Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

ARTICLE TYPE

# Automated patent classification for crop protection via domain adaptation

Dimitrios Christofidellis*[1,3] | Marzena Maria Lehmann[2] | Torsten Luksch[2] | Marco Stenta[2] | Matteo Manica[1]

[1]IBM Research Europe
[2]Syngenta Crop Protection AG
[3]Queen's University Belfast

**Correspondence**
*Dimitrios Christofidellis Email: dchristofidellis01@qub.ac.uk

**Abstract**

Patents show how technology evolves in most scientific fields over time. The best way to use this valuable knowledge base is to use efficient and effective information retrieval and searches for related prior art. Patent classification, i.e., assigning a patent to one or more predefined categories, is a fundamental step towards synthesizing the information content of an invention. To this end, architectures based on Transformers, especially those derived from the BERT family have already been proposed in the literature and they have shown remarkable results by setting a new state-of-the-art performance for the classification task. Here, we study how domain adaptation can push the performance boundaries in patent classification by rigorously evaluating and implementing a collection of recent transfer learning techniques, e.g., domain-adaptive pretraining and adapters. Our analysis shows how leveraging these advancements enables the development of state-of-the-art models with increased precision, recall, and F1-score. We base our evaluation on both standard patent classification datasets derived from patent offices-defined code hierarchies and more practical real-world use-case scenarios containing labels from the agrochemical industrial domain. The application of these domain adapted techniques to patent classification in a multilingual setting is also examined and evaluated.

**KEYWORDS:**
patent classification, NLP, transformers, patent analysis, BERT, domain-adaption

## 1 | INTRODUCTION

Patents corpora[1,2] are a valuable resource that shows how technology evolves over time. This is documented by the volume of granted patents every year, e.g., over 300K patents per year in the last 6 years for the sole USPTO[3]. Monitoring this throughput is critical for capturing trends and developing domain-specific knowledge bases that can be used to organize information and accelerate the discovery process. This calls for the development of automated systems for information retrieval and patent literature search that can be easily adapted to fit the needs of different fields. The classification of patents is a critical step in the design of such systems. To this end, the World Intellectual Property Organization (WIPO) introduced the International Patent Classification (IPC), a hierarchical system of codes (language independent) that classify patents based on the different technological areas that they cover[4].

In this paper, we address the problem of classifying patents relying solely on the content of the inventions and using language models (LMs) and domain adaptation strategies. Domain adaptation in transformers is the preferred strategy for pushing the performance boundaries of transformer-based models in domains that are highly differentiated from the pretraining corpora[5]. We investigate the application of various domain adaptation strategies using several LMs based on Transformer models from the BERT family[6], including SciBERT[7], BERT-like models adapted in the patent domain using adaptive pretraining, BERT-like models fine-tuned using adapters[8] as well as combinations of the above. We evaluate the proposed approaches in terms of precision, recall, and F1-score under two different scenarios. Firstly, we rely on an existing baseline dataset including patents from USPTO. Then, we focus on a specific use case originated from the crop protection industry domain. Our analysis allowed us to identify fine-tuning recipes that ensure robust performance.

Our approach for patent classification outperforms the state-of-the-art. We depart from the standard solely IPC-based model evaluation by introducing an evaluation based on actual use-cases and labels that do not conform to the IPC hierarchy. Furthermore, we examine and evaluate the use of our domain adapted methodologies in a multilingual setup of patent classification. On top of that, we assess the effectiveness of such methods on classifying patents made available in 2021. Finally, we establish two new patent based BERT-like models, namely domain-adaptive patent BERT or dapBERT and domain-adaptive patent SciBERT or dapSciBERT, that can be leveraged for any NLP task related to the patent domain. We have made the code, the models and the dataset of our work available at https://github.com/GT4SD/domain-adaptive-patent-classifier.

## 2 | RELATED WORK

Numerous methods for patent classification have been introduced, as well as many baseline datasets[9]. Early attempts proposed k-Nearest Neighbor[10], support vector machine[10,11], Naive Bayes [10,11] or neural networks [12]. CNNs (Convolutional Neural Networks) and various word embeddings[13,14,15] have also been successfully combined for the task. Recent trends indicate an increased emphasis on fine-tuned pretrained LMs, with ULMFit[16] and BERT[17] based methods being the state of the art. PatentBERT[17] is the work that comes closest to our approach. However, while PatentBERT relies on an existing pretrained BERT model and performs patent classification using standard recipes for fine-tuning, we begin by adapting a BERT-like model to the patent domain before fine-tuning the classification task. This way, we ensure that the LM being used is domain-aware.

The content of an invention is critical when it is used as input for the patent classification method. The methods described previously act on different parts of the document. DeepPatent[13] utilizes titles and abstracts while PatentBERT[17] has been developed using claims and titles or abstracts. The full patent text including title, abstract, description, and claims has being evaluated in other attempts[14]. In general, the title and abstract sections are more informative than the full-text representation of the patent document[18]. Additionally, focusing exclusively on titles and abstracts has the advantage of being the two most easily accessible sections of a patent.

## 3 | PATENT CLASSIFICATION USING BERT-LIKE MODELS

A pretrained BERT model fine-tuned on labelled data for patent classification represents state-of-the-art performance in patent classification. While pretrained LMs have been shown to be more robust to out-of-distribution generalisation than previous models[19], they are still ill-equipped to deal with data that differ significantly from what was observed during pretraining. Patent corpora is a clear example: the unique syntax and vocabulary of patent applications may differ significantly from those used for pretraining. To address the distribution mismatch, we avoid performing an expensive and resource-intensive pretraining of a BERT-like model from scratch and instead examine four alternative approaches (i) adoption of a pretrained BERT-like model trained on corpora with similar vocabulary (vocabulary adaptation) (ii) domain-adaptive pretraining on the domain of interest, (iii) adapters[8] based fine-tuning and (iv) combination of the above options. Figure 1 depicts the 4 different approaches under investigation.

