J. Hlaváčová (Ed.): ITAT 2017 Proceedings, pp. 186–192
CEUR Workshop Proceedings Vol. 1885, ISSN 1613-0073, © 2017 L. Lenc, P. Král

ITAT

Ensemble of Neural Networks for Multi-label Document Classification

Ladislav Lenc[1,2] and Pavel Král[1,2]

[1] Department of Computer Science and Engineering, Faculty of Applied Sciences,
University of West Bohemia, Univerzitní 8, 306 14 Plzeň, Czech Republic
[2] NTIS—New Technologies for the Information Society, Faculty of Applied Sciences,
University of West Bohemia, Technická 8, 306 14 Plzeň, Czech Republic
nlp.kiv.zcu.cz
{llenc,pkral}@kiv.zcu.cz

Abstract: This paper deals with multi-label document classification using an ensemble of neural networks. The assumption is that different network types can keep complementary information and that the combination of more neural classifiers will bring higher accuracy. We verify this hypothesis by an error analysis of the individual networks. One contribution of this work is thus evaluation of several network combinations that improve performance over one single network. Another contribution is a detailed analysis of the achieved results and a proposition of possible directions of further improvement. We evaluate the approaches on a Czech ČTK corpus and also compare the results with state-of-the-art approaches on the English Reuters-21578 dataset. We show that the ensemble of neural classifiers achieves competitive results using only very simple features.

Keywords: Czech, deep neural networks, document classification, multi-label

1 Introduction

This paper deals with multi-label document classification by neural networks. Formally, this task can be seen as the problem of finding a model M which assigns a document $d \in D$ a set of appropriate labels (categories) $c \in C$ as follows $M : d \rightarrow c$ where D is the set of all documents and C is the set of all possible document labels. The multi-label classification using neural networks is often done by thresholding of the output layer [1, 2]. It has been shown that both standard feed-forward networks (FNNs) and convolutional neural networks (CNNs) achieve state-of-the-art results on the standard corpora [1, 2].

However, we believe that there is still some room for further improvement. A combination of classifiers is a natural step forward. Therefore, we combine a CNN and an FNN in this work to gain further improvement in the terms of precision and recall. We support the claim that combination may bring better results by studying the errors of the individual networks. The main contribution of this paper thus consists in the analysis of errors in the prediction results of the individual networks. Then we present the results of several combination methods and illustrate that the ensemble of neural networks brings significant improvement over the individual networks.

The methods are evaluated on documents in the Czech language, being a representative of highly inflectional Slavic language with a free word order. These properties decrease the performance of usual methods. We further compare the results of our methods with other state-of-the-art approaches on English Reuters-21578[1] dataset in order to show its robustness across languages. Additionally we analyze the final F-measure on document sets divided according to the number of assigned labels in order to improve the accuracy of the presented approach.

The rest of the paper is organized as follows. Section 2 is a short review of document classification methods with a particular focus on neural networks. Section 3 describes our neural network models and the combination methods. Section 4 deals with experiments realized on the ČTK and Reuters corpora and then analyzes and discusses the obtained results. In the last section, we conclude the experimental results and propose some future research directions.

2 Related Work

Document classification is usually based on a supervised machine learning. A classifier is trained on an annotated corpus and it then assigns class labels to unlabelled documents. Most works use vector space model (VSM), which generally represents each document as a vector of all word occurrences usually weighted by their tf-idf.

Several classification methods have been successfully used [3], as for instance Bayesian classifiers, maximum entropy, support vector machines, etc. However, the main issue of this task is that the feature space is highly dimensional which decreases the classification results. Feature selection/reduction [4] or better document representation [5] can be used to solve this problem.

Nowadays, "deep" neural nets outperform majority of the state-of-the-art natural language processing (NLP) methods on several tasks with only very simple features. These include for instance POS tagging, chunking, named entity recognition and semantic role labelling [6]. Several different topologies and learning algorithms were proposed. For instance, Zhang et al. [7] propose two convolutional neural nets (CNN) for ontology classification, sen-

[1] http://www.daviddlewis.com/resources/testcollections/reuters21578/

timent analysis and single-label document classification. They show that the proposed method significantly outperforms the baseline approach (bag of words) on English and Chinese corpora. Another interesting work [8] uses in the first layer pre-trained vectors from word2vec [9]. The authors show that the proposed models outperform the state of the art on 4 out of 7 tasks, including sentiment analysis and question classification. Recurrent convolutional neural nets are used for text classification in [10]. The authors demonstrated that their approach outperforms the standard convolutional networks on four corpora in single-label document classification task.

