# Deep Learning based Pipeline with Multichannel Inputs for Patent Classification

Mustafa Sofean

*FIZ Karlsruhe, Hermann-von-Helmholtz-Platz 1. 76344 Eggenstein-Leopoldshafen*
*Mustafa.Sofean@fiz-karlsruhe.de*

**Abstract**

Patent document classification as groundwork has been a challenging task with no satisfactory performance for decades. In this work, we introduce a deep learning pipeline for automatic patent classification with multichannel inputs based on LSTM and word vector embeddings. Sophisticated text mining methods are used to extract the most important segments from patent texts, and a domain-specific pre-trained word embeddings model for the patent domain is developed; it was trained on a very large dataset of more than five million patents. A deep neural network model is trained with multichannel inputs namely embeddings of different segments of patent texts, and sparse linear input of different metadata. A series of patent classification experiments are conducted on different patent datasets, and the experimental results indicate that using the segments of patent texts as well as the metadata as multichannel inputs for a deep neural network model, achieves better performance than one input channel.

*Keywords:* Patent Analysis, Neural Network, Deep Learning, Patent Classification

## 1. Methods

Patent classification is a kind of knowledge management where documents are assigned into predefined categories. Due to the extremely complicated patent language and hierarchical patent classification scheme, many previous studies focused only on whole texts of patent or some general sections such as title, abstract, detailed description and claims [2] [1]. They did not consider the most important sections like background, technical field, summary, and independent claims that need specific text mining tools to extract.

### *1.1. Semantic Structure of patent and Embeddings*

Efficient text mining services are used for semantic structuring of the patent texts [3]. The first service is used to structure the description part of patent text into structured segments such as the technical field, background, summary, and the embodiments [5]. The second service is able to automatically identify the complete claim hierarchy within patent texts [4]. In addition, a domain-specific word and phrase embeddings model is developed for the patent domain. The model is trained on more than five million patent documents and can be used for word/phrase similarity or patent analysis such as classification tasks.

### *1.2. Deep Learning based Pipeline Architecture*

Firstly, we extract the most important segments of patent texts which are title, abstract, technical field, background, summary, and the independent claim. For texts of each segment, a tokenization process is used for breaking the text into individual words, and the sequence length of each segment is set according to the maximum length of each. The deep learning architecture has two components: deep, and wide. It feed-forward neural networks with embeddings of each segment, and uses them as deep layers for deep neural network model, and the patent metadata on the other hand is used as a wide part for the model. Specifically, the architecture is described as follows: for the wide components of the model, we used one-hot representation for patent metadata features (such as inventors, citations, and assignees), these one-hot vectors are fed into separate sub-networks, and

*Published in "Proceedings of The 1**st Workshop on Patent Text Mining and Semantic Technologies (PatentSemTech 2019)**", Eds. L. Andersson, H. Aras, F. Piroi, A. Hanbury, PatentSemTech 2019, 12 September 2019, Karlsruhe, Germany. https://doi.org/10.34726/pst2019.5 |* 

*An extended version has been submitted to the World Patent Information Journal - Virtual Special Issue: Text Mining and Semantic Technologies in the Intellectual Property domain.*

Table 1: Evaluation Results. (TI:title, AB:abstract, TECHF: Technical Fields, BACK: Background, SUMM: Summary, IND_CLAM: Independent Claim, INVs: Inventors, and PAs: Patent Assignees)

| **Input** | **Accuracy** | **Precision** | **Recall** | **F1-score** |
|---|---|---|---|---|
| All texts of segments as one channel | 67% | 84% | 61% | 71% |
| TI, AB, TECHF , BACK, SUMM, and IND_CLAM as multichannels | **74%** | **92%** | **63%** | **75%** |
| TI and TECHF as multichannel inputs | 66% | 83% | 59% | 69% |
| TI and TECHF, INVs, and PAs as multichannel inputs | 68% | 85% | 61% | 71% |

at the end they are represented as deep networks. For the deep components of the model, deep layers are created for the most important patent text segments. These are sequential input to a Long Short-Term Memory (LSTM) network that takes the embeddings as inputs that are obtained by using a pre-trained word embeddings model to encode each segment texts into vectors, and then we feed them into LSTM layers. To avoid network overfitting and help network stability, additional layers are added for each input channel, dropout layer is used to drop out 30% of input in order to prevent neural networks from overfitting, and Batch normalization layer is used to normalize the input layer by adjusting and scaling the activations. The exponential linear unit (ELU) is used as activation function. Finally, we concatenated nine components which are text-based LSTM layers (title LSTM, abstract LSTM, technical field LSTM, background LSTM, summary LSTM, and independent claim LSTM), and metadata-based LSTM layers (inventors, assignees, and citations) into a final set of deep layers with dropout, batch normalization, and softmax activation function for multi-class and sigmoid for multi-label classification task.

### *1.3. Experimental Results*

The dataset in this work is extracted from databases of the European Patent Office (EPO) and the World Intellectual Property Organization (WIPO). All extracted patents contain the title, abstract, detailed description, claims, and at least one IPC label. The total number of extracted records in the dataset is about 1,915,308 patents filed between 1978 and 2016. The segmentation tools [3] [4] were used to extract the most important sections (technical field, background, summary of invention and independent claim from patent texts. All patent documents are classified into related subclass level of IPC, and we used four evaluation measures namely accuracy, precision, recall, and F1. A series of patent classification experiments are conducted on the dataset, and we also studied how the full text, different parts of a patent information, and their combination affect the classification performance. The evaluation results are shown in the table 1. The best performance we obtained is 74%, 92%, 63%, and 75% for accuracy, precision, recall, and F1, respectively. The result in this work indicates that using the segments of patent text as multichannel inputs improved the performance of patent classification in terms of all evaluation criteria.

## 2. Conclusion

In this work, we introduced a deep learning based pipeline for large-scale patent classification. Different parts of patent information are used as multichannel inputs for a Long Short-Term Memory (LSTM) that takes the both vectors (embeddings and one-hot) in order to learn a patent classification model. The experimental results indicated that using the segments of patent texts as well as the metadata as multichannel inputs for a deep neural network model, achieve a good performance.

## 3. References


[1] Assad Abbas, Limin Zhang, S.U.K., 2015. A literature review on the state-of-the-art in patent analysis. World Patent Information 37, 3–13.

[2] Juan Carlos Gomez, M.F.M., 2014. A survey of automated hierarchical classification of patents. Springer International Publishing .

[3] Mustafa Sofean, Hidir Aras, A.A., 2018. A workflow-based large-scale patent mining and analytics framework. Proceedings of 24th International Conference on Information and Software Technologies (ICIST); .

[4] Rene Hackl-Sommer, M.S., 2015. Patent claim structure recognition. Archives of Data Science .

[5] Sofean, M., 2017. Automatic segmentation of big data of patent texts. International Conference on Big Data Analytics and Knowledge Discovery. DaWaK , 343–351.