Multi-Label text classification of EU legislation documents with EUROVOC concepts

Marketos damigos
marda352

Linköping University
January 2023
Text Mining Project
Course Code: 732A81

Abstract

The aim of this project is to experiment with Multi-Label text classification in legal documents. More specifically, the dataset used is the EURLEX57K which consists of 57k legal documents, sourced from EU and annotated with approx. 4.3k EUROVOC labels. Legal documents use domain-specific language, thus making them an interesting NLP task. In this project, we used a pre-trained BERT model [2] to classify the documents. We also experimented with different models and different ways of training the model. The results were evaluated using the F1 score and the accuracy of predicting the labels. LegalBERT seems to outperform classical methods. Also an interesting find was that for classical models the use of body data had a negative impact on the performance in comparison to title data, while LegalBERT performed the same in both cases. Further investigation has to be undergone regarding the underlying reasons behind it.

1 Introduction

Text classification is a well established nlp task. It is used in many applications such as spam detection, sentiment analysis, topic classification, etc. In this project we will focus on multi-label text classification. In this task we have a set of documents and for each document we have a set of labels. The goal is to predict all the relevant labels for each document. Even more specifically, the use of domain-specific documents where large-scale of labels are available for the classification make the task more complex. Multi Label text classification is being used in many applications such as news article classification, movie genre classification, medical records, etc. European Union has as it describes it a "multilingual and multidisciplinary thesaurus" called EUROVOC [1] which consists of approx. 7k labels. These labels are manually being assigned to each legal document released by the EU. The dataset that we will use, EURLEX57k [3] consists of 57k documents which are annotated with the EUROVOC labels. Out of the 7k labels only 4.3k have been assigned to at least one document of the dataset and a little more than 2k have been assigned to more than 10 documents. This makes the dataset a good candidate for also few shot learning and zero shot learning. Unfortunatelly this project will not focus on these tasks but is a good starting point for future work.

2 Theory

2.1 Models

2.1.1 Multinomial Naive Bayes Classifier

The Multinomial Naive Bayes Classifier [5] is a probabilistic classifier based on Bayes Theorem. It is a simple and fast classifier which is often used as a baseline for text classification. It is based on the assumption that the features are independent. The classifier is trained by calculating the probability of each label given the document. The label with the highest probability is the predicted label. The formula for calculating the probability of a label given a document is:

$$P(c|d) = \frac{P(c)\prod P(w_i|c)^{f_i}}{P(d)}$$

where $P(c|d)$ is the probability of a label c given a document d, $P(d|c)$ is the probability of a document d given a label c, $P(c)$ is the probability of a label c and $P(d)$ is the probability of a document d. The probability of a document given a label is calculated by multiplying the probability of each word in the document. The probability of a word given a label is calculated by counting the number of times the word appears in documents with the label and dividing the result by the total number of words in documents. The probability of a label is calculated by counting the number of documents with the label and dividing the result by the total number of documents. The MNB classifier

is a good baseline for text classification because it is fast and simple. It is also a good choice for multi-label text classification because it can predict multiple labels for a document. The main disadvantage of the MNB classifier is that it assumes that the features are independent. This is not always the case in text classification. For example the word "not" is a negation word and it is often used in the context of a positive word. The MNB classifier will not be able to capture this relationship between the words.

2.1.2 BERT

BERT is an open-source pretrained model by GOOGLE which is based on Transformers [4]. BERT originally is trained on the BookCorpus and the English Wikipedia. BERT relies on the Transformer architecture which is a deep learning model that is based on attention mechanism. The Transformer architecture is a stack of self-attention layers and feed-forward layers. The self-attention layers are used to capture the context of the words in the document. The feed-forward layers are used to capture the context of the sentences. This makes the model more robust and it can capture the context of the words and sentences in the document in comparison to traditional NLP methods like word embeddings, with big examples being GloVe and word2vec which map each word to a vector and do not take into account the context in which the word is in. The main disadvantage of the BERT model is that it is a large model and it requires a lot of computational resources to train it. In this project we will use LegalBERT [2] which is a pretrained BERT model that is pretrained on legal documents. LegalBERT is based on the official BERT and has 12 layers with 768 hidden units, 12 heads and 110M parameters and is trained in legal documents from EU, UK and US. A graphical representation of the BERT model can be seen in the figure below.