On the other hand, traditional feed-forward neural net architectures are used for multi-label document classification rather rarely. These models were more popular before as shown for instance in [11]. They build a simple multi-layer perceptron with three layers (20 inputs, 6 neurons in hidden layer and 10 neurons in the output layer, i.e. number of classes) which gives F-measure about 78% on the standard Reuters dataset. The feed-forward neural networks were used for multi-label document classification in [12]. The authors have modified standard backpropagation algorithm for multi-label learning (BP-MLL) which employs a novel error function. This approach is evaluated on functional genomics and text categorization.

A recent study on multi-label text classification was proposed by Nam et al. in [1]. The authors build on the assumption that neural networks can model label dependencies in the output layer. They investigate limitations of multi-label learning and propose a simple neural network approach. The authors use cross-entropy algorithm instead of ranking loss for training and they also further employ recent advances in deep learning field, e.g. rectified linear units activation, AdaGrad learning with dropout [13, 14]. TF-IDF representation of documents is used as network input. The multi-label classification is handled by performing thresholding on the output layer. Each possible label has its own output node and based the final value of the node a final decision is made. The approach is evaluated on several multi-label datasets and reaches results comparable to the state of the art.

Another method [15] based on neural networks leverages the co-occurrence of labels in the multi-label classification. Some neurons in the output layer capture the patterns of label co-occurrences, which improves the classification accuracy. The architecture is basically a convolutional network and utilizes word embeddings for initialization of the embedding layer. The method is evaluated on the natural language query classification in a document retrieval system.

An alternative approach to handling the multi-label classification is proposed by Yang and Gopal in [16]. The conventional representations of texts and categories are transformed into meta-level features. These features are then utilized in a learning-to-rank algorithm. Experiments on six benchmark datasets show the abilities of this approach in comparison with other methods.

Another recent work proposes novel features based on the unsupervised machine learning [17].

A significant amount of work about combination of classifiers was done previously. Our approaches are motivated by the review of Tulyakov et al. [18].

3 Neural Networks and Combination

3.1 Individual Nets

We use two individual neural nets with different activation functions (*sigmoid* and *softmax*) in the output layer. Their topologies are briefly presented in the following two sections.

Feed-forward Deep Neural Network (FDNN) We use a Multi-Layer Perceptron (MLP) with two hidden layers[2]. As the input of our network we use the simple bag of words (BoW) which is a binary vector where value 1 means that the word with a given index is present in the document. The size of this vector depends on the size of the dictionary which is limited by N most frequent words which defines the size of the input layer. The first hidden layer has 1024 while the second one has 512 nodes. This configuration was set based on the experimental results. The output layer has the size equal to the number of categories $|C|$. To handle the multi-label classification, we threshold the values of nodes in the output layer. Only the labels with values larger than a given threshold are assigned to the document.

Convolutional Neural Network (CNN) The input is a sequence of words in the document. We use the same dictionary as in the previous approach. The words are then represented by the indexes into the dictionary. The architecture of our network (see Figure 1) is motivated by Kim in [8]. However, based on our preliminary experiments, we used only one-dimensional (1D) convolutional kernels instead of the combination of several sizes of 2D kernels. The input of our network is a vector of word indexes of the length L where L is the number of words used for document representation. The issue of the variable document size is solved by setting a fixed value (longer documents are shortened and the shorter ones padded). The second layer is an embedding layer which represents each input word as a vector of a given length. The document is thus represented as a matrix with L rows and EMB columns where EMB is the length of the embedding vectors. The third layer is the convolutional one. We use N_C convolution kernels of the size $K \times 1$ which means we do 1D convolution over one position in the embedding vector over K input words. The following layer performs max-pooling over the length $L-K+1$ resulting in N_C $1 \times EMB$ vectors.

[2] We have also experimented with an MLP with one hidden layer with lower accuracy.

