Categorisation of Bulgarian Legislative Documents

Nikola Obreshkov **Martin Yalamov** **Svetla Koeva**
Institute for Bulgarian Language "Prof. Lyubomir Andreychin"
Bulgarian Academy of Sciences
{nikola,martin,svetla}@dcl.bas.bg

Abstract

The paper presents the categorisation of Bulgarian MARCELL corpus in top-level EuroVoc domains. The Bulgarian MARCELL corpus is part of a recently developed multilingual corpus representing the national legislation in seven European countries. We performed several experiments with JEX Indexer, with neural networks and with a basic method measuring the domain-specific terms in documents annotated in advance with IATE terms and EuroVoc descriptors (combined with grouping of a primary document and its satellites, term extraction and parsing of the titles of the documents). The evaluation shows slight overweight of the basic method, which makes it appropriate as the categorisation should be a module of a NLP Pipeline for Bulgarian that is continuously feeding and annotating the Bulgarian MARCELL corpus with newly issued legislative documents.

Keywords: document categorisation, document classification, legislative domain

1. Introduction[1]

The paper presents the categorisation of Bulgarian MARCELL corpus in top-level EuroVoc domains[2]. The Bulgarian MARCELL corpus is part of a recently developed multilingual corpus representing the national legislation in seven European countries. The presented work is an outcome of the CEF Telecom project Multilingual Resources for CEF.AT in the Legal Domain1 (MARCELL) aiming to enhance the eTranslation system of the European Commission by supplying large-scale domain specific data in seven languages (Bulgarian, Croatian, Hungarian, Polish, Romanian, Slovak and Slovenian).

The Bulgarian MARCELL corpus consists of 27,283 documents (at the beginning of April 2020), which are classified into fifteen types: Administrative Court, Agreements, Amendments (Legislative acts), Compacts, Conventions, Decrees, Decrees of the Council of Ministers, Decisions of the Central Election Commission, Decisions of the Constitutional Court, Decisions of the Council of Ministries, Guidelines, Instructions, Laws (Acts), Memorandums and Resolutions.

Classifying the national legislation documents into EuroVoc classes serves the purpose of compiling multilingual domain-specific corpora corresponding to top-level EuroVoc domains. Only a few of the national legislations in the seven countries have been (manually) classified so far according to the EuroVoc Thesaurus (Croatian and Slovenian). The initial task was to categorise national legislation documents with the JRC EuroVoc indexer software – JEX (Steinberger et al., 2012). The high number of categories used by JEX Indexer, combined with a very unevenly balanced training set, is a big challenge for a multi-label categorisation task and even bigger for a one-label classification task. We show that

[1] Sections 1, 2, 3, 4, 7 are written by Sv. Koeva.
[2] https://op.europa.eu/en/web/eu-vocabularies/th-dataset/-/resource/dataset/eurovoc

taking into consideration the specific properties of legislative documents (namely, the specific terminology and the structure of the titles used in legislative documents) can be exploited for the document classification task.

The paper is organised as follows: in Section 2, we present in brief the related work in the field of categorisation of legislative documents. Sections 3 and 4 describe the specific tasks we are going to solve and the preliminary processing of the documents. The methods we have used for categorisation of legislative documents are presented in Section 5, and the evaluation of the results is presented in Section 6. Finally, Section 7 presents conclusions for our results and explains how the categorisation of legislative documents will be further enriched. The target result is a large-scale monolingual corpus of Bulgarian national legislation organised by EuroVoc top-level domains in thematically related sets of documents.

2. Related work

The EUR-Lex2 database of legal documents of the European Union served as a document collection for several classification methods. (Mencia and Frnkranz, 2007) studied multi-label classification problems, the largest being the categorisation of the EUR-Lex legal documents into the EuroVoc concept hierarchy with almost 4,000 classes. Three algorithms were evaluated: (i) the binary relevance approach, which independently trains one classifier per label; (ii) the multi-class multi-label perceptron algorithm, which respects dependencies between the base classifiers; and (iii) the multi-label pairwise perceptron algorithm, which trains one classifier for each pair of labels, the latest showing a good predictive performance.