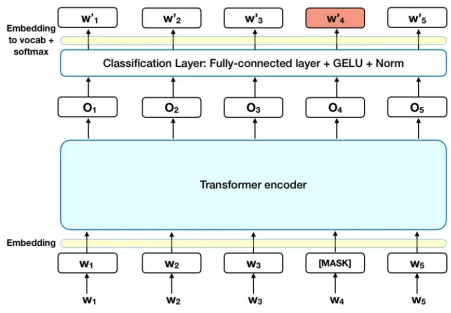

Figure 1: BERT Graph

3 Data

EURLEX57k [3] is a dataset that consists of 57k documents which have been sourced and labeled by the EU Publication Office. The documents are in English and have been labeled with multiple concepts from EUROVOC [1], which is the thesaurus labels of EU regarding their deocuments. EUROVOC contains approximately 7k different keywords, organized in 21 domains and 127 subdomains. Out of the 7k available concepts, only 4.3k have been assigned to at least one of the 57k documents. Moreovere only a litle over than 2k have been assigned to more than 10 documents. The average length of the body of each document is 547 words with a median of 399 and a maximum of 3479 words. The dataset is split into 3 subsets: train, validation and test. The train set contains 45k documents, the validation and test set contain 6k documents each.

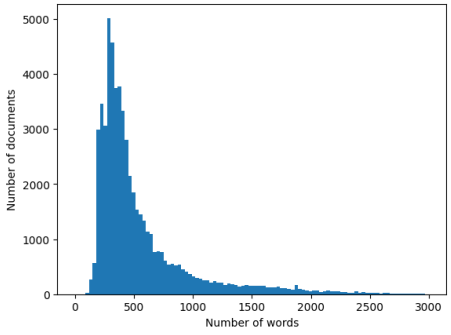

Figure 2: Distribution of number of words per document.

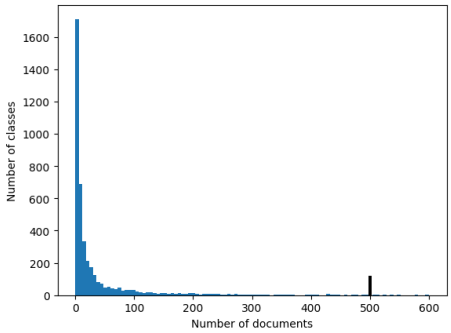

Figure 3: Number of classes per document.

	id	title
0	1000	financing
1	1005	EU financing
2	1006	compensatory financing
3	1008	financing of aid
4	1015	excise duty

Figure 4: Preview of EUROVOC concepts.

3.1 Data Preprocessing

The preprocessing of the data is an important step in the training of the model. The preprocessing of the data is done in the following steps:

1. Remove stop words, punctuation, special characters and numbers and keep only the works.
2. Remove labels that have less than 10 documents.
3. Remove the documents that have no labels (because of the previous step).
4. Tokenize the documents with tf-idf for Naive Bayes and BERT tokenizer for BERT.
5. Binary encode the labels.

All the data we saved into disk after the lematization and labels filtering in two different datases. One including only titles and the other the body. The exact same steps were apllied to both versions.

4 Training

4.1 Multinomial Naive Bayes Classifier

The Multinomial Naive Bayes was trained with the use of MultiOutputClassifier from sklearn. The model was trained with the tf-idf vectorizer and the binary encoded labels. No special parameters were passed to the model.

4.2 LegalBERT

Since BERT has a limit of 512 tokens we truncated the documents to 512 tokens. The vectorizer used in this case is the BERT tokenizer. An extra Linear layer was added to the model to transform the output of the BERT model to the number of labels. The model was trained with the AdamW optimizer and the BCEWithLogitsLoss loss function. A batch size of 16 was used in both the body and titles datasets, with 300 epochs and learning rate of 2e-05. In the case of the titles dataset, Mixed Precision was used to speed up the training process, but in the case of the body dataset the GPU used, did not support it. The top 3 models were saved and the one with the best validation loss was used for the predictions. Both models had also early stopping with patience of 5 epochs. The training was done in a RTX 3070 8GB (title dataset since body dataset did not fit) and a GTX 1080TI 11GB (body dataset). Mixed precision significantly reduced the training time of the model.