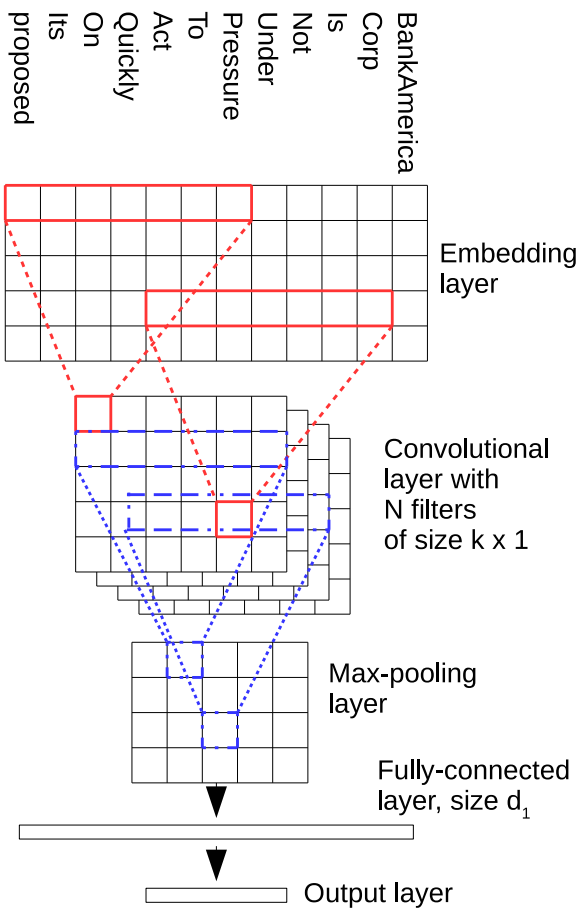

Figure 1: CNN architecture

The output of this layer is then flattened and connected with the output layer containing $|C|$ nodes. The final result is, as in the previous case, obtained by the thresholding of the network outputs.

3.2 Combination

We consider that the different nets keep some complementary information which can compensate recognition errors. We also assume that similar network topology with different activation functions can bring some different information and thus that all nets should have its particular impact for the final classification. Therefore, we consider all the nets as the different classifiers which will be further combined.

Two types of combination will be evaluated and compared. The first group does not need any training phase, while the second one learns a classifier.

Unsupervised Combination The first combination method compensates the errors of individual classifiers by computing the average value from the inputs. This value is thresholded subsequently to obtain the final classification result. This method is called hereafter *Averaged thresholding*.

The second combination approach first thresholds the scores of all individual classifiers. Then, the final classification output is given as an agreement of the majority of the classifiers. We call this method as *Majority voting with thresholding*

Supervised Combination We use another neural network of type multi-layer perceptron to combine the results. This network has three layers: $n \times |C|$ inputs, hidden layer with 512 nodes and the output layer composed of $|C|$ neurons (number of categories to classify). n value is the number of the nets to combine. This configuration was set experimentally. We also evaluate and compare, as in the case of the individual classifiers, two different activation functions: *sigmoid* and *softmax*. These combination approaches are hereafter called *FNN with sigmoid* and *FNN with softmax*. According to the previous experiments with neural nets on multi-label classification, we assume better results of this net with sigmoid activation (see first part of Table 1).

4 Experiments

In this section we first describe the corpora that we used for evaluation of our methods. Then, we describe the performed experiments and the final results.

4.1 Tools and Corpora

For implementation of all neural nets we used Keras toolkit [19] which is based on the Theano deep learning library [20]. It has been chosen mainly because of good performance and our previous experience with this tool. All experiments were computed on GPU to achieve reasonable computation times.

4.2 Czech ČTK Corpus

For the following experiments we used first the Czech ČTK corpus. This corpus contains 2,974,040 words belonging to 11,955 documents. The documents are annotated from a set of 60 categories as for instance agriculture, weather, politics or sport out of which we used 37 most frequent ones. The category reduction was done to allow comparison with previously reported results on this corpus where the same set of 37 categories was used. We have further created a development set which is composed of 500 randomly chosen samples removed from the entire corpus. Figure 2 illustrates the distribution of the documents depending on the number of labels. Figure 3 shows the distribution of the document lengths (in word tokens). This corpus is freely available for research purposes at `http://home.zcu.cz/~pkral/sw/`.

