



**GERMAN NEWS ARTICLE CLASSIFICATION :
A MULTICHANNEL CNN APPROACH**

Shantipriya Parida^a Petr Motlicek
Satya Ranjan Dash

Idiap-RR-09-2020

MAY 2020

^aIdiap Research Institute

German News Article Classification: A Multichannel CNN Approach

Shantipriya Parida¹, Petr Motlicek², and Satya Ranjan Dash^{3*}

¹Idiap Research Institute, Martigny Switzerland.

{shantipriya.parida}@idiap.ch

²Idiap Research Institute, Martigny Switzerland.

{petr.motlicek}@idiap.ch

³School of Computer Applications, KIIT University, Odisha, INDIA. {sdashfca}@kiit.ac.in

Abstract. Nowadays, more and more people are gaining interest in news, and social media networks, and are also sharing their opinions freely in different languages. Such kind of activities leads to interesting topics of research that scientists are working on. Considering news, it must be classified and easily accessible by the users for the information of their interest. In comparison to traditional machine learning techniques, deep learning approaches have achieved surpassing results on natural language processing tasks. Convolutions neural networks (CNNs) have shown promising performance, which extracts n-grams as features to represent the input.

In this work, we build a multi-channel CNN for German news article classification. The model can classify different categories of news articles with an accuracy of 99.2% on training and 81.4% on the test dataset. We also perform a comparative study with single-channel CNN and have found that the multi-channel approach outperforms the single-channel by +6.3% absolute on the test set.

Keywords: Convolution neural network · Multi-channel CNN · News article classification.

1 Introduction

The objective of text classification is to automatically classify documents by assigning one or more predefined tags/categories based on their content.

In the past few years, deep learning methods were found to be effective for natural language processing (NLP) related tasks. In several domains related to NLP applications including text classification, deep learning techniques better results as compared to traditional machine learning approaches [1] - [2]. Two main types of

deep neural network architectures that are widely explored for handling various NLP tasks and provide competitive results are CNN and recurrent neural network (RNN) [3]. CNNs are generally used in computer vision and have shown to achieve better performance on text

classification tasks [4] - [6]. The convolutional layers of CNN are extracting the features using geometrically fixed filters and can be regarded as an implementation of the n-gram language model [7]. CNN generally outperformed RNN in capturing high-level features in short text [8].

A series of experiments with CNN for the task of sentence classification built on top of “word2vec” (with little hyperparameter tuning) has shown excellent results over multiple benchmarks [4]. Even one-layer CNN performs excellently for sentence classification [9]. Deep CNN architectures operating directly on character level representation/input have shown an improvement in text classification tasks [10], however for inputs such as patent text containing full of technical and legal terminologies the performance may not be adequate [11]. The paper is organized as follows. Section 1 describes related work on text classification. Section 2 explains the proposed model architecture. Section 3 explains the dataset used in our experiment. Section 4 explains the experimental settings: pre-processing and hyper-parameters. Section 5 provides evaluation results with analysis and discussion. The paper is concluded in Section 6.

2 Model

When applying CNN to text rather than an image, we have a 1-dimensional array representing text and the architecture changed to 1D convolutional-and-pooling operations [12].

We define single- and multi-channel models. The multi-channel CNN is a combination of many versions of the standard model. The standard model contains an embedding layer as input, next to a one-dimensional CNN, a pooling layer followed by a prediction output layer with kernels of different sizes. This enables the text to be processed for various n-grams (groups of words) at a time, while the model learns how to best integrate these interpretations [4].

We define a model containing 3 input channels for processing different n-grams (4-grams, 6-grams, and 8-grams) of the input text as shown in Figure 1. Each of the channels consists of the following elements:

- *Input layer* defines the input sequences length.
- *Embedding layer* is set to the vocabulary size, and dimension 100 to store real-valued representations.
- *One-dimensional convolutional layer* having a filter size of 32 and the kernel size equal to the count of words read at once.
- *Max pooling layer* consolidates the output from the convolutional layer.

- *Flatten layer* maps from 3-dimensional output to 2-dimensional output, required for concatenation.

The output obtained from the three channels is concatenated into a single vector. Then it is processed by a dense and output layer respectively. The architecture of the multi-channel network is shown in Figure 2.

