Establishing a Baseline for Arabic Patents Classification: A Comparison of Twelve Approaches

Taif Al-Omar[1], Hend Al-Khalifa[2] and Rawan Al-Matham[3]
iWAN Research Group, King Saud University
[1]taifalomar@gmail.com, [2]hendk@ksu.edu.sa, [3]r.almatham@gmail.com

Abstract

Nowadays, the number of patent applications is constantly growing and there is an economical interest on developing accurate and fast models to automate their classification task. In this paper, we introduce the first public Arabic patent dataset called ArPatent and experiment with twelve classification approaches to develop a baseline for Arabic patents classification. To achieve the goal of finding the best baseline for classifying Arabic patents, different machine learning, pre-trained language models as well as ensemble approaches were conducted. From the obtained results, we can observe that the best performing model for classifying Arabic patents was ARBERT with F1 of 66.53%, while the ensemble approach of the best three performing language models, namely: ARBERT, CAMeL-MSA, and QARiB, achieved the second best F1 score, i.e., 64.52%.

1 Introduction

Over the past few years, there has been an increased focus on improving patent classification systems. This is due to the growing recognition of the importance of classification in improving the efficiency of patent examination and in providing better access to information for users of patent databases.

Currently, patent examiners manually classify patents, which is a time-consuming process. If a method could be developed for automatically classifying patents, it would greatly reduce the amount of time needed to examine a patent application. This is an important area of research because it has the potential to significantly improve the efficiency of patent examination.

Current research efforts in patent classification are focusing on improving the efficiency and accuracy of the classification process. One area of research is exploring the use of machine learning algorithms to automatically classify patents (Aristodemou & Tietze, 2018). Another area of research is looking at ways to improve the use of prior art information in the classification process (Harris et al., 2010).

While classifying patent text is applied widely for some languages, such as, English. The Arabic version of the problem requires the availability of Arabic patent annotated corpus with considerable size as well as experimenting with different classification models. Therefore, the contributions of this paper include:

1- Constructing the first Arabic Patent dataset (called ArPatent) labeled with the International Patent Classification (IPC).
2- Evaluating twelve classification approaches in order to achieve a baseline for Arabic patents classification.

The rest of the paper is organized as follows: sections 2 and 3 provide background and related work on patents and their classification; section 4 presents the data collection and preprocessing process. Sections 5 and 6 discuss the used methods and the obtained results. Finally, section 7 concludes the paper with limitation and research outlook.

2 Background

A patent is a form of intellectual property that gives its owner the legal right to exclude others from making, using, or selling an invention for a limited period of time.

Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP), pages 287 - 294

A patent document typically includes a title, an abstract, a classification, a background section, a brief summary of the invention, a detailed description of the invention, one or more claims and drawings.

There are two schemes used for patent classification: (1) International Patent Classification (IPC) (International Patent Classification (IPC), n.d.) and (2) Cooperative Patent Classification (CPC) (Office, n.d.).

IPC scheme is a hierarchical patent classification system used in over 100 countries to classify the content of patents in a uniform manner; hence it is usually used for Arabic patents classification. It was established by the World Intellectual Property Organization (WIPO) in 1971.

Each patent publication is given one classification Section (see Table 1) identifying the topic to which the invention relates [1]. Further classification sections and indexing codes may be given to provide further information about the contents.

Section (Class)	Topic
A	Human Necessities
B	Performing Operations, Transporting
C	Chemistry, Metallurgy
D	Textiles, Paper
E	Fixed Constructions
F	Mechanical Engineering, Lighting, Heating, Weapons
G	Physics
H	Electricity

Table 1: IPC eight classification sections and topics

3 Related Work

There are many studies in the literature that tackled automated patent classification. The earliest studies employed classical Natural Language Processing (NLP) and Machine Learning (ML) approaches with feature engineering. For instance, (Fall et al., 2003) did many experiments in two levels, classes and subclasses. They compared the precision of using many classifiers, Naïve Bayes (NB), Support Vector Machine (SVM), Spars Network of Windows (SNoW) and K-Nearest Neighbor (KNN) classifiers on their self-collected WIPO [2] dataset named WIPO-alpha, with performing stop words removal, stemming, and term selection using information gain as a preprocessing step. The best result in classes-level experiments was obtained with NB classifier (79%). While in the subclasses-level the best result was with KNN (62%).