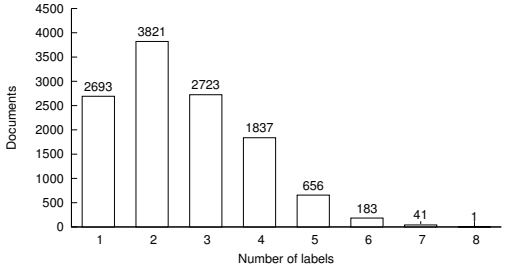

Figure 2: Distribution of documents depending on the number of labels assigned to the documents

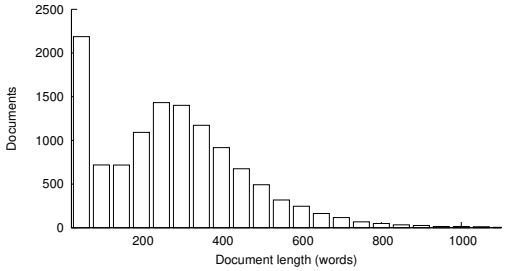

Figure 3: Distribution of the document lengths

We use the five-folds cross validation procedure for all experiments on this corpus. The optimal value of the threshold is determined on the development set. For evaluation of the multi-label document classification results, we use the standard recall, precision and F-measure (*F1*) metrics [21]. The values are micro-averaged.

Reuters-21578 English Corpus The Reuters-21578[3] corpus is a collection of 21,578 documents. This corpus is used to compare our approaches with the state of the art. As suggested by many authors, the training part is composed of 7769 documents, while 3019 documents are reserved for testing. The number of possible categories is 90 and average label/document number is 1.23.

4.3 Results of the Individual Nets

The first experiment (see Table 1) shows the results of the individual neural nets with sigmoid and softmax activation functions against the baseline approach proposed by Brychcín et al. [17]. These nets will be further referenced by the method number.

This table demonstrates very good classification performance of both individual nets and that the classification results are very close to each other and comparable. This table also shows that softmax activation function is slightly better for FDNN, while sigmoid activation function gives significantly better results for CNN.

Another interesting fact regarding to these results is that the approaches no. 1 - 3 have comparable precision and recall, while the best performing method no. 4 has significantly better precision than recall ($\Delta \sim 4\%$).

This table further shows that three individual neural networks outperform the baseline approach.

Table 1: Results of the individual nets with sigmoid and softmax activation functions against the baseline approach

No.	Network/activation		Prec.	Recall	F1 [%]
1.	FDNN	softmax	84.4	82.1	83.3
2.		sigmoid	83.0	81.2	82.1
3.	CNN	softmax	80.6	80.8	80.7
4.		sigmoid	86.3	81.9	**84.1**
	Baseline [17]		89.0	75.6	81.7

Error Analysis To confirm the potential benefits of the combination we analyze the errors of the individual nets. As already stated, we assume that different classifiers retain different information and thus they should bring different types of errors which could be compensated by a combination. Following analysis shows the numbers of incorrectly identified documents for two categories. We present the numbers of errors for all individual classifiers and compare it with the combination of all classifiers.

The upper part of Figure 4 is focused on the most frequent class - *politics*. The graph shows that the numbers of errors produced by the individual nets are comparable. However, the networks make errors on different documents and only few ones (384 from 2221 are common for all the nets.

The lower part of Figure 4 is concentrated on the less frequent class - *chemical industry*. This analysis demonstrates that the performances of the different nets significantly differ, the sigmoid activation function is substantially better than the softmax and the different nets provide also different types of errors. The number of the common errors is 49 (from 232 in total).

To conclude, both analysis clearly confirm our assumption that the combination should be beneficial for improvement of the results of the individual nets.

4.4 Results of Unsupervised Combinations

The second experiment shows (see Table 2) the results of *Averaged thresholding* method. These results confirm our assumption that the different nets keep complementary information and that it is useful to combine them. This experiment further shows that the combination of the nets with lower scores (particularly with net no. 2) can degrade the final classification score (e.g. combination 1 & 2 vs. individual net no. 1).