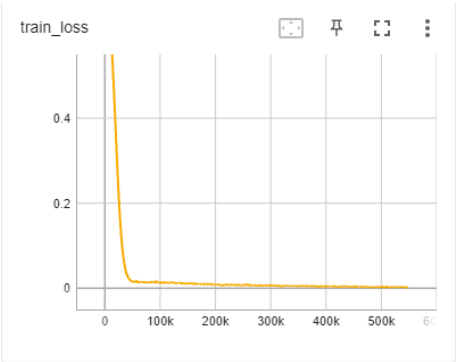

Figure 5: Training Loss

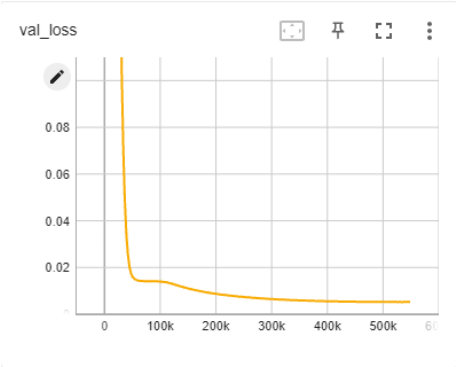

Figure 6: Validation Loss

5 Evaluation

For both models the evaluation is based on the F1-score and the accuracy. The F1-score is the harmonic mean of the precision and recall. The precision is the number of true positives divided by the number of true positives and false positives. The recall is the number of true positives divided by the number of true positives and false negatives. The accuracy is the number of true positives and true negatives divided by the total number of predictions. The evaluation was done with the use of the classification_report from sklearn.metrics. In both cases, the predictions for each class were converted to 0 and 1 based on their probability with a threshold of 0.5. So every class with a probability higher than 0.5 was classified as 1. Then the f1 score was calculated for a range of thresholds from 0.1 to 1 with 0.05 increments. The reasoning behind the use of a threshold is that we have probabilities for all the classes we need a minimum limit that make the class assign to each case. Another possible solution is to implement PR@K and F1@K where K is the K classes with the highest probability.

5.0.1 Body

	precision	recall	f1-score	support
0	1.00	1.00	1.00	12254176
1	0.86	0.07	0.13	29579
accuracy			1.00	12283755
macro avg	0.93	0.54	0.57	12283755
weighted avg	1.00	1.00	1.00	12283755

Figure 7: Classification Report for MNB with body data

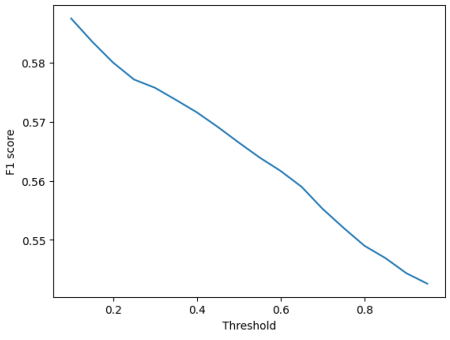

Figure 8: F1 score vs threshold for MNB with body data

5.0.2 Titles

	precision	recall	f1-score	support
0	1.00	1.00	1.00	12254176
1	0.76	0.59	0.66	29579
accuracy			1.00	12283755
macro avg	0.88	0.79	0.83	12283755
weighted avg	1.00	1.00	1.00	12283755

Figure 9: Classification Report for LegalBERT with body data

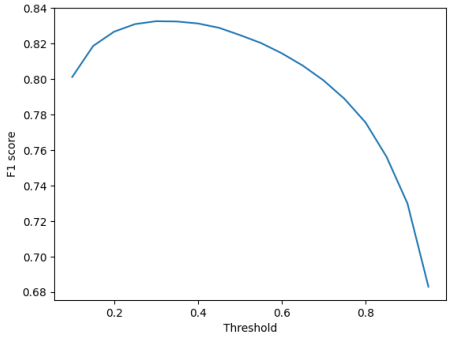

Figure 10: F1 score vs threshold for LegalBERT with body data

	precision	recall	f1-score	support
0	1.00	1.00	1.00	12254176
1	0.85	0.18	0.29	29579
accuracy			1.00	12283755
macro avg	0.92	0.59	0.65	12283755
weighted avg	1.00	1.00	1.00	12283755

Figure 11: Classification Report for MNB with titles data

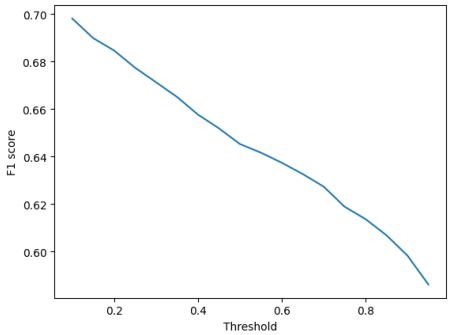

Figure 12: F1 score vs threshold for MNB with titles data

	precision	recall	f1-score	support
0	1.00	1.00	1.00	12254176
1	0.74	0.58	0.65	29579
accuracy			1.00	12283755
macro avg	0.87	0.79	0.83	12283755
weighted avg	1.00	1.00	1.00	12283755

