Text Analytics: 2nd Assignment

Tsirmpas Dimitris
Drouzas Vasilis

February 11, 2024

Athens University of Economics and Business
MSc in Data Science

Contents

1 Introduction

This report will briefly discuss the theoretical background, implementation details and decisions taken for the construction of MLP models for sentiment analysis and POS tagging tasks.

This report and its associated code, analysis and results were conducted by the two authors. Specifically, the sentiment analysis task was performed by Drouzas Vasilis, and the POS-tagging task by Tsirmpas Dimitris. This report was written by both authors.

Note that due to the relative custom code complexity, most of the code used in this section was developed and imported from python source files located in the ‘tasks‘ module. In-depth documentation and implementation details can be found in these files.

2 POS Tagging

POS tagging is a language processing task where words in a given text are assigned specific grammatical categories, such as nouns, verbs, or adjectives. The objective is to analyze sentence structure.

In this section we describe how we can leverage pre-trained word embeddings to create a context-aware MLP classifier.

2.1 Dataset

Acquiring and preprocessing our data with the goal of eventually acquiring a sufficient representation of our text is the most difficult and time-consuming task. We thus split it in distinct phases:

- Original dataset acquisition and parsing
- Qualitative analysis and preprocessing
- Transformation necessary for the NLP task

Each of these distinct steps are individually analyzed below.

2.1.1 Acquisition

We select the English EWT-UD tree, which is the largest currently supported collection for POS tagging tasks for the English language.

This corpus contains 16622 sentences, 251492 tokens and 254820 syntactic words, as well as 926 types of words that contain both letters and punctuation, such as ’s, n’t, e-mail, Mr., ’s, etc). This is markedly a much higher occurrence than its siblings, and therefore may lead to a slightly more difficult task.

The dataset is made available in `conllu` format, which we parse using the recommended `conllu` python library. We create a dataframe for every word and its corresponding POS tag and link words belonging to the same sentences by a unique sentence ID. The data are already split to training, validation and test sets, thus our own sets correspond to the respective split files.

We are interested in the UPOS (Universal Part of Speech) tags for English words.

Set	Mean	Std	Min	25%	50%	75%	Max
Training	16.31	12.4	1	7	14	23	159
Validation	12.56	10.41	1	5	10	17	75
Test	12.08	10.6	1	4	9	17	81

Table 1: Summary and order statistics for the number of words in the sentences of each data split.

Set	Total Word Count	Total Sentence Count
Training	204614	12544
Validation	25152	2001
Test	25096	2077

Table 2: Total text volume of each data split.

2.1.2 Qualitative Analysis

Our training vocabulary is comprised of 16654 words. We include qualitative statistics on the sentences of our dataset in Tables 1 and 2. The splits are explicitly mentioned separately because the splitting was performed by the dataset authors and not by random sampling. We would therefore like to confirm at a glance whether their data are similar.

2.1.3 Preprocessing

Given the nature of our task we can not implement preprocessing steps such as removing punctuation marks, stopwords or augmenting the dataset. Thus, the only meaningful preprocessing at this stage would be converting the words to lowercase. We believe that the context of each word will carry enough information to distinguish its POS tag regardless of case.

Another issue we need to address before continuing is that of words being part of (depending on) other words for their POS tag. Those would be words such as ”don’t”, ”couldn’t” or ”you’re”. In the standard UPOS schema these are defined as two or more separate words, where the first is represented by its standard POS tag, and the rest as part of that tag (UPOS tag ”PART”). For instance, ”don’t” would be split into ”do” and ”n’t” with ”AUX” and ”PART” tags respectively. In our dataset, these words are represented both in the manner described above followed by the full word (”don’t”) tagged with the pseudo-tag ”_”. We remove the latter representation from the working dataset.

The general algorithm to calculate the window embeddings on our dataset can be found in Algorithm 1. The algorithm uses a few external functions which are not described here for the sake of brevity. `get_window()` returns the context of the word inside a sentence, including padding where needed, `embedding()` returns the word embedding for a single word and `concatenate` returns a single element from a list of elements. The rest of the functions should be self-explanatory. Note that this algorithm does not represent the actual python implementation.