Similarly, (Tikk et al., 2008) used stemming, dimensionality reduction, stop word removal and removal of rare terms with a neural network called HITEC which was evaluated on WIPO-alpha corpus and Espace A/B [3] corpora. there results outperform other state-of-the art classifiers significantly (by 6.5~14.5%). Additionally, (Lim & Kwon, 2016) created a list of stop words specifically for the patent domain and used a TF-IDF weighing system to choose their feature set. The patent document classification was examined using a multi-label model and 564,793 registered Korean patents at the IPC subclass level. They achieved a precision rate of 87.2% when using titles, abstracts, claims, technical fields, and backgrounds.

Around 2017, the focus of automated patent classification research changed to Deep Learning (DL) approaches. (Grawe et al., 2017) trained Espace A/B patent dataset with Word2Vec and fed it to LSTM classifier. At the level of subclasses, they achieved an accuracy rate of 63%. Likewise, (Xiao et al., 2018) applied similar approach (Word2Vec and LSTM) with their self-collected domain-specific patent datasets for security. Their approach achieved 93.48% of accuracy. On the other hand, (Risch & Krestel, 2019) employed bi-directional GRUs (another type of RNN) to improve classification performance compared to Word Embeddings trained with Word2Vec on Wiki pages and the FastText embedding that was trained on different datasets (WIPO-alpha ,USPTO[4]-2M and USPTO-5M). Their approach increased the average precision for patent classification by 17 percent compared to state-of-the-art approaches.

Moreover, (Sofean, 2021) created a self-trained Word Embedding that was trained on a million patents collected from multiple patents datasets (European, German, Japanese and Chinese patents), and then classified them using LSTM

[1] https://ipcpub.wipo.int/
[2] https://www.wipo.int/
[3] https://www.epo.org/
[4] https://bulkdata.uspto.gov/

network. He obtained an accuracy of 67%. Another paper by (S. Li et al., 2018) developed DeepPatent algorithm for patent classification by combining Word Embeddings with CNN. It was tested on USPTO-2M dataset. Their approach achieved precision of 73.88%. Similarly, (Zhu et al., 2020) experiments were applied to classify Chinese short text patent using Word Embedding with CNN. Their results outperformed traditional RNN.

However, using CNNs alone has gained the best results in patent classification task. In (Abdelgawad et al., 2020), they achieved an accuracy of 52.02% at the subclass level with the WIPO-alpha dataset.

(Lee & Hsiang, 2020) fine-tuned a pre-trained BERT model on a their self-collected patent dataset which contains three million patents. They focused on patent claims without other parts and the best results they achieved was 66.83% for F1.

Ensemble techniques were also among the used approaches in patent classifications. (Eleni Kamateri et al., 2022) experimented with CLEF-IP 2011 test collection to compare the accuracy of classifying English patents using different individuals' classifiers (CNN, Bi-LSTM, Bi-GRU, LSTM, and GRU) and using ensemble approach with three classifiers. Their highest accuracy (64.85%) was gained by using ensemble approach that combined three of their best performing classifiers.

From the previous studies we noticed that there is a huge interest in the community for patent classification in languages such as English and Korean. Furthermore, the best results were obtained with fine-tuning BERT language models and ensemble approach. Therefore, in this paper we created the first Arabic patent dataset and applied the best classification approaches mentioned previously in the literature.

4 Data Acquisition and Preprocessing

4.1 Data Acquisition

Delivering an Arabic patent dataset is one of the contributions in this paper. The dataset was acquired by scraping the granted Arabic patents at the Saudi Authority of Intellectual Property (SAIP)[5] website using Selenium[6]. All the available patents at the website were retrieved, mainly 9772 patents. All patents were typically composed of the following sections: title, abstract, applicant, the International Patent Classification (IPC) and other details (see Figure 1). The size of the acquired data was approximately 413 MB with 1.58M tokens — words. We named our dataset "ArPatent" and it is publicly available through our Github repository[7].