Some of our experiments are performed with JEX Indexer – a free, multi-class, multi-label classification tool (Steinberger et al., 2012) provided with pre-trained models for 27 languages, including Bulgarian. The document to be indexed is represented as a vector of the same features (inflected word forms, n-grams, etc.) with their frequency in the document. The training documents (22,692 for Bulgarian covering 2,147 EuroVoc categories) are represented as a log-likelihood-weighted list of features, using the training document set as the reference corpus. The most appropriate categories for the new document are found by ranking the category vector representations according to their cosine similarity with the vector representation of the document to be indexed. JEX uses large numbers of stop words (332 for Bulgarian) that are ignored in the classification process. In order to optimise the profile generation for each class, a number of different parameter settings were optimised by selecting the best-performing setting within a range of values. The following are some of the most important parameters used: How many training documents there must be at least for a class to be trained; How long these training documents must be at least; How often words need to be found in the corpus in order to be used as associates; How statistically relevant a word must be in a training document in order to be considered; How to weigh words depending on the number of descriptors assigned to each training document (Steinberger et al., 2012). The reported precision for Bulgarian is 0,4619, recall – 0,5120 and F1 – 0,4940.

Filtz et al. (2019) uses different approaches to compare the performance of text classification algorithms on existing datasets and corpora of legal documents. For the EUR-Lex legal datasets, the authors show that exploiting the hierarchy of the EuroVoc thesaurus helps to improve classification performance by reducing the number of potential classes while retaining the informative value of the classification itself. Their results suggest that the advantage of using neural networks for the legal document classification problem is lower compared to text classification in other domains.

There are many examples for classification using variants of recurrent or convolutional neural networks (Howard, 2018; Jacovi, 2018). Some recent efforts are towards the so-called Extreme multi-label text classification (XMTC) – the most relevant class labels from an extremely large label collection are assigned to each document (Liu et al., 2017: 115). Kim (2014) reported on a series of experiments with convolutional neural networks (CNN) trained on top of pre-trained word vectors for sentence-level classification tasks and showed that a simple CNN with little hyperparameter tuning and static vectors achieved good results on multiple benchmarks. Liu et al. (2017) applied also deep learning to XMTC,

with a family of Convolutional Neural Network models, which are tailored for multi-label classification and reported results on several benchmark datasets, including EUR-Lex.

Chalkidis et al. (2019) released a new dataset of 57k legislative documents from EUR-Lex, annotated with concepts from EuroVoc. The dataset is substantially larger than previous EUR-Lex datasets and suitable for XMTC. Experiments with several neural classifiers were performed, and it is claimed that BIGRU with self-attention outperforms the current multi-label state-of-the-art methods, which employ label-wise attention.

To sum up, although the neural networks are widely used in classification tasks, there are results showing that the neural networks might not be very appropriate for particular domains, including legislative documents. It is very difficult to produce or reuse large training datasets in the legal domain, and such do not exist (to the best of our knowledge) for legislative documents (in Bulgarian).

3. Problem and Proposed Approach

Our efforts are directed towards the categorisation of Bulgarian legislative documents in top-level EuroVoc domains. Most of the classification approaches use a limited number of classification labels. The EuroVoc thesaurus contains 7,139 descriptors (labels) and is appropriate for the classification of documents in a multi-label classification. In contrast, our task is the classification of legislative documents into one of the top-level domains of EuroVoc: Politics, International relations, European Union, Law, Economics, Finance, Social questions, Education and Communications, Science, Business and Competition, Employment and Working conditions, Transport, Environment, Agriculture, Forestry and Fisheries, Agri-foodstuffs, Production, Technology and Research, Energy, Industry, Geography, International organisations. The limitations of the EuroVoc thesaurus are: it has been designed to meet the needs of systems of general documentation on the activities of the European Union; it cannot cover the various national situations at a sufficiently detailed level[3]. We reduced the classes to 19, excluding Geography and International organisations, as they are not representative for the national legislation.