Algorithm 1 Window Embedding creation algorithm from raw-text sentences.

Input sentences, window_lim: a list of sentences and an upper bound of windows to be computed

Output tuple(windows, targets): the window embeddings and the POS tag corresponding to the median word of each window

```
windows = list()
targets = list()

for sentence in sentences do
    for word in sentence do
        window = get_window(word, sentence)
        target = get_tag(word)
        windows.add(window)
        targets.add(target)
    end for
end for

window_embeddings = list()
for window in windows do
    if window_embeddings.size ≥ window_lim then
        break
    end if

    word_embeddings = list()
    for word in window do
        if word is PAD_TOKEN then
            word_embeddings.add(zeros(embedding_size))
        else
            word_embeddings.add(embedding(word))
        end if
    end for
    window_embedding = concatenate(word_embeddings)
    window_embeddings.add(window_embedding)
end for

targets_vec = list()
for target in targets do
    targets_vec.add(one_hot(target))
end for

return window_embeddings, targets_vec
```

2.2 Baseline Classifier

We create our own classifier which classifies each token by the majority label associated with it. The classifier is defined as a subclass of sklearn's classifier superclass and thus can seamlessly use it in most sklearn-provided functions such as `classification_report()` and its implementation can be found in the `tasks.models` module.

The results of the classifier can be found in Figure TODO. These results make intuitive sense, since most words in the English language can be classified in a single label, irrespective of context. For example, "is" will always be classified as "AUX", and all punctuation marks will be classified as "PUNCT".

Thus, besides quantitative statistics such as categorical accuracy and f1-score, we should pay close attention to the precision and recall statistics for the more variable POS tags such as "NOUN" or "VERB" in order to properly evaluate our MLP classifier.

2.3 MLP Classifier

2.3.1 Hyper-parameter tuning

We use the `keras_tuner` library to automatically perform random search over various hyper-parameters of our model.

The parameter search consists of:

- The depth of the model (the number of layers)
- The height of the model (the number of parameters by layer)
- The learning rate

The parameter search does NOT consist of:

- Dropout rate, since dropout rarely changes the final result of a neural network, but rather tunes the trade-off between training time and overfit avoidance
- Activation functions, since they rarely significantly influence the model's performance

With this scheme we hope to maximize the area and granularity of our search to the hyper-parameters that are most likely to significantly influence the final results.

We implement early stopping and set a maximum iteration limit of 70. We assume that if a model needs to go over that limit, it may be computationally inefficient, and thus less desirable compared to a slightly worse, but much more efficient model. Additionally, we use a relatively large batch size to improve training times since this operation is very computationally heavy. We don not yet aim to create the best classifier, so slightly suboptimal weights are not a problem for the purposes of the hyperparameter search. We use a relatively very large batch size to improve training times since this operation is very computationally heavy. We don't yet aim to create the best classifier, so slightly suboptimal weights are not a problem for the purposes of the hyperparameter search.

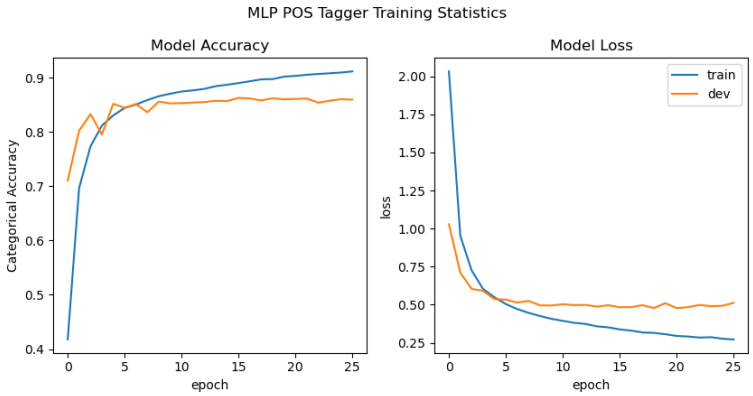

Figure 1: Loss and accuracy on the training and validation sets depending on the number of epochs.