### 3.1 | Vocabulary adaptation

A domain-specific vocabulary is crucial to generate meaningful word embeddings, and it is thus essential in the majority of NLP applications. There is a strong correlation between poor NLP models performance on unfamiliar domains and the effects of out-of-vocabulary words[19]. However, including a significant number of domain-specific tokens in a BERT-like model requires the

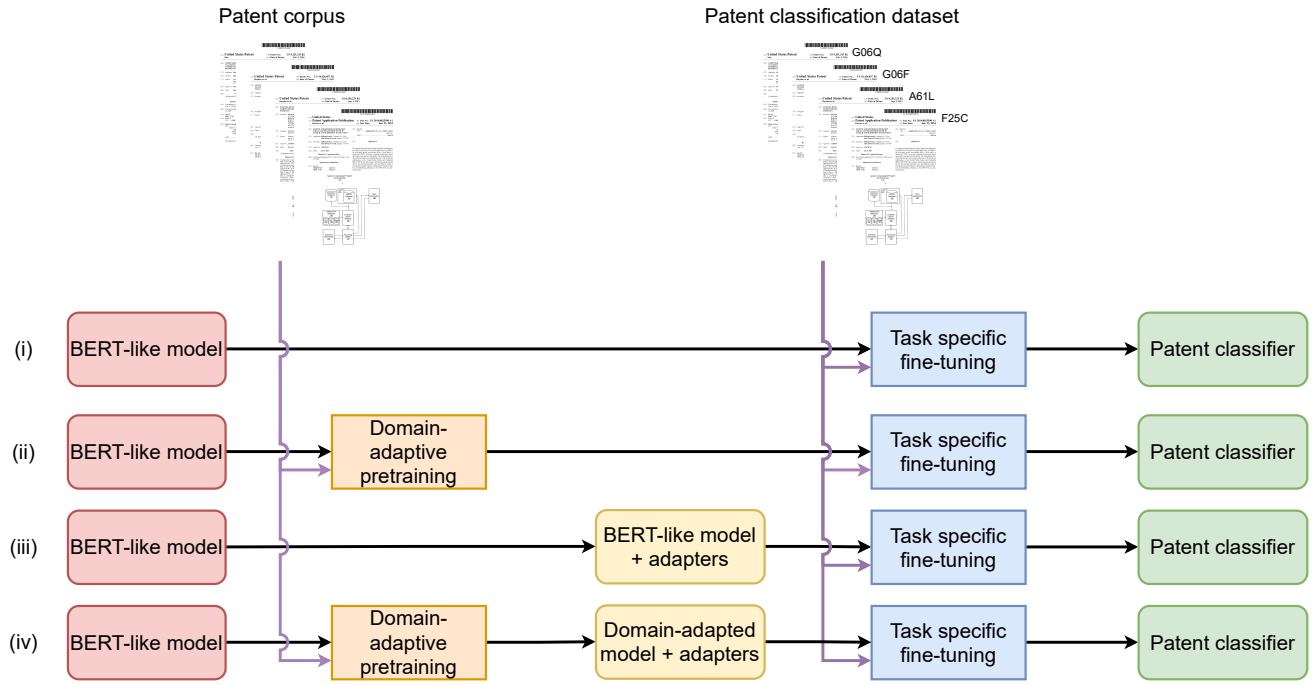



**FIGURE 1** Patent classification based on the 4 different approaches. Case (i) depicts the standard task specific fine-tuning approach that can be used leveraging any BERT-like model. Method of case (ii) performs a domain adaption of a BERT-like model prior to the task specific fine-tuning. Cases (iii) and (iv) utilize adapters without or with domain-adaptive pretraining respectively.

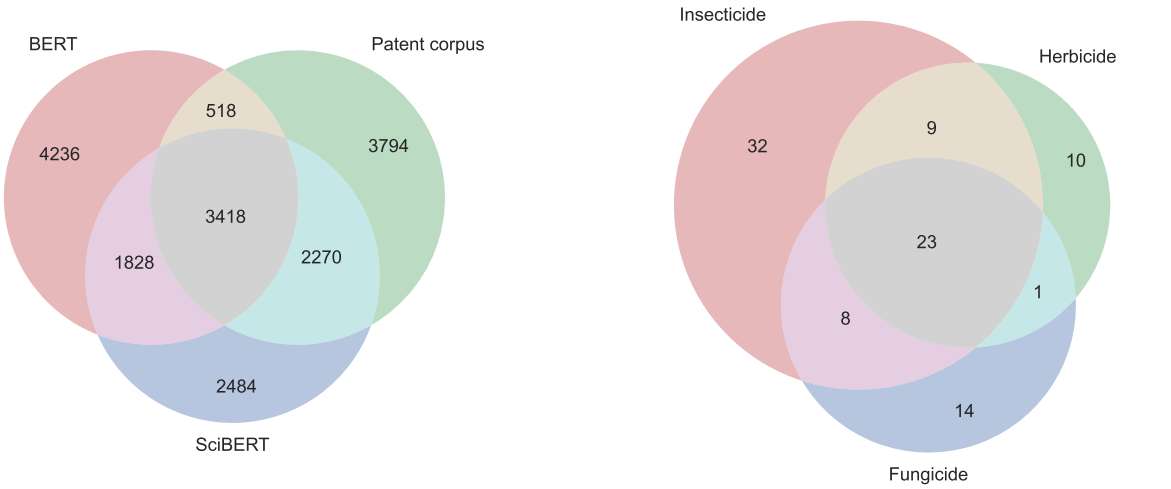
inclusion of the same number of uninitialized embedding vectors in the model that must be learned during training. Incorporating such a large number of embedding vectors is impractical in the standard fine-tuning scenario, and pre-training from scratch would be required to learn appropriate new token representations. As an alternative, we use the existing BERT variant that is closest to the patent domain as base for standard fine-tuning. In the following, we consider SciBERT[7]. SciBERT was trained on scientific documents, and while their structure and syntax do not resemble those of a patent, their vocabulary is more relevant to the domain of interest than a standard pretrained BERT. To demonstrate this, we analysed the vocabulary overlap of a patent corpus and corpora similar to those used in the trainings of BERT and SciBERT models. Specifically, we used 150K patent abstracts as the patent corpus, 150K texts from the BookCorpus[20] and English Wikipedia as the BERT corpus, and 150K texts from Semantic Scolar[21] as the SciBERT corpus. For each of them, we found the 10K most common words excluding stop-words, and then we examined the overlap of the respective sets. Figure 2a depicts the analysis of the vocabulary overlap resulting in a greater similarity between the patent corpus and SciBERT than BERT.