[19] المملكة العربية السعودية SA
مدينة الملك عبدالعزيز للعلوم والتقنية
[11] رقم البراءة: ١٢٧٨
[45] تاريخ المنح: ١٤٢٧/٠٨/٢٦هـ
الموافق: ٢٠٠٦/٠٩/١٩م
[12] براءة اختراع
[51] التصنيف الدولي: Int. Cl.7:C07D 301/12
[56] المراجع:
براءة امريكية ٥٣٧٤٧٤٧ ١٩٩٤/١٢/٢٠م
براءة امريكية ٥٤١٢١٢٢ ١٩٩٥/٠٥/٠٢م
براءة اوروبية ٧١٢٨٥٢ ١٩٩٦/٠٥/٢٢م
براءة امريكية ٥٥٩١٨٧٥ ١٩٩٧/٠١/٠٧م
طلب دولي ٩٨٤٣٧٣٦ ١٩٩٨/١٠/٠٨م
اسم الفاحص: فهيد بن خلف السبيعي
[72] اسم المخترع: جان- بيار كاتينا، ميشال ستريبيل
[73] مالك البراءة: سولفاي
عنوانه: (شارع) رو دو برنس البير، ٣٣، بروكسيل ب-١٠٥٠، بلجيكا
[74] الوكيل: احمد عبدالعزيز الراجحي
[21] رقم الطلب: ٠٠٢٠٠٨٩٣
[22] تاريخ الإيداع: ١٤٢٠/١٠/١٩ هـ
الموافق: ٢٠٠٠/٠١/٢٦ م
[54] اسم الاختراع: عملية لصنع اوكسيران oxirane
[57] الملخص: عملية مستمرة لصنع اوكسيران oxirane و التي بموجبها يتفاعل اوليفين olefin عند درجة حرارة تفوق ٣٥ ° مئوية و لمدة تزيد عن ٤٨ ساعة مع مركب بيروكسيدي peroxide compound بوجود مساعدة حفازة catalyst اساسها الزيوليت zeolite و بموجود ملح فلزي ، و التي فيها لا تخضع المادة الحفازة لعملية تجديد و التي تكون فيها سرعة اخماد deactivation المادة الحفازة ، المعبر عنها كنسبة مئوية لتحول المركب البيروكسيدي المفقود لكل غرام من الاوكسيران المنتج لكل غرام زيوليت و المحددة بعد اتمام ظروف التفاعل اي بعد استهلاك ٢٠٥ جزيء غرامي من الوظيفة البيروكسيدية لكل غرام زيوليت، دائما اقل من او مساوية لـ ٢٠،١٥ ٪ .

Figure 1: A screenshot of an Arabic patent document from SAIP[5]

4.2 Preprocessing

We chose to classify the patents into one of the eight IPC sections by training the models using titles and abstracts onlybelow; as they are the common sections among almost all patents. Nevertheless, six patents had no abstract hence were removed from the dataset. Moreover, one patent had no accurate IPC section and was excluded, so the remaining patents reached 9765. To explore and understand the dataset, we calculated the unigram and bigram frequencies after the data has been preprocessed as shown in Table 2 and Table 3.

The preprocessing steps consisted of removing Arabic and English punctuations or digits, removing elongation and normalizing Arabic letters by replacing different forms of alif "آ،إ،أ،ا" into the simple form "ا", replacing "ة" into "ه", and finally replacing any "ى" into "ي". It is noteworthy to mention that words with English letters were

[5] https://www.saip.gov.sa/en
[6] https://www.selenium.dev/
[7] https://github.com/iwan-rg/Arabic-Patents

	Word	Count		Word	Count
1	الاختراع	9726	**7**	ماده	2479
2	الحالي	6557	**8**	يكون	2384
3	الاقل	5073	**9**	مجموعه	2131
4	يمكن	2893	**10**	جزء	2062
5	تشتمل	2558	**11**	باستخدام	1910
6	بواسطه	2517	**12**	تتضمن	1904

Table 2: The top 15 bigram words in the Arabic patent dataset along with their frequencies.

