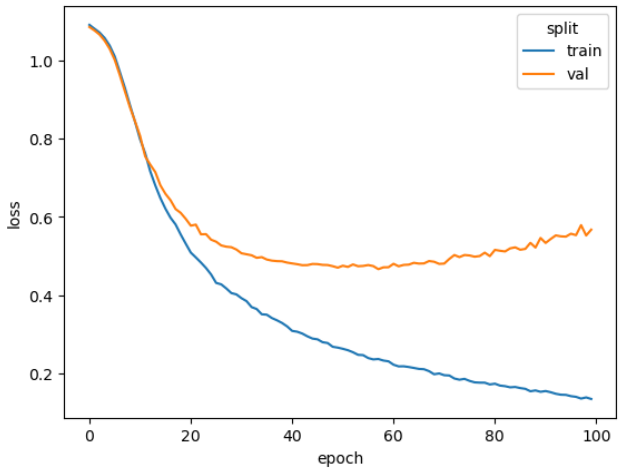
Figure 1: The model's visual representation

3.1 Model's parameters

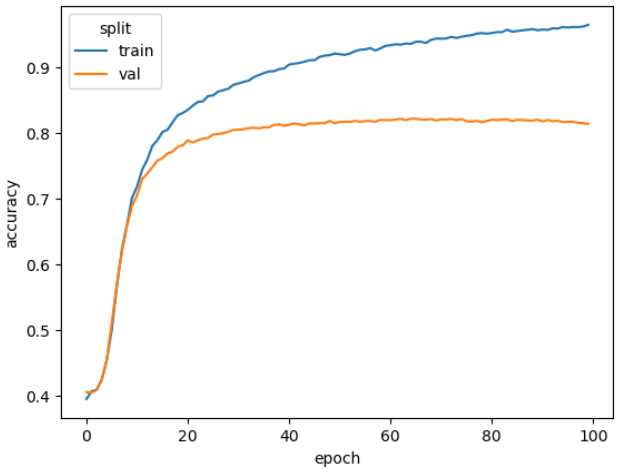
Learning rate	0.001
Epochs	100
Embedding dimension	64
Batch size	256
Dropout rate	0.35

Table 3: Hyper-parameters

3.2 Learning graphs



(a) Loss graph



(b) Accuracy graph

3.3 Results

The values of the different metrics using the weights on the epoch with best validation accuracy gave the following results :

Accuracy	F1 score	Precision	Recall
0.822	0.821	0.826	0.818

Table 4: The results of new architecture.

We notice that the results for all the metrics are better than the results obtained in the previous lab.

4 Conclusion

In this tow labs we explored the main feature extraction and word embedding techniques as well as the importance of experimenting with them,and how can a deep learning model act as a feature extractor and classifier at the same time and how can these problem-oriented features help in improving the performance on a given task.