## 3.2 | Domain-adaptive pretraining

Domain-adaptive pretraining[5] involves fine-tuning the model on additional unlabelled data, originated from the domain of interest, using a pretraining objective prior to task-specific fine-tuning. This operation aims to shift a pretrained model towards the domain of interest and is achieved by performing just a few extra epochs of training on the domain-specific data in order to avoid degradation of the model's general language capabilities. Here, the training dataset consists of 10.000.000 patent abstracts downloaded from WIPO and USPTO.

## 3.3 | Adapters

Adapters[8] represent an alternative fine-tuning strategy that relies on optimizing a small set of additional newly initialized weights at every layer of the transformer. These weights are trained during fine-tuning, while the pretrained parameters of the transformer




**(a)** Vocabulary overlaps of the Patent, BERT and SciBERT corpora.

**(b)** IPC codes overlap between the patents that have been classified as insecticide, fungicide and herbicide.

**FIGURE 2** Overlap between different corpora vocabularies and patent labels using Venn diagrams.

model are frozen. This strategy has two significant advantages. To begin, by freezing the entire base model, we can significantly reduce the amount of computation required during training. Additionally, training many task-specific adapters using the same base model enables an efficient parameter sharing between the different tasks. In this case, the memory footprint of the application can be reduced because we only need to store the extra weights for each task, not the entire fine-tuned model.

### 3.4 | Combining methods

All of the methods outlined above can also be combined in a two-phase approach to improve performance. The first phase focuses on domain adaptation. The selected model is adapted to the patent domain by performing domain-adaptive pretraining. This phase produces a LM that is specialized for a patent corpus. This model is not tailored for patent classification and can be used for any downstream NLP task. In the second phase, we perform task-specific fine-tuning using a cross-entropy loss. At this point, we have two alternatives to perform the classification: utilizing adapters or following the standard approach of attaching a classification head as output in the existing architecture.

## 4 | RESULTS

We used a newly established patent dataset focusing on the crop industry domain and a standard USPTO based dataset for the evaluation of the different models. The former data set reveals the performance of the models in actual use cases belonging to the agrochemical domain. The latter dataset serves as a standard baseline in the IPC classification task, the most documented case in the literature.

As baselines, we used the CNN+embedding based models presented in Roudsari et al[15] and a standard fine-tuned BERT model similar to Lee et al[17]. We compared them with SciBERT, as a vocabulary adaptation approach, dapBERT, as a domain adapted pretraining method and BERT fine-tuned using adapters. Also, we included dapSciBERT, SciBERT+adapters, dapBERT+adapters, and dapSciBERT+adapters methods to highlight possible combinations of methods that can take place. All approaches that are based on standard task-specific fine-tuning have a classification head consisting of a dense layer, with ReLU

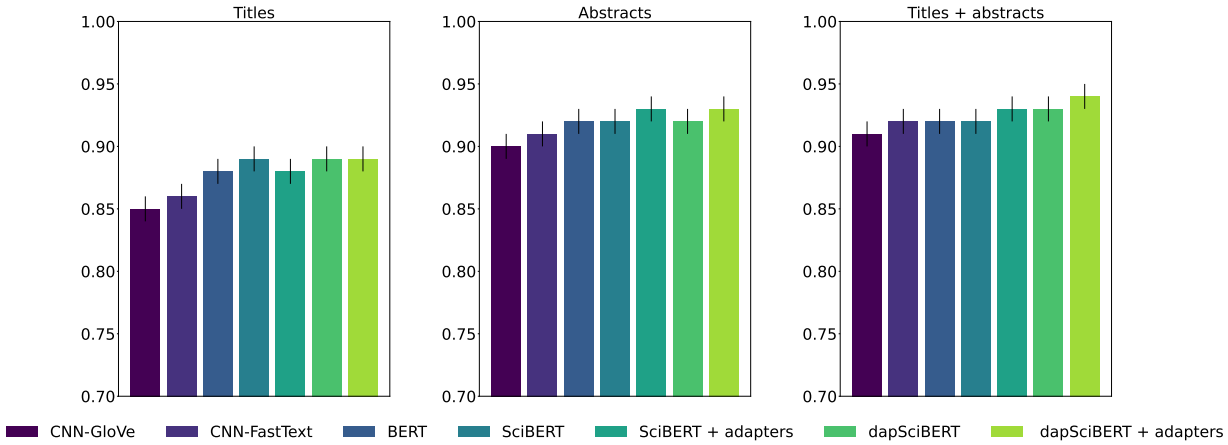



**FIGURE 3** Macro F1-score of the presented models in the the agrochemical related case based on different input text.

activation function and dropout, plus the output layer. The fine-tuning was performed for 5 epochs. When it comes to the methods that include adapters, the classification head consisted of only the output layer and the adapters-based fine-tuning had been performed for 30 epochs. For each method, we performed 10 independent runs and we presented the mean value of them.

DapBERT and dapSciBERT are two BERT-like models trained based on the domain adaptive pretraining method described above for the patent domain. Even though Gururangan et al[5] indicates that only a single pass on the domain dataset is required, we investigated whether additional training is beneficial for domain adaptation. We evaluated the versions of the adapted models that were trained for 3 epochs, as these versions yielded the best results. Additional training epochs did not improve the performance (see Appendix B). To train these models, we relied on the GT4SD library[22] and its LM trainer.