2.3.2 Training

We now re-train our model with a much smaller batch size and keep track of the training history and best weights by validation loss.

Unfortunately, the different batch size means we can not rely on the hyper parameter search to get an estimation of training epochs. Thus, we rely on early-stopping on the validation data to ensure our model does not overfit as a result of training time.

We use the categorical accuracy stopping criterion instead of loss. This may lead to situations where validation loss increases, but so does accuracy [1]. This represents a trade-off between our model being more confidently incorrect about already-misclassified instances, but better at edge cases where the classification is more ambiguous. We previously discussed how the strength of a context-aware classifier lies in these kinds of distinctions, which justifies our choice of favoring correct edge-case classifications in the expense of more confidently incorrect misclassifications.

This phenomenon is demonstrated in Figure 1.

2.3.3 Results

The results of our MLP classifier compared to the baseline model mentioned above can be found in Tables 3, 4 and 5. We include precision, recall and F1 scores for each individual tag, as well as their macro average denoted by the ”MACRO” tag in the tables.

We note an increase in all metrics for our MLP classifier, especially in tags such as ”SCONJ” (subordinating conjunction), and VERB and NOUN (which we hypothesized at the beginning of the report). The only notable exception is the ”X” tag (other) which is attributed to unintelligible material, foreign words and word fragments. This is an acceptable drawback, since these are easily caught by preprocessing or weak classifiers anyway.

Table 3: Results on the training dataset.

model	tag	precision	recall	f1
Baseline	ADJ	1.000	0.886	0.940
Baseline	ADP	1.000	0.826	0.905
Baseline	ADV	1.000	0.841	0.913
Baseline	AUX	1.000	0.924	0.960
Baseline	CCONJ	1.000	0.994	0.997
Baseline	DET	1.000	0.958	0.979
Baseline	INTJ	1.000	0.917	0.957
Baseline	NOUN	1.000	0.908	0.952
Baseline	NUM	1.000	0.985	0.992
Baseline	PART	1.000	0.884	0.938
Baseline	PRON	1.000	0.959	0.979
Baseline	PROPN	1.000	0.894	0.944
Baseline	PUNCT	1.000	0.997	0.999
Baseline	SCONJ	1.000	0.455	0.626
Baseline	SYM	1.000	0.835	0.910
Baseline	VERB	1.000	0.899	0.947
Baseline	X	1.000	0.676	0.806
Baseline	MACRO	0.863	0.873	0.862
MLP	ADJ	1.000	0.961	0.980
MLP	ADP	1.000	0.910	0.953
MLP	ADV	1.000	0.921	0.959
MLP	AUX	1.000	0.970	0.985
MLP	CCONJ	1.000	0.997	0.998
MLP	DET	1.000	0.988	0.994
MLP	INTJ	1.000	0.900	0.947
MLP	NOUN	1.000	0.962	0.980
MLP	NUM	1.000	0.994	0.997
MLP	PART	1.000	0.979	0.989
MLP	PRON	1.000	0.971	0.985
MLP	PROPN	1.000	0.928	0.963
MLP	PUNCT	1.000	0.998	0.999
MLP	SCONJ	1.000	0.712	0.832
MLP	SYM	1.000	0.910	0.953
MLP	VERB	1.000	0.955	0.977
MLP	X	1.000	0.364	0.534
MLP	MACRO	0.936	0.907	0.915

Table 4: Results on the validation dataset.