We decided to make several experiments: a) with JEX indexer converting its multi-label categorisation to one-label categorisation; b) with neural networks using an unbalanced training set for Bulgarian annotated with IATE terms and EuroVoc descriptors; c) with a method measuring the domain-specific IATE terms and EuroVoc descriptors in the documents. The last approach is combined with term extraction, grouping of the primary document and its secondments and categorization by the titles of the documents.

4. Pre-processing of Documents

4.1. Part-of-Speech Tagging and Lemmatisation

For pre-processing Bulgarian legislative documents, we use the pipeline that integrates a sentence splitter, a tokeniser, a part-of-speech tagger, a lemmatiser, a named entity recogniser, a noun phrase parser, an IATE term annotator and a EuroVoc descriptor annotator. All tools are self-contained and part of them are designed to work in a chain, i.e. the output of the previous component is the input for the next component, starting from the sentence splitter and following the strict order for the tokeniser, the POS tagger and the lemmatiser (Bulgarian Language Processing Chain – BGLPC). In particular, we use enhanced versions of the sentence splitter, the tokeniser, the part-of-speech tagger and the lemmatiser for tagging and lemmatisation (Koeva and Genov, 2011). The output is in the CoNLL- U Plus format[4].

4.2. Term Recognition

The term recognition7 is performed via automatic text analysis methods in order to identify words and multiword expressions fulfilling the criteria for terms. The focus texts and the reference texts (texts from literature and news that are supposed not to contain terms) are tagged for part-of-speech and lemmatised. This ensures that each multiword term in the focus texts can be matched against the following linguistic filters (N, AN, AAN, NRN, ANRN, NRAN, ANRAN, NN, ANN, NAN, where A is adjective, N –

[3] https://eur-lex.europa.eu

[4] https://universaldependencies.org/format.html

noun, R – preposition) and that frequencies can be calculated correctly when terms are used in different word forms. For each sequence of part-of-speech tags in the focus texts matching one of the linguistic filters and for each adjective from the reference corpus the following information was indexed: the number of texts in which they occur and the number of all occurrences. Multiword term candidates that contain indexed reference adjectives are eliminated.

To compare the number of occurrences of term candidates in the focus texts with the number of their occurrences in the reference corpus TF-IDF and Log Likelihood algorithms are implemented[5]. The threshold for TF-IDF is set to 0.02. We use the union of the results from Tf-IDF and Log Likelihood. To increase the results the algorithm Dice is applied, which identifies terms similar to those already recognised (with a threshold set to 0.85). Processing the legislative corpus, we extracted 813,118 term candidates.

4.3. Annotation with IATE Terms and EuroVoc Descriptors

The Bulgarian MARCELL corpus has been annotated with terminology from two terminology repositories: IATE – 'Interactive Terminology for Europe'[6], the EU's terminology database used in the EU institutions and agencies, and EuroVoc[7], a multilingual, multidisciplinary thesaurus covering the activities of the EU.