## 4.1 | Crop Protection Industry dataset

The discovery of fungicides, insecticides, and herbicides is the primary focus for crop protection research[23]. To keep up to date with emerging trends, the recent literature has to be followed. The manual identification of relevant articles is time-consuming and therefore, an automated identification and correct categorization mechanism would deliver a key element towards a more efficient process.

In patents, the IPC hierarchy provides a general classification of the patents based on their topics, yet this classification is not always aligned with important domain specific categories. Figure 2b highlights this aspect focusing on insecticides, herbicides, and fungicides as the three main crop protection categories. These three categories, have a high degree of overlap in terms of IPC codes, indicating that we cannot distinguish these labels relying solely on the IPC hierarchy.

There is a large amount of patent data available in the public domain. Patents starting from 2012 to 2020 were analyzed and classified manually into three categories - insecticide, fungicide, and herbicide – plus an extra no-class label with irrelevant patents, leading to a data set with 9,976 entries. This data set was used to build and evaluate various 4 class models based on the methodologies described above. The evaluation included a 10-fold cross-validation.

Figure 3 presents the results of the evaluation focusing on the macro F1-score. We examined three distinct input cases: title, abstract, and title+abstract, and included only SciBERT based adapted models in the results due to their superior performance. Tables with the full evaluation, including all the investigated models and all the metrics, can be found in Appendix A. The results show that any kind of domain adaptation can be beneficial for the task and all of them outperform the baselines. The best performance is observed by dapSciBERT+adapters method using as input both the title and the abstract of a patent. An approach that has the extra advantage of fewer storage requirements in case of multiple classification cases should be accommodated by the same system. Furthermore, the results dictate that the abstract is a way more informative part of a patent than the title for the classification task. To verify the significance of the difference between the baselines and our proposed alternatives we performed statistical tests to compare the mean value of the best baseline method which is the finetuned BERT model and our best variant in terms of both performance and methodological advantages which is the dapSciBERT+adapters. A Wilcoxon Signed-Ranks Test indicated that dapSciBERT+adapters method has greater mean value than BERT methods using as input title (statistic=50.0,

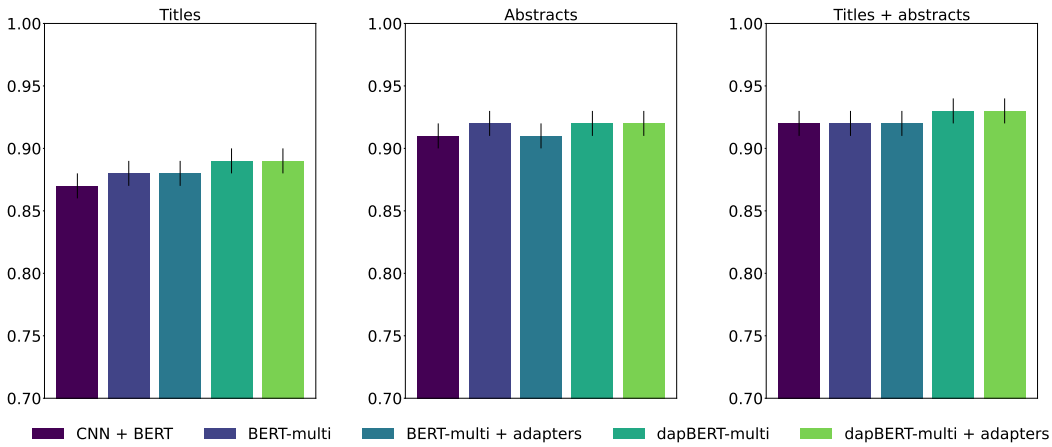



**FIGURE 4** Macro F1-score of the presented models in the multilingual versions of the the agrochemical case based on different input text.

pvalue<0.05), abstract (statistic=55.0, pvalue<0.05) and title+abstract (statistic=52.0, pvalue<0.05). In general, the differences in terms of mean F1-score between the methods are not large, yet even a 1% or 2% increase in such scenarios is highly important as this improvement is translated into tens of thousands of further correctly classified incoming patents in automated streams that should handle every year millions of patents.

## Multilingual patent classification

We also extended the models evaluations to determine the applicability of the previously described methodologies in a multilingual scenario. In fact, a reliable multilingual classifier could be a game changer for the field, as it could classify patents submitted to a large number of different speaking patent offices in a massively, quickly, and inexpensive way, without the need for additional tools such as translators. We relied on BERT's multilingual version and re-evaluated all the different adaptation techniques described previously (adapters, domain adapted models, and their combination), training also an adapted BERT multilingual model for the patent domain called dapBERT-multi. For the multilingual case, there are no available LMs trained on scientific or other relevant domains, thus we confined our experiment to the multilingual BERT variant. We used a multilingual corpus of 17,476,660 patent abstracts for the adapted pretraining. The patents collected covered a variety of languages, including English, Chinese, French, Korean, Japanese, and German. We trained the model for 3 epochs.

The methods were evaluated using the same case and dataset as in the English version. Nonetheless, a sizable portion of the patents was chosen in their alternate non-English language version. In total the dataset had 9,989 patents with 47% not written in English. Figure 4 presents the performance of our proposed methods. We chose the multilingual BERT as a baseline, as well as a CNN approach based on the multilingual BERT's embeddings. We used the same evaluation strategy as described previously for monolingual models. In general, the adapted methods outperform the baselines, with the performance resembling closely the same pattern as the English version. The combination of an adapted model and adapters (dapBERT-multi + adapters) presents the best results which is more than 2% better than the baselines. Wilcoxon Signed-Ranks Test also verifies our findings, specifically it proves that dapBERT-multi + adapters method has greater mean value than multiBERT method using as input title (statistic=50.0, pvalue<0.05), abstract (statistic=48.0, pvalue<0.05) and title+abstract (statistic=53.0, pvalue<0.05). We also observe that the multilingual models have slightly worse performance than the monolingual English models even if the amount of data that has been used is the same and the dominant language in the multilingual dataset is still English. This observation underlines the power of language generalization that these LMs hold and their potential in such cases and domains.