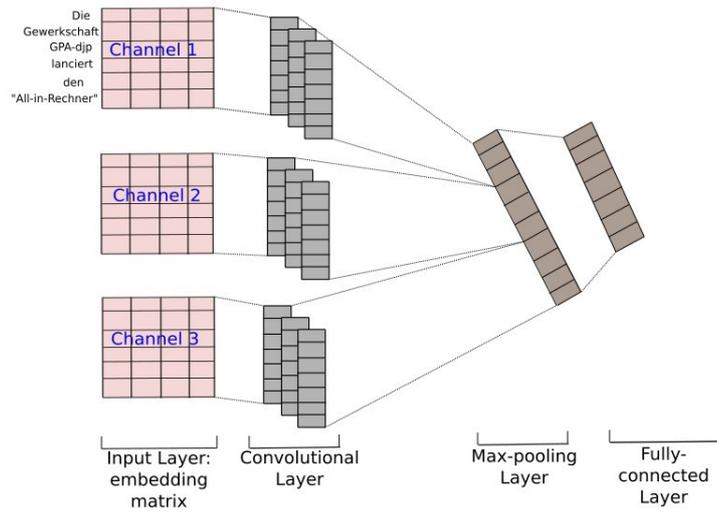


Fig.1. Model architecture with three channels.

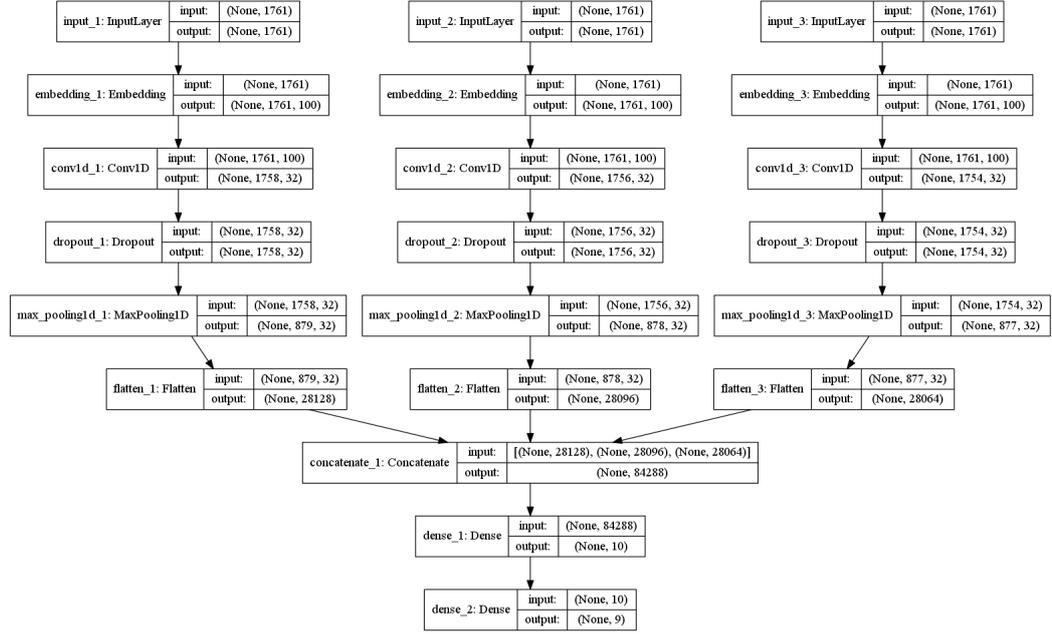


Fig.2. The proposed multi-channel CNN architecture.

2.1 Regularization

We use dropout for regularization which operates on randomly dropping out a variable proportion of the hidden units [13].

3 Dataset

We use 10k German News Articles Dataset (10kGNAD)¹ in our experiment. It has 10,273 German-language news articles collected by an Austrian online newspaper, categorized into 9 topics [14]. The articles and per-category distributions are shown in Figure 3. As can be seen, the class distribution of the 10kGNAD is not balanced.

¹ <https://github.com/tblock/10kGNAD>

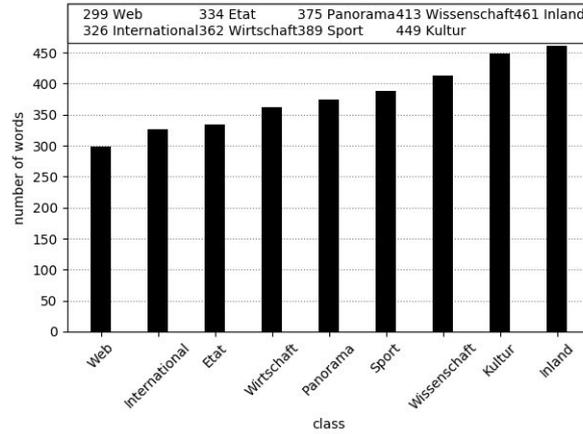


Fig.3. Articles per class. [Image is taken from 10k German News Articles Dataset website. <https://tblock.github.io/10kGNAD>]

4 Experimental Setup

This section describes the proposed setup for the experiments.