Another interesting, somewhat surprising, observation is that the CNN with the lowest classification accuracy can have some positive impact to the final classification

[3] http://www.daviddlewis.com/resources/testcollections/reuters21578/

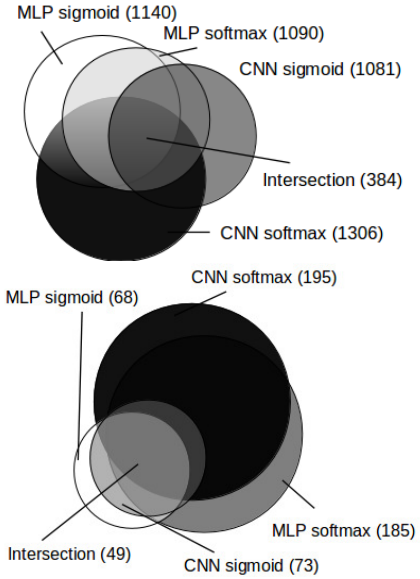

Figure 4: Error analysis of the individual nets for the most frequent (top, *politics*) and for the less frequent (bottom, *chemical industry*) classes, numbers of incorrectly identified documents in brackets

Table 2: Combinations of nets by *Averaged thresholding*

Net combi.	Precision	Recall	F1 [%]
1 & 2	83.0	82.4	82.7
1 & 3	83.2	84.6	83.9
1 & 4	85.7	84.3	85.0
2 & 3	86.2	79.6	82.8
2 & 4	84.9	83.5	84.2
3 & 4	87.3	81.7	84.4
1 & 2 & 3	84.8	81.9	83.3
1 & 2 & 4	90.1	79.6	84.5
1 & 3 & 4	86.7	83.5	**85.1**
2 & 3 & 4	89.3	80.5	84.6
1 & 2 & 3 & 4	89.7	80.5	84.9

(e.g. combination 1 & 3). However, the FDNN no. 2 (with significantly better results) brings only very small positive impact to any combination.

The next experiment which is depicted in Table 3 deals with the results of the second unsupervised combination method, *Majority voting with thresholding*. Note, that we consider an agreement of at least one half of the classifiers to obtain unambiguous results. Therefore, we evaluated the combinations of at least three networks.

This table shows that this combination approach brings also positive impact to document classification and the results of both methods are comparable. However, from the point of view of the contribution of the individual nets, the net no. 2 contributes better for the final results as in the previous case.

Table 3: Combinations of the nets by *Majority voting with thresholding*

Net combi.	Precision	Recall	F1 [%]
1 & 2 & 3	86.1	82.9	84.6
1 & 2 & 4	87.5	82.6	**85.0**
1 & 3 & 4	86.5	82.9	84.6
2 & 3 & 4	86.9	82.7	84.8
1 & 2 & 3 & 4	84.1	85.7	84.9

4.5 Results of Supervised Combinations

The following experiments show the results of the supervised combination method with an FNN (see Sec 3.2). We have evaluated and compared the nets with both sigmoid (see Table 4) and softmax (see Table 5) activation functions.

These tables show that these combinations have also positive impact on the classification and that sigmoid activation function brings better results than softmax. This

Table 4: Combinations of the nets by *FNN with sigmoid*

Net combi.	Precision	Recall	F1 [%]
1 & 2	86.1	82.1	84.1
1 & 3	87.1	81.5	84.2
1 & 4	88.4	81.9	85.0
2 & 3	86.6	81.4	83.9
2 & 4	87.7	82.0	84.7
3 & 4	89.3	80.0	84.4
1 & 2 & 3	86.9	82.4	84.6
1 & 2 & 4	87.9	82.8	85.3
1 & 3 & 4	88.2	82.5	85.2
2 & 3 & 4	87.9	82.2	85.0
1 & 2 & 3 & 4	88.0	82.8	**85.3**

Table 5: Combinations of the nets by *FNN with softmax*

Net combi.	Precision	Recall	F1 [%]
1 & 2	85.3	81.6	83.4
1 & 3	85.4	81.8	83.6
1 & 4	86.3	82.6	84.4
2 & 3	85.4	80.9	83.1
2 & 4	86.1	82.0	84.0
3 & 4	86.7	81.3	83.9
1 & 2 & 3	85.0	82.7	83.9
1 & 2 & 4	85.7	83.2	84.4
1 & 3 & 4	85.8	83.3	84.5
2 & 3 & 4	85.6	82.9	84.3
1 & 2 & 3 & 4	85.7	83.6	**84.6**

is a similar behaviour as in the case of the individual nets. Moreover, as supposed, this supervised combination slightly outperforms both previously described unsupervised methods.