For IATE term and EuroVoc descriptor annotation, a dedicated instrument called TextAnnotator was developed (Koeva et al., 2020). The TextAnnotator[8] calls dictionaries of terms and finds occurrences of these terms in the documents. Both the documents and the dictionaries are structured in the CoNLL-U Plus format (token, part-of-speech tag, lemma, extended grammatical tags) and each token is associated with a term descriptor. The annotation tool matches sequences of lemmas and part-of-speech tags of dictionary entries and lemmas and part-of-speech tags of document tokens. The matching procedure is based on a hash table indexing. For each dictionary entry, a hash key is generated concatenating lemmas and part-of-speech tags within it. All hash keys for a given dictionary are grouped into length classes based on the number of tokens they contain. The algorithm gives a priority to the longest length classes, which ensures the selection of longest matches. When a match is found, the corresponding tokens in the document are annotated with a term (Identification numbers of IATE terms or EuroVoc descriptors) and the processing continues from the end index of the match. The identification numbers (IDs) of the IATE terms point also to the relevant EuroVoc domains and subdomains. There are 45,592 IATE terms for Bulgarian. The annotation takes into account that several terms can be related with one and the same IATE ID (synonymy) and one term can be related with different IDs (polysemy). There are also IATE terms in Bulgarian, which describe concepts specific for other languages, i.e. община (obshtina) IATE ID: 3553038 'regions of Poland'. Such terms were excluded from the annotation (4,641 terms altogether). As a result, 13,799,334 IATE terms and 3,386,437 EuroVoc descriptors were annotated. IATE terms and EuroVoc descriptors are numbered within a given sentence (starting from 1) and the number is repeated for each token belonging to the term. The IATE identification number for the term is listed followed by the numbers of the corresponding EuroVoc descriptor(s).

4.4. Grouping of Documents

The documents are related to a primary legislative act, if such exists (i.e., a law and an instruction to this law). The documents' titles are divided into two parts: a general part that describes the type of the document (i.e. Закон за висшето образование (Zakon za visshetoto obrazovanie – 'The higher education act') and a differential part that describes the topic of the document (i.e. Закон за висшето образование (Zakon za visshetoto obrazovanie 'The higher education act'). The title of a primary document contains exactly two parts, while the titles of satellite documents contain at least two general parts and one differential part. The documents build a group if their differential parts and the nearest general parts match. The titles are lemmatised; then only lemmas are further used for matching. There

[5] The Term Recognition is developed by DImitar Georgiev.

[6] https://iate.europa.eu/home

[7] https://op.europa.eu/en/web/eu-vocabularies/th-dataset/-/resource/dataset/eurovoc

[8] The TextAnnotator is developed by Nikola Obreshkov.

might be differences in using punctuation marks, brackets, capital letters, abbreviations, dates, etc. in the title of the primary legislative act and its secondments. Several procedures are performed to predict possible differences and to match correctly thematically related documents.

4.5. Test Dataset

A manually annotated test corpus was developed including a total of 667 documents. 275 documents[9] were manually annotated with multiple labels among which the most appropriate label was also selected. For the annotation of the remaining 392 documents, 2000 documents were automatically annotated in advance with multiple labels which assisted the manual annotation with the most appropriate label. The features of the EuroVoc thesaurus discussed above were reflected in the fact that in some cases it was difficult even for a human expert to classify a given legislative document to the EuroVoc top-domains (many domains that are subject of legislation are not present, i.e. Health, Culture, etc.).

Code	EuroVoc Domain Name	Number of Documents
04	Politics	129
24	Finance	61
66	Energy	57
08	International relations	56
12	Law	52
44	Employment and working conditions	51
28	Social questions	37
52	Environment	32
56	Agriculture, forestry and fisheries	32
20	Economics	30
32	Education and communications	28
60	Agri-foodstuffs	20
48	Transport	18
68	Industry	15
10	European Union	11
40	Business and competition	10
64	Production, technology and research	10
16	Economics	6
36	Science	3

Table 1: The distribution of documents in the test corpus

5. Experiments

5.1. Categorisation with JEX Indexer

For a given document, JEX Indexer assigns several labels among more than 6,700 EuroVoc descriptors with corresponding likelihood weights. The default settings of the system were used: at least 4 training

[9] The 275 documents were manually annotated by Ts. Dimitrova and V. Stefanova.

documents for a class; at least 100 words per training document; at least 4 occurrences of a word in the corpus which are used as associates; minimum log-likelihood value of 5 to consider a word statistically relevant in a training document. Example 1 shows a document categorised with six labels (category code is the EuroVoc descriptors' ID) and the assigned log-likelihood weight.

```
<document id="bg-81407.txt">
    <category code="4585" weight="0.18519803030161675 "></category>
    <category code="1684" weight="0.18393845477668894 "></category>
    <category code="1234" weight="0.16739820210223233 "></category>
    <category code="365" weight="0.13740067398049766 "></category>
    <category code="1021" weight="0.11888676884171023 "></category>
    <category code="2900" weight="0.11647743754115164 "></category>
</document>
```

Figure 1: JEX Indexer categorisation output for document bg-81407.txt

The one-label categorisation of Bulgarian legislative texts was performed in two steps:

Each document was annotated with weighted EuroVoc descriptors using the JEX Indexer tool.