Figure 13: Classification Report for LegalBERT with titles data

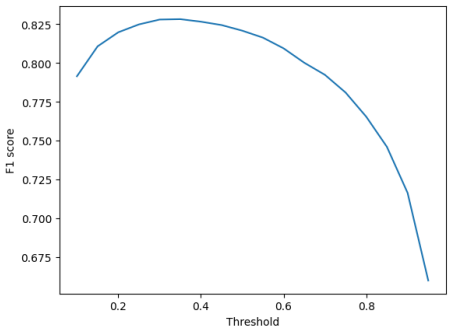

Figure 14: F1 score vs threshold for LegalBERT with titles data

	Body	Actual Tags	Predicted Tags
2833	[Commission, Regulation, EC, October, amend, R...	(1031, 1115, 1602, 2723, 4172, 862)	(1115, 1201, 1602)
3562	[Commission, Implementing, Regulation, EU, Oct...	(2687, 3191, 4080, 4315, 4317, 4319)	(2173, 2687, 2733, 3191, 4080, 4315, 4317, 4319)
158	[EEC, Commission, Decision, implementation, re...	(2970, 2971, 2979, 3774, 889, 980)	(1504, 2970, 2979, 3774, 889, 980)
837	[Commission, Regulation, EC, December, concern...	(2282, 2437, 2879, 4788, 544, 863)	(2282, 2308, 2437, 2879, 5254)
2838	[Commission, Regulation, EC, October, amend, r...	(1654, 1744, 2193, 3568, 3732, 4059, 4215, 449...	(3568, 3732, 5360)
3766	[Commission, Regulation, EC, November, repeal,...	(2282, 2879, 313, 4022, 4038, 4320, 4790, 544,...	(2282, 2879, 313, 4320, 4790, 544)
3643	[EC, decision, European, Central, Bank, Decemb...	(2149, 2447, 3259, 4763, 5455, 5883)	(2149, 2447, 3259, 4763, 5455, 5883)
5525	[Commission, Regulation, EEC, April, establish...	(3611, 3759, 4164, 4385)	(3611, 3759, 4385)
1856	[Commission, Regulation, EEC, July, establish,...	(2331, 2783, 3611, 4385)	(3611, 4385)
548	[Commission, Regulation, EC, July, fix, repres...	(1309, 1863, 2687, 4080, 4314)	(1309, 1863, 2687, 4080, 4314)

Figure 15: Sample of predictions from test set.

6 Discussion

In our results we can notice that in both title and body data there seems to not be significant difference in the results. This may be related with the fact that title data are comprehensive and share parts with the body data. At threshold of 0.5 LegalBERT perforfs significantly better than Multinomial Naive Bayes with a weighted average F1-score of 0.83 versus 0.57 respectively for body data and 0.83 versus 0.65 for title data. Here it is important to note the big difference in MNB for title data versus body data with 0.09 in favor of title data. The training time for the two models was 30 minutes for the MNB with title data while 600 minutes for the body data. There does not seem to be a clear explanation behind why this happens but it is interesting to be further investigated. In both cases and models we notice the same plot for the F1-score in a range of thresholds from 0.1 to 1 with an interval of 0.05 but very different between MNB and LegalBERT. In the first we have a constant decline while LegalBERT has an incline up to 0.4 with a local maximum and then declines. This pattern applies in both body and titles data. LegalBERT seems to outperform classical models in terms of accuracy and be a suitable model for multiclass classification of legal documents. The world of NLP is a fast evolving one and the legal domain can be benefited significantly from the advances in the field. Such models can be used to classify legal cases or even help the lawyers to find the right documents for their cases, solve legal issues or even help the judges to make better decisions and in shorter time. Another topic of that domain would be the summarization of such long documents. All these can be considered as future work.

7 Conclusion

This project intented to explore the use of NLP and Transformers that are pre-trained with legal data. The results seem promising and the models can be used in the future for other tasks in the legal domain. With the emerging advances in the field of NLP, the legal domain can be benefited significantly. The models from the above project can be further fine-tuned and used with bigger datasets. An interesting future work would be to modify the model and use it for the summarization of the documents. This would be a great help for the lawyers and judges to make better decisions and in shorter time.

8 Code

The Github repository for this project can be found here.

References

[1] Publications office of the european union.

[2] Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. LEGAL-BERT: The muppets straight out of law school. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904, Online, November 2020. Association for Computational Linguistics.

[3] Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, and Ion Androutsopoulos. Large-scale multi-label text classification on EU legislation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6314–6322, Florence, Italy, 2019. Association for Computational Linguistics.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.

[5] Shuo Xu, Yan Li, and Zheng Wang. Bayesian multinomial naïve bayes classifier to text classification. In James J. (Jong Hyuk) Park, Shu-Ching Chen, and Kim-Kwang Raymond Choo, editors, *Advanced Multimedia and Ubiquitous Engineering*, pages 347–352, Singapore, 2017. Springer Singapore.