4.6 Final Results Analysis

Finally, we analyze the results for the different document types. The main criterion was the number of the document labels. We assume that this number will play an important role for classification and intuitively, the documents with less labels will be easier to classify. We thus divided the documents into five distinct classes according to the number of labels (i.e. the documents with one, two, three and four labels and the remaining documents). Then, we tried to determine an optimal threshold for every class and report the F-measure. This value is compared to the results obtained with *global* threshold identified previously (one threshold for all documents).

The results of this analysis are shown in Figure 5. We have chosen two representative cases to analyze, the individual *FDNN with softmax* (left side) and the combination by *Averaged thresholding* method (right side). The adaptive threshold means that the threshold is optimized for each group of documents separately. The fixed threshold is the one that was optimized on the development set. This figure confirms our assumption. The best classification results are for the documents with one label and then they decrease. Moreover, this analysis shows that this number plays a crucial role for document classification for all cases. Hypothetically, if we could determine the number of labels for a particular document before the thresholding, we could improve the final F-measure by 1.5%.

4.7 Results on English Corpus

This experiment shows results of our methods on the frequently used Reuters-21578 corpus. We present the results on English dataset mainly for comparison with other state-of-the-art methods while we cannot provide such comparison on Czech data. Table 6 shows the performance of proposed models on the benchmark Reuters-21578 dataset. The bottom part of the table provides comparison with other state-of-the-art methods.

5 Conclusions and Future Work

In this paper, we have used several combination methods to improve the results of individual neural nets for multi-label document classification of Czech text documents. We have also presented the results of our methods on a standard English corpus. We have compared several popular (unsupervised and also supervised) combination methods.

[1] Approach proposed by Zhang et al. [12] and used with ReLU activation, AdaGrad and dropout.

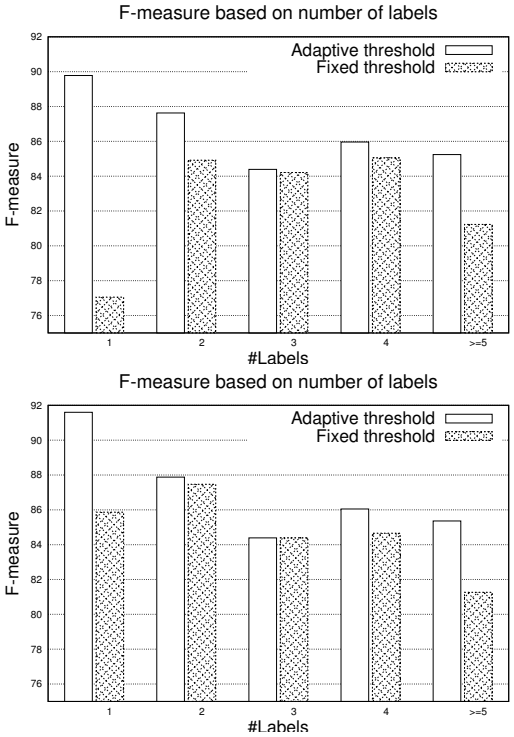

Figure 5: F-measure according to the number of labels for adaptive and fixed thresholds, the upper graph shows the results for MLP with softmax while the lower one is for the combination of all nets

Table 6: Results on the Reuters-21578 dataset

Method	**Precision**	**Recall**	**F1 [%]**
MLP/softmax	89.08	80.6	85.0
MLP/sigmoid	89.6	82.7	86.0
CNN/softmax	87.8	84.1	85.9
CNN/sigmoid	89.4	81.3	85.2
Supervised combi	91.4	84.1	87.6
NN_{AD} [1]	90.4	83.4	86.8
$BP-MLL_{TAD}$ [1]	84.2	84.2	84.2
BR_R [22]	89.8	86.0	87.9

The experimental results have confirmed our assumption that the different nets keep different information. Therefore, it is useful to combine them to improve the classification score of the individual nets. We have also proved that the thresholding is a good method to assign the document labels of multi-label classification. We have further shown that the results of all the approaches are comparable. However, the best combination method is the supervised one which uses an FNN with sigmoid activation function. The F-measure on Czech is 85.3% while the best result for English is 87.6%. Results on both languages are thus at least comparable with the state of the art.