## Evaluation under real-life conditions

In a real case scenario, a model is trained on past data and is applied to predict incoming novel data. To this end, the classification model trained on data up to 2020 was applied to patents made available in 2021. This exercise focused on patents related to human necessities or chemistry, which are defined by their assigned IPC codes, leading to a corpus of 78712 patents. For

**TABLE 1** Evaluation of the performance of different monolingual or multilingual classifiers in a patent stream of patents that made available in 2021. The groundtruth is originated from manual annotations made by experts. The utilized metrics are the F1-score, the number of correct classified patents (correctly classified patents of interest), the total number of found patents (found patents of interest including cases with wrong label assignment) and the number of false positive patents (generally irrelevant patents that be assigned to one of the categories).

| Model | F1-score | Correct Found | Total Found (correct + missclassified) | False positives |
|---|---|---|---|---|
| BERT | 0.94 | 754 | 812 | 4786 |
| dapSciBERT + adapters | 0.98 | 742 | 786 | 1870 |
| BERT-multi | 0.94 | 753 | 811 | 4521 |
| dapBERT-mulit + adapters | 0.96 | 739 | 785 | 3352 |
| Total patents of interest in 2021 | | | 839 | |

the patents in the list that were not written in English, we relied on the respective English version that was retrieved through Google patents. The best performing model (dapSciBERT + adapters) was subsequently applied to label the patents into the three classes (fungicide, herbicide, insecticide) or to provide the no-class label. The results of the predictions were compared to the classification obtained from subject matter experts. An interesting observation is that using our model we could identify errors that occurred in five cases during the manual annotation. In addition, 108 patents that the model classified as insecticide, herbicide, or fungicide contained relevant keywords but were not classified accordingly by the subject matter experts. Of course, patents may contain relevant keywords without being relevant, but this could indicate that some relevant patents were missed during the manual assessment. Undoubtedly, the volume of false positive examples indicates that a patent classifier cannot yet be used as a standalone method to cherry-pick patents of interest, given an incoming patent stream. Nevertheless, such a model can facilitate the process and significantly reduce the volume of patents that require manual inspection. The model was also compared to a baseline model. As a baseline, we relied again on a fine-tuned BERT model for patent classification. Furthermore, to investigate multilingual extensions, we repeated the same experiment using a multilingual version of the same corpus in which the original language of each patent is used as input text, corresponding to 58% of the patents. Table 1 presents the overall results. The confusion matrices of classification results for each model can be found in Appendix A. For both only-English and multilingual cases, our model outperforms the baseline in terms of the F1 score. The fine-tuned BERT baselines retrieve a few more correct instances in both cases, yet our proposed models manage to provide significantly fewer false positive instances. Thus, the percentage of actual patents of interest predicted with our models is much higher than the baselines. IPC-code filtering or ensemble classifiers could improve even more the performance and reduce or eliminate the need for manual intervention. The comparison between English and multilingual classifiers' performance seconds the previous investigations' findings. It indicates that even if both variants have remarkable performance, using the English-only classifier leads to slightly better performance. This performance can be attributed to the fact that we relied on a standard multilingual BERT as there is no available multilingual BERT-like model trained solely on scientific domains.

## 4.2 | USPTO dataset

We further evaluated our methods based on an USPTO based dataset used in Roudsari et al[15]. It contained 235,858 patents submitted in 2014 as training set and 42,321 patents submitted in 2015 as testing set. As dataset's target was given both the title and the abstract of a patent to identify the associated IPC subclass labels of it. The target labels were 89 IPC subclasses. More information about the dataset generation process can be found in Roudsari et al[15].

Table 2 summarizes the comparison of the different models based on micro averaging precision, recall, and F1-score as well as coverage error. Coverage error is a metric that depicts how far we need to go down a ranked list of categories on average to account for all the true positive categories. The results for the embedding-based methods have been extracted from Roudsari et al[15].

Micro averaging of the above metrics, as well as evaluation at top-1 and top-5 predictions, have been also examined following the exact same evaluation process of Roudsari et al[15], and the results reveal a similar performance with the macro averaging results, are available in the Appendix B.

**TABLE 2** Coverage error and micro averaging precision, recall and F1-score of all the evaluated methods, considering the USPTO dataset benchmark proposed in Roudsari et al[15]. In bold the top-3 values for each metric.

| Model | Macro | | | Coverage Error |
|---|---|---|---|---|
| | Precision | Recall | F1 | |
| CBOW | 0.64 | 0.46 | 0.52 | 4.50 |
| GloVe | 0.68 | 0.42 | 0.51 | 4.36 |
| Skip-gram | 0.74 | 0.42 | 0.52 | 3.92 |
| FastText | 0.76 | 0.51 | 0.60 | 3.87 |
| GPT-2 | 0.76 | 0.49 | 0.59 | 3.90 |
| BERT | 0.76 | 0.52 | 0.60 | 3.52 |
| SciBERT | 0.76 | 0.53 | 0.62 | 3.46 |
| BERT+adapters | 0.77 | 0.52 | 0.61 | 3.40 |
| SciBERT+adapters | **0.78** | 0.53 | 0.62 | 3.29 |
| dapBERT | 0.77 | **0.54** | **0.63** | 3.31 |
| dapSciBERT | 0.77 | **0.55** | **0.63** | **3.28** |
| dapBERT+adapters | **0.79** | 0.53 | 0.62 | **3.19** |
| dapSciBERT+adapters | **0.78** | **0.54** | **0.63** | **3.15** |

The USPTO evaluation suggests that BERT-like approaches outperform the competition and that all of the presented domain adaptation based methods have the potential to improve the performance even further. The leveraging of SciBERT or adapters for vocabulary adaptation offers marginal benefits in the overall task performance in comparison to standard BERT finetuning. The improvement was becoming more significant when we utilized the domain-adapted models as dapBERT and mainly dapSciBERT had more than 2% improvement comparing to BERT in the majority of the utilized metrics. The combination of Patent based models and adapters can improve even more the performance of the models, especially in terms of coverage error. Overall, based on the experiments the best approach was the dapSciBERT+adapters which achieved a significant improvement in comparison to the rest CNN+embedding methods and the BERT baseline.