The annotated descriptors were grouped into one top-domain by the hierarchical relations to broaden terms up to the top-level as well as by the associative relations to related terms.

If more than one descriptor points to a top-domain, the weights of descriptors are summed up. For example, the document bg-81407.txt is classified to the following top-domains:

24 0.725 FINANCE

20 0.184 TRADE

The weight 0.725 for top-domain Finance is calculated by summing the weights of 5 descriptors, assigned by the JEX indexer: 4585 данък върху добавената стойност (danak varhu dobavenata stoynost) 'value added tax', 1234 фискална хармонизация (fiskalna harmonizaciya) 'tax harmonisation', 365 данъчно облекчение (danachno oblekchenie) 'tax relief', 1021 данъчна система (danachna sistema) 'tax system' and 2900 данъчна основа (danachna osnova) 'basis of tax assessment'. The experiment was repeated by setting a threshold for summed log-likelihood values to 0.1, 0.2, and 0.3 respectively.

5.2. Neural Categorisation

First, we tested the JEX language model trained with the European legal texts, but the results were not good. Therefore, we opted to train a new model based on the Bulgarian MARCELL corpus. We use the version of the corpus annotated with IATE terms and EuroVoc descriptions. The annotated terms were linked to the EuroVoc top-domains and the documents were sorted by the number of the top-domain associations. For every EuroVoc top-domain up to 500 training documents were selected; however, for some top-domains the number of associated documents was very small: Energy – 82, Production, Technology and Research – 10, International organisations – 4 and so on.

The generated training corpus was used to train a neural model with TensorFlow[10] and Keras[11]. The built-in Keras tokeniser is used for tokenisation. First, a vocabulary is constructed containing unique tokens in the documents without taking into account their frequency. Then each document is transformed into a sequence of integers based on the presence of its words in the vocabulary. The selected design of the neural network returns the credibility for each label candidate. A threshold is set up to specify the level of credibility.

[10] https://www.tensorflow.org/

[11] https://keras.io/

5.3. Basic Categorisation

The basic categorisation relies on pre-processing: sentence splitting, tokenisation, part-of-speech tagging, lemmatisation, annotation with IATE terms and EuroVoc descriptors. The IATE subject fields link IATE terms with EuroVoc descriptors; the fields of knowledge in which the IATE concepts are used (one IATE term can be linked with several EuroVoc descriptors). The annotated EuroVoc descriptors were grouped into top-domains by the associative relations and the hierarchical relations linking the EuroVoc descriptors. For each document the obtained top-domains are summed up and the numbers of associations to top-domains are sorted and the top-domain with highest number of associations is selected as the document category. Figure 1 shows the basic categorisation pipeline.

5.4. Basic Categorisation Combined with Term Extraction

The legislative domain is a domain with unique or specialised terminology. As we do not analyse the context and disambiguate the word senses, we decided to experiment with term extraction. The basic categorisation is performed in the same manner with the only difference that IATE and EuroVoc dictionaries used for the IATE term and EuroVoc descriptor annotation are filtered to contain only the obtained term candidates.

5.5. Basic Categorisation Combined with Documents' Grouping

We also use the groups formed by a primary legislative act and its satellite documents. The assumption is that all thematically related documents should belong to one category. Based on this assumption two different experiments were performed on top of the Basic classification:

Normal Grouping - All documents within the group are considered as a single document. This experiment is implemented on top of the basic categorisation. For every such document, the EuroVoc term annotations are combined and the EuroVoc top-domain associations are recalculated.