model	tag	precision	recall	f1
Baseline	ADJ	1.000	0.751	0.858
Baseline	ADP	1.000	0.838	0.912
Baseline	ADV	1.000	0.748	0.856
Baseline	AUX	1.000	0.922	0.960
Baseline	CCONJ	1.000	0.990	0.995
Baseline	DET	1.000	0.951	0.975
Baseline	INTJ	1.000	0.690	0.816
Baseline	NOUN	1.000	0.901	0.948
Baseline	NUM	1.000	0.704	0.827
Baseline	PART	1.000	0.865	0.928
Baseline	PRON	1.000	0.964	0.982
Baseline	PROPN	1.000	0.395	0.566
Baseline	PUNCT	1.000	0.989	0.994
Baseline	SCONJ	1.000	0.453	0.624
Baseline	SYM	1.000	0.812	0.896
Baseline	VERB	1.000	0.746	0.855
Baseline	X	1.000	0.020	0.040
Baseline	MACRO	0.795	0.749	0.759
MLP	ADJ	1.000	0.819	0.901
MLP	ADP	1.000	0.867	0.929
MLP	ADV	1.000	0.726	0.841
MLP	AUX	1.000	0.908	0.952
MLP	CCONJ	1.000	0.973	0.986
MLP	DET	1.000	0.965	0.982
MLP	INTJ	1.000	0.529	0.692
MLP	NOUN	1.000	0.858	0.924
MLP	NUM	1.000	0.887	0.940
MLP	PART	1.000	0.951	0.975
MLP	PRON	1.000	0.932	0.965
MLP	PROPN	1.000	0.666	0.799
MLP	PUNCT	1.000	0.995	0.997
MLP	SCONJ	1.000	0.589	0.742
MLP	SYM	1.000	0.870	0.930
MLP	VERB	1.000	0.831	0.908
MLP	X	1.000	0.122	0.218
MLP	MACRO	0.822	0.793	0.797

Table 5: Results on the test dataset.

model	tag	precision	recall	f1
Baseline	ADJ	1.000	0.754	0.860
Baseline	ADP	1.000	0.826	0.905
Baseline	ADV	1.000	0.810	0.895
Baseline	AUX	1.000	0.910	0.953
Baseline	CCONJ	1.000	0.997	0.998
Baseline	DET	1.000	0.956	0.978
Baseline	INTJ	1.000	0.758	0.862
Baseline	NOUN	1.000	0.897	0.946
Baseline	NUM	1.000	0.584	0.737
Baseline	PART	1.000	0.887	0.940
Baseline	PRON	1.000	0.971	0.985
Baseline	PROPN	1.000	0.430	0.602
Baseline	PUNCT	1.000	0.990	0.995
Baseline	SCONJ	1.000	0.438	0.609
Baseline	SYM	1.000	0.826	0.905
Baseline	VERB	1.000	0.766	0.867
Baseline	X	0.000	0.000	0.000
Baseline	MACRO	0.779	0.753	0.755
MLP	ADJ	1.000	0.829	0.906
MLP	ADP	1.000	0.880	0.936
MLP	ADV	1.000	0.792	0.884
MLP	AUX	1.000	0.915	0.956
MLP	CCONJ	1.000	0.986	0.993
MLP	DET	1.000	0.965	0.982
MLP	INTJ	1.000	0.439	0.610
MLP	NOUN	1.000	0.851	0.919
MLP	NUM	1.000	0.871	0.931
MLP	PART	1.000	0.933	0.966
MLP	PRON	1.000	0.934	0.966
MLP	PROPN	1.000	0.666	0.800
MLP	PUNCT	1.000	0.990	0.995
MLP	SCONJ	1.000	0.609	0.757
MLP	SYM	1.000	0.861	0.925
MLP	VERB	1.000	0.843	0.915
MLP	X	1.000	0.194	0.324
MLP	MACRO	0.819	0.797	0.801

References

[1] Soltius (https://stats.stackexchange.com/users/201218/soltius). *How is it possible that validation loss is increasing while validation accuracy is increasing as well*. Cross Validated. URL:https://stats.stackexchange.com/q/341054 (version: 2023-03-28). eprint: https://stats.stackexchange.com/q/341054. URL: https://stats.stackexchange.com/q/341054.