4.1 Pre-processing

The downloaded 10kGNAD training and test datasets contain articles along with their respective categories (i.e. Web, Panorama, etc.). We mapped each article to one of 9 (0-8) categories for the training and test dataset. The statistics of the dataset are shown in Table 1.

Table 1. Statistics of the experimental (10kGNAD) data.

Dataset	#Articles	#Category
Train	9245	9
Test	1028	9

During pre-processing, we performed the following text-normalization operations:

- Divide the tokens based on the occurrence of the white space.
- Removing of punctuation present in the words.
- Removing words which do not contain fully alphabetical characters.
- Removing German stop words.
- Truncate words having a length of ≤ 1 character.

4.2 Hyper-parameters

The configuration parameters are shown in Table 2. The parameters are similar for the single- and multi-channel experiments. We manually explored and tuned the hyper-parameters such as drop-out (for avoiding over-fitting), and batch size (for improving performance). The training also follows an early termination if validation loss does not improve for 3 epochs [15].

Table 2. Configuration parameters: “Relu” refers to the rectified linear unit, a common CNN activation function [16].

Description	Values
Filter	32
Feature maps	100
Activation function	Relu
Pooling	1-max pooling
Dropout rate	0.5
Loss	categorical cross-entropy
Optimizer	Adam
Epoch	15
Batch size	16

5 Evaluation and Discussion

We evaluate the proposed text classification system on the test dataset (10kGNAD) and shown the results in Table 3. In addition to classification accuracies, we also show the training and test accuracy and loss w.r.t. each epoch of training in Figure 4 and Figure 5. The maximum article length (in words) is 1,761 and the total number of unique words in the training dataset is 197,762. We observe a performance improvement comparing multi-channel CNN with a single-channel CNN. We also experimented with different kernel sizes (2-gram, 4-gram, and 6-grams) for multi-channel CNN, while not observing any further improvement. As our experiments are limited to news articles German dataset, the performance of our model for other domains needs investigation.

Table 3. Evaluation results of text-classification on the test dataset (10kGNAD).

Model	Dataset	Accuracy
CNN-single-channel	Train	99.5
	Test	75.1
CNN-multi-channel	Train	99.2
	Test	81.4

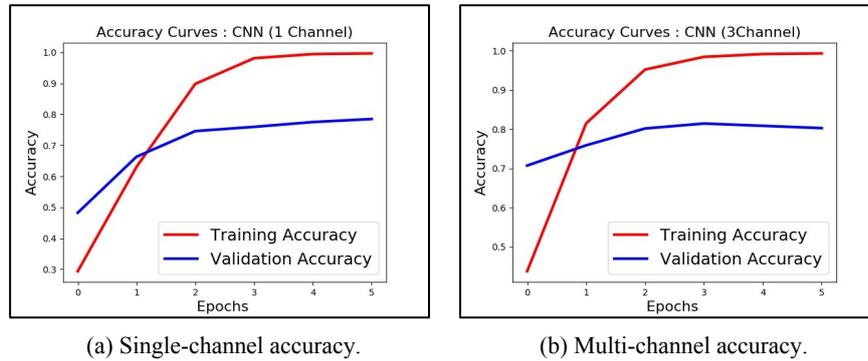


Fig.4. Classification accuracies of news-article categories on training and test datasets.

6 Conclusion

We propose a multi-channel CNN approach for German news article classification. We found that CNN's can well capture the textual feature information for classifying text in multilingual scenarios [17]. The proposed model uses multiple parallel CNN which reads the German news articles using different n-gram sizes (4-grams, 6-grams, and 8-grams) and the multi-channel CNN boosts text classification accuracy and able to classify different news categories for the news articles better than single-channel CNN. Even, with a dataset size of 10K, the proposed model achieves good validation accuracy.

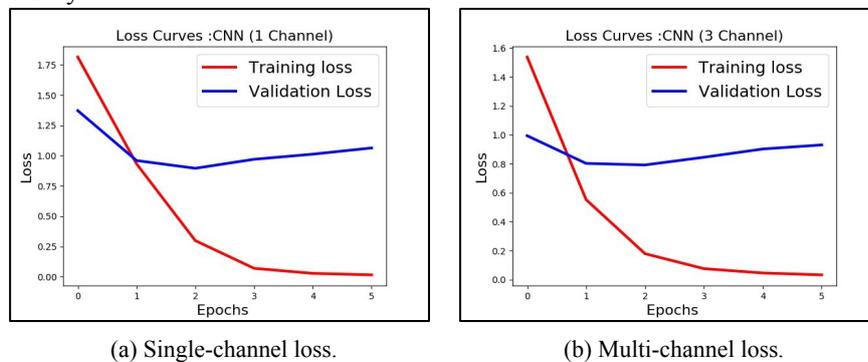


Fig.5. Loss computed using categorical cross-entropy function measured on training and test datasets.