Hierarchical Grouping - Only the primary document is used to represent the whole group. Its EuroVoc term annotations are used to select a class that is then assigned to its related satellite documents.

5.6. Basic Categorisation Combined with Titles' Classification

The method uses the EuroVoc descriptors occurring within the documents' titles and their association with EuroVoc top-domains. The results are scored based on counting the descriptors' lemma matches in different combinations. The most probable top-level domain is the one that has the highest score, which is calculated with the following formula:

score = TotalPhraseMatches * 10 + TotalDescriptorMatches + TotalDescriptorSequenceMatches/100 where:

TotalPhraseMatches is defined as the total number of branches pointing to a top-domain where a lemmatised descriptor is matched with a part of the title (regardless of the word order if the descriptor is a multiword term);

TotalDescriptorsMatches is defined as the total number of single lemmatised descriptors matched in the title for each top-domain;

TotalDescriptorSequenceMatches is defined as the total number of branches pointing to a top-domain where a lemmatised descriptor is matched with a part of the title keeping the word order if the descriptor is a multiword term.

The top category assigned by the titles' categorisation is compared with the category candidates obtained by the basic categorisation. Two scenarios are possible:

The top category of the titles' categorisation is not among the category candidates of the basic categorisation. In this case, the basic categorisation is not affected.

The top category of the titles' categorisation overlaps with one of the category candidates of the basic categorisation. In this case, the respective top-domain annotations count is multiplied by a predefined weight. This may result in reordering the category candidates and promoting a different top category for the current document.

6. Evaluation of Results

For every document the manually annotated class (category) from test corpus was compared with the suggested class by the selected classification method. Initially, per-class precision and recall were evaluated, where

per-class precision = correctly predicted documents with this class / all predicted documents with this class

per-class recall = correctly predicted documents with this class / all documents with this class

Then a macro-averaged precision and a macro-averaged recall were calculated for the whole classifier. This was done with arithmetic mean of the per-class precision and recall. Finally, the harmonic mean of the macro-averaged precision and the macro-averaged recall was used for the F1-score estimation. F1 = 2 * precision * recall / (precision + recall).

The results of the described methods for text categorisation of the Bulgarian legislative documents are presented in Table 2.

Categorization method	Accuracy	Precision	Recall	F1
JEX Indexer	46.36	44.58	45.43	45
Neural model	16.71	15.53	17.38	16.04
Categorisation by Titles	37.39	51.34	40.98	45.58
Basic method	54.42	52.64	70.18	61.16
Basic method + Term Extraction	36.28	40.53	45.42	42.84
Basic method + Normal Grouping	54.03	51.84	69.91	59.54
Basic method + Hierarchical Grouping	54.64	52.08	70.13	60.24
Basic method + Categorisation by Titles (1.3)	57.83	55.04	72.44	62.55
Basic method + Categorisation by Titles (1.5)	58.3	55.23	72.41	62.66
Basic method + Categorisation by Titles (2)	55.66	53.91	57.97	55.87
Basic method + Hierarchical Grouping + Categorisation by Titles (1.5)	58.51	55.37	72.52	62.8

Table 2: The evaluation results

The results achieved with the JEX indexer and the Neural model were not optimal and their improvement may require manual annotation of a large training corpus. For the current task these methods were omitted and the focus were pointed towards the basic method which doesn't need a training corpus. Given the challenges of this classification task, the baseline results achieved by the Basic method were reasonable and seven additional experiments were performed in pursuit of further improvement. The Term Extraction filtering resulted in a dropout of the measured scores. Similarly, the Normal Grouping did not improve the results indicating that the primary document of the group holds the essential (most relevant) information. This hypothesis was also confirmed by the Hierarchical Grouping method where only the primary document was used. It produced better results than the Normal Grouping method and a slight increase of precision and a slight decrease of recall compared to the Basic method. Three experiments were performed with the combination of the Basic method and the Titles classification. For every experiment a different weight was used to control the importance of the Title class when applied to the Basic method classification. The best results were achieved with a weight of 1.5. The last experiment was to combine the Basic method with the Hierarchical Grouping and with the Titles classification weighted with 1.5. This combination led to the best results from all experiments. It

can be seen that the addition of Hierarchical Grouping improved the result of the Basic method and Titles classification both in precision and recall.