## 5 | CONCLUSION

Patent classification is a fundamental step in many patent analysis or patent generation pipelines. Improving the performance of classification methods is a critical task, and domain adaptation of transformers appears to be a promising direction. In this paper, we propose and investigate different methods for patent classification relying mainly on domain adaptation. The domain adaptive pretraining demonstrated the best results in the test cases and its performance can be furthermore improved if we select a pretrained base model with vocabulary closer to our domain, such as SciBERT. Additionally, the domain-adapted LM generated in the first phase can be fine-tuned and used for any downstream NLP task. When combined with already domain-adapted models, the use of adapters results in the same or even better performance. This finding, combined with their lightweight characteristics, such as requiring fewer training resources and less storage space, makes them an appealing option, particularly when it is required the development of multiple classification schemes for a single domain. We further utilized and examined the use of domain adaptation techniques for multilingual patent classification. The multilingual performance reflects the results of the English-written patents. The same level of performance in both multilingual and English cases highlights the strength of the proposed methods for patent classification and the great application potential that they can serve. To push further performance boundaries, future steps may include the exploration of additional domains[5], vocabulary[24] adaptation methods or the use of patents' metadata.

# APPENDIX

## A CROP PROTECTION INDUSTRY DATASET

Tables A1 presents the full evaluation of the models in the agrochemical related case. Furthermore, Tables A2 and A3 present the full evaluation of the models in the multilingual version of the respective dataset. Lastly, Tables A4, A5, A6 and A7 present the confusion matrices of the classification results obtained by our models or baselines in the evaluation using the patents published in 2021 and as groundtruth the annotations made by experts of the field. All tables have in bold the top value for each metric.

**TABLE A1** Performance of the evaluated models in the agrochemical dataset.

| Model | Titles | | | Abstracts | | | Titles+Abstracts | | |
|---|---|---|---|---|---|---|---|---|---|
| | Precision | Recall | F1 | Precision | Recall | F1 | Precision | Recall | F1 |
| CNN+GloVe | 0.86±0.01 | 0.84±0.01 | 0.85±0.01 | 0.91±0.01 | 0.89±0.01 | 0.90±0.01 | 0.92±0.01 | 0.90±0.01 | 0.91±0.01 |
| CNN+FastText | 0.87±0.01 | 0.86±0.01 | 0.86±0.01 | 0.92±0.01 | 0.90±0.01 | 0.91±0.01 | 0.92±0.01 | 0.91±0.01 | 0.92±0.01 |
| BERT | 0.89±0.02 | 0.88±0.01 | 0.88±0.01 | 0.92±0.01 | 0.91±0.01 | 0.92±0.01 | 0.92±0.01 | 0.92±0.01 | 0.92±0.01 |
| SciBERT | 0.89±0.02 | 0.88±0.01 | **0.89±0.01** | 0.92±0.01 | 0.92±0.01 | 0.92±0.01 | 0.92±0.01 | 0.92±0.01 | 0.92±0.01 |
| BERT+adapters | 0.89±0.02 | **0.89±0.01** | **0.89±0.01** | 0.92±0.01 | 0.92±0.01 | 0.92±0.01 | **0.93±0.01** | 0.93±0.01 | 0.93±0.01 |
| SciBERT+adapters | 0.89±0.02 | 0.88±0.01 | 0.88±0.01 | **0.93±0.01** | **0.93±0.01** | **0.93±0.01** | **0.93±0.01** | 0.93±0.01 | 0.93±0.01 |
| dapBERT | 0.89±0.01 | **0.89±0.01** | **0.89±0.01** | 0.91±0.01 | 0.92±0.01 | 0.91±0.01 | **0.93±0.01** | 0.93±0.01 | 0.92±0.01 |
| dapSciBERT | **0.90±0.01** | **0.89±0.01** | **0.89±0.01** | 0.92±0.01 | 0.92±0.01 | 0.92±0.01 | **0.93±0.01** | 0.93±0.01 | 0.93±0.01 |
| dapBERT+adapters | 0.89±0.01 | **0.89±0.01** | **0.89±0.01** | 0.92±0.01 | 0.92±0.01 | 0.92±0.01 | **0.93±0.01** | 0.93±0.01 | 0.93±0.01 |
| dapSciBERT+adapters | 0.89±0.01 | **0.89±0.01** | **0.89±0.01** | **0.93±0.01** | **0.93±0.01** | **0.93±0.01** | **0.93±0.01** | **0.94±0.01** | **0.94±0.01** |