One perspective for further work is to improve the com-

bination methods while the error analysis has shown that there is still some room for improvement. We have also shown that knowing the number of classes could improve the result. Another perspective is thus to build a classifier with thresholds dependent on the number of labels.

Acknowledgements

This work has been supported by the project LO1506 of the Czech Ministry of Education, Youth and Sports. We also would like to thank the Czech New Agency (ČTK) for support and for providing the data.

References

[1] Nam, J., Kim, J., Mencía, E.L., Gurevych, I., Fürnkranz, J.: Large-scale multi-label text classification—revisiting neural networks. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer (2014) 437–452

[2] Lenc, L., Král, P.: Deep neural networks for czech multi-label document classification. CoRR **abs/1701.03849** (2017)

[3] Della Pietra, S., Della Pietra, V., Lafferty, J.: Inducing features of random fields. IEEE Transactions on Pattern Analysis and Machine Intelligence **19**(4) (1997) 380–393

[4] Yang, Y., Pedersen, J.O.: A comparative study on feature selection in text categorization. In: Proceedings of the Fourteenth International Conference on Machine Learning. ICML '97, San Francisco, CA, USA, Morgan Kaufmann Publishers Inc. (1997) 412–420

[5] Ramage, D., Hall, D., Nallapati, R., Manning, C.D.: Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora. In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1. EMNLP '09, Stroudsburg, PA, USA, Association for Computational Linguistics (2009) 248–256

[6] Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., Kuksa, P.: Natural language processing (almost) from scratch. The Journal of Machine Learning Research **12** (2011) 2493–2537

[7] Zhang, X., LeCun, Y.: Text understanding from scratch. arXiv preprint arXiv:1502.01710 (2015)

[8] Kim, Y.: Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882 (2014)

[9] Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. In: Proceedings of Workshop at ICLR. (2013)

[10] Lai, S., Xu, L., Liu, K., Zhao, J.: Recurrent convolutional neural networks for text classification. (2015)

[11] Manevitz, L., Yousef, M.: One-class document classification via neural networks. Neurocomputing **70**(7-9) (2007) 1466–1481

[12] Zhang, M.L., Zhou, Z.H.: Multilabel neural networks with applications to functional genomics and text categorization. Knowledge and Data Engineering, IEEE Transactions on **18**(10) (2006) 1338–1351

[13] Nair, V., Hinton, G.E.: Rectified linear units improve restricted boltzmann machines. In: Proceedings of the 27th international conference on machine learning (ICML-10). (2010) 807–814

[14] Srivastava, N., Hinton, G.E., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. Journal of Machine Learning Research **15**(1) (2014) 1929–1958

[15] Kurata, G., Xiang, B., Zhou, B.: Improved neural network-based multi-label classification with better initialization leveraging label co-occurrence. In: Proceedings of NAACL-HLT. (2016) 521–526

[16] Yang, Y., Gopal, S.: Multilabel classification with meta-level features in a learning-to-rank framework. Machine Learning **88**(1-2) (2012) 47–68

[17] Brychcín, T., Král, P.: Novel unsupervised features for Czech multi-label document classification. In: 13th Mexican International Conference on Artificial Intelligence (MICAI 2014), Tuxtla Gutierrez, Chiapas, Mexic, Springer (16-22 November 2014) 70–79

[18] Tulyakov, S., Jaeger, S., Govindaraju, V., Doermann, D.: Review of classifier combination methods. In: Machine Learning in Document Analysis and Recognition. Springer (2008) 361–386

[19] Chollet, F.: keras. `https://github.com/fchollet/keras` (2015)

[20] Bergstra, J., Breuleux, O., Bastien, F., Lamblin, P., Pascanu, R., Desjardins, G., Turian, J., Warde-Farley, D., Bengio, Y.: Theano: a cpu and gpu math expression compiler. In: Proceedings of the Python for scientific computing conference (SciPy). Volume 4., Austin, TX (2010) 3

[21] Powers, D.: Evaluation: From precision, recall and f-measure to roc., informedness, markedness & correlation. Journal of Machine Learning Technologies **2**(1) (2011) 37–63

[22] Rubin, T.N., Chambers, A., Smyth, P., Steyvers, M.: Statistical topic models for multi-label document classification. Machine learning **88**(1-2) (2012) 157–208