As the next step, we plan to investigate more on: i) applying different configurations (e.g. different n-grams, varying channels, deeper networks, varying

dropout rates). ii) experiment with our model with other languages and across domains. iii) compare the performance of our model with other supervised and hybrid approaches [18] - [22].

Acknowledgments

The work was supported by an innovation project (under an InnoSuisse grant) oriented to improve the automatic speech recognition and natural language understanding technologies for German. Title: "SM2: Extracting Semantic Meaning from Spoken Material" funding application no. 29814.1 IP-ICT. And also supported by the EU H2020 project "Real-time network, text, and speaker analytics for combating organized crime" (ROXANNE), grant agreement: 833635.

References

1. Lai, S., Xu, L., Liu, K., Zhao, J.: Recurrent convolutional neural networks for text classification. In: Twenty-ninth AAAI conference on artificial intelligence (2015)
2. Maas, A.L., Hannun, A.Y., Ng, A.Y.: Rectifier nonlinearities improve neural network acoustic models. In: Proc. icml. vol. 30, p. 3 (2013)
3. Hughes, M., Li, I., Kotoulas, S., Suzumura, T.: Medical text classification using convolutional neural networks. *Studies in health technology and informatics* 235,246 (2017)
4. Xiao, L., Zhang, H., Chen, W., Wang, Y., Jin, Y.: Transformable convolutional neural network for text classification. In: IJCAI. pp. 4496–4502 (2018)
5. Li, W., Liu, P., Zhang, Q., Liu, W.: An improved approach for text sentiment classification based on a deep neural network via a sentiment attention mechanism. *Future Internet* 11(4), 96 (2019)
6. Goldberg, Y.: A primer on neural network models for natural language processing. *Journal of Artificial Intelligence Research* 57, 345–420 (2016)
7. Hu, J., Li, S., Hu, J., Yang, G.: A hierarchical feature extraction model for multilabel mechanical patent classification. *Sustainability* 10(1), 219 (2018)
8. Hsu, S.T., Moon, C., Jones, P., Samatova, N.: A hybrid CNN-RNN alignment model for phrase-aware sentence classification. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers. pp. 443–449 (2017)
9. Conneau, A., Schwenk, H., Barrault, L., Lecun, Y.: Very deep convolutional networks for text classification. In: European Chapter of the Association for Computational Linguistics EACL'17 (2017)
10. Yoon, J., Kim, H.: Multi-channel lexicon integrated CNN-BiLSTM models for sentiment analysis. In: Proceedings of the 29th Conference on Computational Linguistics and Speech Processing (ROCLING 2017). pp. 244–253 (2017)

11. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research* 15(1), 1929–1958 (2014)
12. Shimura, K., Li, J., Fukumoto, F.: HFT-CNN: Learning hierarchical category structure for multi-label short text categorization. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. pp. 811–816 (2018)
13. Wang, S., Huang, M., Deng, Z.: Densely connected CNN with multi-scale feature attention for text classification. In: *IJCAI*. pp. 4468–4474 (2018)
14. Schabus, D., Skowron, M., Trapp, M.: One million posts: A data set of german online discussions. In: *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*. pp. 1241–1244. Tokyo, Japan (Aug 2017). <https://doi.org/10.1145/3077136.3080711>
15. Prechelt, L.: Early stopping-but when? In: *Neural Networks: Tricks of the trade*, pp. 55–69. Springer (1998)
16. Zhang, Y., Wallace, B.: A sensitivity analysis of (and practitioners’ guide to) convolutional neural networks for sentence classification. In: *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. pp. 253–263 (2017)
17. Mahata, D., Friedrichs, J., Ratn Shah, R., et al.: # phrama covigilance-exploring deep learning techniques for identifying mentions of medication intake from twitter. *arXiv preprint arXiv:1805.06375* (2018)
18. Yin, W., Kann, K., Yu, M., Schu̇tze, H.: Comparative study of CNN and RNN for natural language processing. *arXiv preprint arXiv:1702.01923* (2017)
19. Jacovi, A., Shalom, O.S., Goldberg, Y.: Understanding convolutional neural networks for text classification. *EMNLP 2018* p. 56 (2018)
20. Guggilla, C., Miller, T., Gurevych, I.: CNN-and LSTM-based claim classification in online user comments. In: *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. pp. 2740–2751 (2016)
21. Kim, Y.: Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882* (2014)
22. Kowsari, K., Jafari Meimandi, K., Heidarysafa, M., Mendu, S., Barnes, L., Brown, D.: Text classification algorithms: A survey. *Information* 10(4), 150 (2019)