7. Conclusions

The categorisation of legislative documents is a part of the NLP pipeline for Bulgarian, which continuously feeds the Bulgarian MARCELL corpus with newly issued legislative documents and makes changes to the data format, organises data in structures, accumulates data with linguistic information, analyses data and provides explicit links between different data segments. The absence of a relevant training corpus with legislative data in Bulgarian (and the stumbling block for creation of a training dataset relating with the relatively small number of documents in the legislative domain) presupposes the limited performance of neural methods of any supervised machine learning approaches. The target result - a large-scale monolingual corpus of Bulgarian national legislation categorised by EuroVoc top-level domains - is achieved by applying a Basic method which relies on the annotation of IATE terms and EuroVoc descriptors within a document. Some restrictions have been applied to reduce the ambiguity effect (manual removal of inappropriate annotations with more than 4 IATE terms and use of the IATE subject fields that link IATE terms with EuroVoc descriptors). The assumption that the legislative documents that are linked to a primary legislative act must belong to the same category, and that the titles of legislative documents contain information about their category led to the performance of seven experiments. The results show that the specific properties of legislative documents (legislative terminology, relations between legislative acts and the structure of the titles used in legislative documents) can be successfully exploited for the document classification task in the legislative domain.

Acknowledgements

The presented work is an outcome of the CEF Telecom project Multilingual Resources for CEF.AT in the Legal Domain1 (MARCELL).

The Term Recognition is developed by Dimitar Georgiev.

References

Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N., Androutsopoulos, I. (2019). *Extreme Multi-Label Legal Text Classification: A case study in EU Legislation.* CoRR abs/1905.10892.

Filtz, E., Kirrane, S., Polleres, A., Wohlgenannt, G. (2019). *Exploiting EuroVoc's Hierarchical Structure for Classifying Legal Documents.* Lecture Notes in Computer Science On the Move to Meaningful Internet Systems: OTM 2019 Conferences, pages 164-181.

Howard, J. and Ruder, S. (2018). *Fine-tuned language models for text classification.* CoRR abs/1801.06146, `http://arxiv.org/abs/1801.06146`.

Jacovi, A., Shalom, O.S., Goldberg, Y. (2018). *Understanding convolutional neural networks for text classification.* CoRR abs/1809.08037.

Kim, Y. (2014). *Convolutional Neural Networks for Sentence Classification.* Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1746– 1751.

Koeva, S. and Genov, A. (2011). *Bulgarian Language Processing Chain.* In Proceedings of Integration of multilingual resources and tools in Web applications. Workshop in conjunction with GSCL 2011, University of Hamburg.

Koeva, S., Obreshkov, N., Yalamov, M. (2020). *Natural Language Processing Pipeline to Annotate Bulgarian Legislative Documents.* Proceedings of The 12th Language Resources and Evaluation Conference. European Language Resources Association, 2020, 6988–6994.

Liu, J., Chang, W., Wu, Y. and Yang, Y. (2017). *Deep Learning for Extreme Multi-label Text Classification.* Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '17, pages 115–124.

Mencia, E. L. M. and Frnkranz, J. (2007). *Efficient Multilabel Classification Algorithms for Large-Scale Problems in the Legal Domain.* Proceedings of the LWA 2007, pages 126–132.

Steinberger R., Ebrahim, M. and Turchi, M. (2012). *JRC EuroVoc Indexer JEX - A freely available multi-label categorisation tool.* Proceedings of the 8th international conference on Language Resources and Evaluation (LREC'2012), Istanbul, 21-27 May 2012.