**TABLE A2** Performance of the evaluated models in the multilingual agrochemical dataset.

| Model | Titles | | | Abstracts | | | Titles+Abstracts | | |
|---|---|---|---|---|---|---|---|---|---|
| | Precision | Recall | F1 | Precision | Recall | F1 | Precision | Recall | F1 |
| CNN+BERT | 0.88±0.01 | 0.87±0.01 | 0.87±0.01 | 0.91±0.01 | 0.91±0.01 | 0.91±0.01 | 0.92±0.01 | 0.91±0.01 | 0.92±0.01 |
| BERT-multi | **0.89±0.01** | 0.88±0.01 | 0.88±0.01 | **0.92±0.01** | 0.91±0.01 | **0.92±0.01** | 0.92±0.01 | 0.92±0.01 | 0.92±0.01 |
| BERT-multi + adapters | **0.89±0.01** | 0.88±0.01 | 0.88±0.01 | **0.92±0.01** | 0.91±0.01 | 0.91±0.01 | 0.92±0.01 | 0.92±0.01 | 0.92±0.01 |
| dapBERT-multi | **0.89±0.01** | 0.88±0.01 | **0.89±0.01** | **0.92±0.01** | 0.91±0.01 | **0.92±0.01** | **0.93±0.01** | 0.92±0.01 | **0.93±0.01** |
| dapBERT-multi + adapters | **0.89±0.01** | **0.89±0.01** | **0.89±0.01** | **0.92±0.01** | **0.92±0.01** | **0.92±0.01** | **0.93±0.01** | **0.93±0.01** | **0.93±0.01** |

**TABLE A3** Performance of the evaluated models in the multilingual datasets focusing on patents not written in English and using abstract and title as input.

| Model | Precision | Recall | F1 |
|---|---|---|---|
| CNN+BERT | **0.91±0.01** | 0.88±0.01 | 0.89±0.01 |
| BERT-multi | **0.91±0.01** | 0.89±0.01 | 0.90±0.01 |
| BERT-multi + adapters | 0.90±0.01 | **0.90±0.01** | 0.90±0.01 |
| dapBERT-multi | **0.91±0.01** | 0.89±0.01 | 0.90±0.01 |
| dapBERT-multi + adapters | **0.91±0.01** | **0.90±0.01** | **0.91±0.01** |

**TABLE A4** Confusion matrix of results obtained by the fined-tuned BERT in the 2021 patent stream.

| | | Actual | | | |
|---|---|---|---|---|---|
| | | Fungicide | Herbicide | Insecticide | No class |
| Predicted | Fungicide | 175 | 0 | 2 | 763 |
| | Herbicide | 6 | 166 | 13 | 918 |
| | Insecticide | 30 | 7 | 413 | 3105 |
| | No class | 7 | 3 | 12 | 73092 |

**TABLE A5** Confusion matrix of results obtained by our dapSciBERT+adapters approach in the 2021 patent stream.

| | | Actual | | | |
|---|---|---|---|---|---|
| | | Fungicide | Herbicide | Insecticide | No class |
| Predicted | Fungicide | 207 | 6 | 14 | 697 |
| | Herbicide | 0 | 160 | 1 | 206 |
| | Insecticide | 17 | 6 | 375 | 967 |
| | No class | 13 | 4 | 30 | 76009 |

**TABLE A6** Confusion matrix of results obtained by the fined-tuned multilingual BERT in the 2021 patent stream.

| | | Actual | | | |
|---|---|---|---|---|---|
| | | Fungicide | Herbicide | Insecticide | No class |
| Predicted | Fungicide | 186 | 1 | 14 | 672 |
| | Herbicide | 4 | 161 | 5 | 794 |
| | Insecticide | 24 | 10 | 406 | 3055 |
| | No class | 10 | 5 | 13 | 73361 |

**TABLE A7** Confusion matrix of results obtained by our multilingual dapBERT-multi + adapters approach in the 2021 patent stream.

| | | Actual | | | |
|---|---|---|---|---|---|
| | | Fungicide | Herbicide | Insecticide | No class |
| Predicted | Fungicide | 195 | 4 | 12 | 554 |
| | Herbicide | 0 | 157 | 2 | 293 |
| | Insecticide | 21 | 7 | 387 | 2505 |
| | No class | 18 | 8 | 28 | 74530 |

## B USPTO DATASET

Table B8 presents the results of the different models at top 1 and top 5 predictions. Specifically, we first predict 1 and 5 labels for each patent, and then we calculate the precision, recall, and F1-score. The results for the embedding-based methods have been extracted from Roudsari et al[15]. In addition, table B9 depicts the comparison of the different models based on micro averaging precision, recall, and F1-score as well as coverage error. Coverage error is a metric that depicts how far we need to go down a ranked list of categories on average to account for all the true positive categories. Various dapBERT and dapSciBERT checkpoints have been added to the table. Specifically, we compare the versions taken after 1, 3, and 10 training epochs. As it can be extracted, the checkpoints taken after 3 and 10 raining epochs performs equally, which indicates that 3 training epochs of adapted pretraining is enough and no further performance can be seen for the patent classification task with further training. Both tables have in bold the top value for each metric.

**TABLE B8** Performance of all the evaluated methods at top-1 and top-5 predictions considering the USPTO dataset benchmark proposed in Roudsari et al[15].

| Model | @1 (%) | | | @5 (%) | | |
|---|---|---|---|---|---|---|
| | Precision | Recall | F1 | Precision | Recall | F1 |
| CBOW | 75.80 | 56.15 | 61.90 | 27.62 | 88.26 | 40.33 |
| GloVe | 76.51 | 56.71 | 62.51 | 27.90 | 89.14 | 40.73 |
| Skip-gram | 78.80 | 58.46 | 64.42 | 28.49 | 90.68 | 41.54 |
| FastText | 78.87 | 58.49 | 64.46 | 28.48 | 90.70 | 41.53 |
| GPT-2 | 80.52 | **59.97** | **65.99** | 28.51 | 90.57 | 41.55 |
| BERT | 82.25 | 50.68 | 62.71 | 29.10 | 89.65 | 43.94 |
| SciBERT | 83.20 | 51.26 | 63.44 | 29.28 | 90.22 | 44.21 |
| BERT±adapters | 82.85 | 51.05 | 63.18 | 29.22 | 90.02 | 44.12 |
| SciBERT±adapters | 83.63 | 51.52 | 63.77 | 29.46 | 90.76 | 44.48 |
| dapBERT | 83.95 | 51.73 | 64.01 | 29.45 | 90.74 | 44.47 |
| dapSciBERT | 84.28 | 51.93 | 64.26 | 29.56 | 91.08 | 44.64 |
| dapBERT±adapters | 84.35 | 51.97 | 64.32 | 29.60 | 91.20 | 44.70 |
| dapSciBERT±adapters | **84.53** | 52.09 | 64.46 | **29.68** | **91.45** | **44.82** |

