Network and Practical

Introduction to ML for Natural Language Processing

Daniele Passabì [221229], Data Science

Multi-class Classification of European Laws in Five Languages
How do Statistical Models and Recurrent Neural Networks Behave?

Contents

1 Introduction

Natural Language Processing (NLP) originated in the 1950s as the intersection of artificial intelligence and linguistics [3]. It arose in response to the many challenges presented by natural language.

The ambiguity of human language makes it extremely difficult to write software that accurately determines the meaning of text or speech data. Creating hand-written algorithms and rules that can work on any text is very complex, given the presence of homonyms, homophones, sarcasm, unspoken content, idioms, metaphors, grammar and usage exceptions in virtually any text.

NLP combines computational linguistics with statistical models and machine learning in order to successfully process human language in the form of text or speech, understanding its meaning in ways that traditional algorithms are unable to do [2].

Natural Language Processing is used to solve a wide range of tasks: in the following paper we are interested in the one of multi-class text classification. Furthermore, our attention will be drawn to a particularly problematic aspect of the NLP field: the fact that most algorithms are designed and implemented for only 7 of the more than 7000 languages spoken in the world. These are English, Chinese, Urdu, Farsi, Arabic, French, and Spanish [4].

Our objective aims to test whether different languages can influence architectures such as neural networks or statistical models. To this end, we used a dataset containing 65000 European laws, officially translated into several languages, on which we were able to compare the behaviour of various architectures, both with regard to fine-tuning and their performance.

The project code is public and available on Github.

2 Data

The chosen dataset, available on HuggingFace, comprises 65000 European laws officially translated into 23 languages.

Of these available languages, five were selected, based on their adoption at European level:

- *English*, with 51% EU speakers
- *German*, with 32% EU speakers
- *Italian*, with 16% EU speakers
- *Polish*, with 9% EU speakers
- *Swedish*, with 3% EU speakers

For each text in the dataset, there are one or more labels annotated. However, as can be understood by reading the paper on the dataset, these annotations are *granular* [1]. This allowed us to keep only the first level of labels, transforming the problem from a multi-label task to a multi-class one. This operation both simplified the problem and facilitated the evaluation of the architectures results.

This resulted in 21 different classes being initially present, shown in Figure 1 together with the number of occurrences (texts) available for each of them.

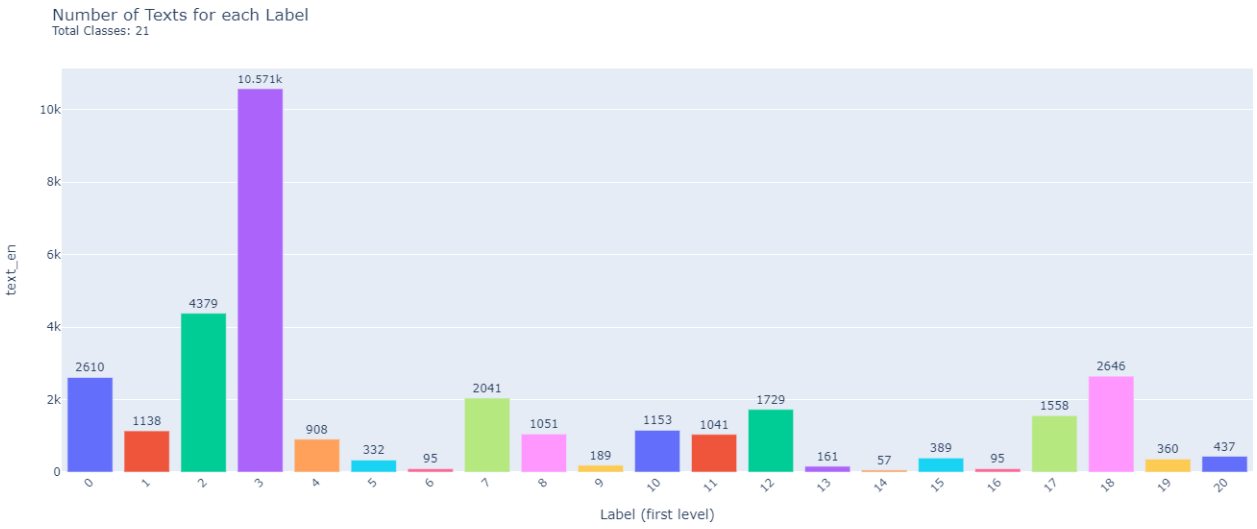

Figure 1: MultiEURLEX Dataset - Distribution of Classes and Texts Occurrences

A preliminary analysis of the data, however, revealed that each law averaged around 1200 words, with slight fluctuations depending on the considered language. Being aware that this factor would be computationally burdensome during the training phase of the models, it was decided to choose 2000 random laws from the three classes with the most observations, i.e. class 2, 3 and 18.

Ultimately, the final dataset consists of 6000 laws, officially translated into 5 languages and belonging to 3 distinct classes.

3 Model

In order to test the impact of different languages, various architectures were chosen, belonging both to the world of neural networks and statistics:

- *Long Short Term Memory*
- *Convolutional Neural Network*
- *Linear Support Vector Classifier*

TODO: DESCRIVERE LE ARCHITETTURE

4 Experimental Setup

4.1 Data Cleaning and Preprocessing

4.2 Feature Extraction

4.3 Implementing the Architectures

4.4 Training Regime

5 Results

6 Discussion

References

[1] Ilias Chalkidis, Manos Fergadiotis, and Ion Androutsopoulos. MultiEURLEX - A multi-lingual and multi-label legal document classification dataset for zero-shot cross-lingual transfer. In *EMNLP*, 2021.

[2] IBM. Natural Language Processing (NLP). https://www.ibm.com/cloud/learn/natural-language-processing. Accessed: 21-05-2022.

[3] Prakash M Nadkarni, Lucila Ohno-Machado, and Wendy W Chapman. Natural language processing: an introduction. *Journal of the American Medical Informatics Association*, 18(5):544–551, 2011.

[4] Towards Data Science. The Importance of Natural Language Processing for Non-English Languages. https://towardsdatascience.com/the-importance-of-natural-language-processing-for-non-

Accessed: 21-05-2022.