**TABLE B9** Performance of different domain adapted checkpoints in the classification task.

| Model | Micro | | | Coverage Error |
|---|---|---|---|---|
| | Precision | Recall | F1 | |
| CBOW | 0.71 | 0.55 | 0.62 | 4.50 |
| GloVe | 0.75 | 0.51 | 0.61 | 4.36 |
| Skip-gram | 0.80 | 0.51 | 0.62 | 3.92 |
| FastText | 0.80 | 0.51 | 0.62 | 3.87 |
| GPT-2 | 0.80 | 0.56 | 0.66 | 3.90 |
| BERT | 0.80 | 0.59 | 0.68 | 3.52 |
| SciBERT | 0.80 | 0.61 | 0.69 | 3.46 |
| $\text{dapBERT}_1$ | **0.81** | 0.59 | 0.68 | 3.51 |
| $\text{dapSciBERT}_1$ | 0.80 | **0.62** | 0.70 | 3.33 |
| $\text{dapBERT}_3$ | 0.80 | 0.61 | 0.70 | 3.31 |
| $\text{dapSciBERT}_3$ | **0.81** | **0.62** | **0.71** | **3.28** |
| $\text{dapBERT}_{10}$ | **0.81** | 0.61 | 0.70 | 3.31 |
| $\text{dapSciBERT}_{10}$ | **0.81** | 0.61 | **0.71** | 3.31 |

## References


1. WIPO - website . https://www.wipo.int/portal/en/index.html; 2021. Accessed: 2021-09-10.
2. USPTO - website . https://www.uspto.gov; 2021. Accessed: 2021-09-10.
3. website UPS. https://www.uspto.gov/web/offices/ac/ido/oeip/taf/us_stat.htm; 2021. Accessed: 2021-09-10.
4. WIPO . Guide to the International Patent Classification.; 2020.
5. Gururangan S, Marasović A, Swayamdipta S, et al. Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. In: Association for Computational Linguistics; 2020; Online: 8342–8360
6. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Association for Computational Linguistics; 2019; Minneapolis, Minnesota: 4171–4186

7. Beltagy I, Lo K, Cohan A. SciBERT: A Pretrained Language Model for Scientific Text. In: Association for Computational Linguistics; 2019; Hong Kong, China: 3615–3620

8. Houlsby N, Giurgiu A, Jastrzebski S, et al. Parameter-Efficient Transfer Learning for NLP.. In: Chaudhuri K, Salakhutdinov R., eds. *ICML*. 97 of *Proceedings of Machine Learning Research*. PMLR; 2019: 2790-2799.

9. Krestel R, Chikkamath R, Hewel C, Risch J. A survey on deep learning for patent analysis. *World Patent Information* 2021; 65: 102035. doi: https://doi.org/10.1016/j.wpi.2021.102035

10. Fall CJ, Törcsvári A, Benzineb K, Karetka G. Automated categorization in the international patent classification. In: . 37. ACM New York, NY, USA. ; 2003: 10–25.

11. D'hondt E, Verberne S, Koster C, Boves L. Text representations for patent classification. *Computational Linguistics* 2013; 39(3): 755–775.

12. Trappey AJ, Hsu FC, Trappey CV, Lin CI. Development of a patent document classification and search platform using a back-propagation network. *Expert Systems with Applications* 2006; 31(4): 755–765.

13. Li S, Hu J, Cui Y, Hu J. DeepPatent: patent classification with convolutional neural networks and word embedding. *Scientometrics* 2018; 117(2): 721–744.

14. Abdelgawad L, Kluegl P, Genc E, Falkner S, Hutter F. Optimizing neural networks for patent classification. In: Springer. ; 2019: 688–703.

15. Roudsari AH, Afshar J, Lee S, Lee W. Comparison and Analysis of Embedding Methods for Patent Documents. In: IEEE. ; 2021: 152–155.

16. Hepburn J. Universal Language model fine-tuning for patent classification. In: ; 2018: 93–96.

17. Lee JS, Hsiang J. Patent classification by fine-tuning BERT language model. *World Patent Information* 2020; 61: 101965.

18. D'hondt E, Verberne S. CLEF-IP 2010: Prior Art Retrieval Using the Different Sections in Patent Documents.. In: Braschler M, Harman D, Pianta E., eds. *CLEF (Notebook Papers/LABs/Workshops)*. 1176 of *CEUR Workshop Proceedings*. CEUR-WS.org; 2010.

19. Hendrycks D, Liu X, Wallace E, Dziedzic A, Krishnan R, Song D. Pretrained Transformers Improve Out-of-Distribution Robustness. In: Association for Computational Linguistics; 2020; Online: 2744–2751

20. Zhu Y, Kiros R, Zemel RS, et al. Aligning Books and Movies: Towards Story-Like Visual Explanations by Watching Movies and Reading Books.. In: IEEE Computer Society; 2015: 19-27.

21. Ammar W, Groeneveld D, Bhagavatula C, et al. Construction of the Literature Graph in Semantic Scholar. In: Association for Computational Linguistics; 2018; New Orleans - Louisiana: 84–91

22. Manica M, Cadow J, Christofidellis D, et al. GT4SD: Generative Toolkit for Scientific Discovery. *arXiv preprint arXiv:2207.03928* 2022.

23. Umetsu N, Shirai Y. Development of novel pesticides in the 21st century. *Journal of Pesticide Science* 2020; 45(2): 54–74.

24. Tai W, Kung HT, Dong X, Comiter M, Kuo CF. exBERT: Extending Pre-trained Models with Domain-specific Vocabulary Under Constrained Training Resources. In: Association for Computational Linguistics; 2020; Online: 1433–1439