	Word	Count		Word	Count
1	الاختراع, الحالي	6095	**7**	الاختراع, الراهن	458
2	يتعلق, الاختراع	1175	**8**	اكسيد, الكربون	399
3	درجه, حراره	815	**9**	الحالي, بجهاز	362
4	واحده, الاقل	788	**10**	الحالي, بنظام	342
5	الحالي, بطريقه	621	**11**	الحالي, بتوفير	332
6	الحالي, بعمليه	459	**12**	الاختراع, بطريقه	332

Table 3: The top 15 bigram words in the Arabic patent dataset along with their frequencies.

preserved as they might represent scientific or technical terms, therefore keeping them might enhance the classification performance and prevent information loss.

Furthermore, the min/max/average of number of words in titles and abstracts were computed and resulted in 4/237/53 for titles, and 10/35371/681 for abstracts. We can say that our dataset is considered an imbalanced since the number of patents in each IPC section was far from equal, see Table 4.

Section (Class)	No. of Patents
A	2169
B	2038
C	2783
D	61
E	733
F	741
G	764
H	476

Table 4: Number of patents in each IPC section

5 Methodology

To achieve the goal of finding the best baseline for classifying Arabic patents, different machine learning approaches were considered. First, from the traditional approaches, SVM were implemented with different word Embeddings techniques; namely: TF-IDF, and Skip-Gram Word2Vec. For Word2Vec, we used the pre-trained Embeddings AraVec (Soliman et al., 2017), and we also trained our own Word2Vec version, named ArPatent-Word2vec[8], on the entire preprocessed Arabic patent text.

Moreover, since BERT-based models have shown state-of-the-art performance at language understanding, we fine-tuned multiple Arabic BERT-based models on the unprocessed titles and abstracts of ArPatent for the task of patent classification. The used fine-tuning hyperparameters were: learning rate of 2e-5 using Adam's optimizer, A dropout layer of 0.1 at the feed-forward classifier, a batch size of 32 and 3 epochs. Moreover, the max length of tokens was set to 350 tokens for the tokenizer, truncating any input beyond that length. Furthermore, stratified split was used for splitting the data into 10/10/80 for testing, validation, and training respectively. The evaluation metrics used are accuracy, macro-precision, macro-recall and macro-F1 score, having the last as the primary metric since the dataset is highly imbalanced.

The following subsections give a summary of the used pre-trained models.

5.1 CAMeL-MSA

CAMeL-MSA (Inoue et al., 2021) is a pre-trained BERT model on Modern Standard Arabic (MSA) corpus that is comprised of the following public Arabic corpora: Abu El-Khair Corpus, dump of the Arabic Wikipedia on February 01, 2019[9], The unshuffled version of the Arabic OSCAR corpus, the Arabic Gigaword Fifth Edition and lastly the OSIAN corpus. The resulting dataset consisted of 107GB of text, yielding 12.6B tokens. It is worth

[8] https://github.com/iwan-rg/Arabic-Patents

[9] https://archive.org/details/arwiki-20190201

mentioning that the authors removed lines that had no Arabic characters from the text.

5.2 CAMeL-MIX

Same authors of CAMeL-MSA have contributed to releasing a model that was pre-trained on different Arabic variants, i.e., the same MSA corpus mentioned earlier, a range of dialectal Arabic corpora and a classical Arabic corpus from OpenITI. The model was pre-trained on text size of 167GB with 17.3B tokens, which is the largest number of tokens among all other BERT variants mentioned in this paper.

5.3 ARBERT

ARBERT (Abdul-Mageed et al., 2021) is also a BERT-based model pre-trained on 61GB of MSA text with 6.5B tokens, gathered from 1,800 Arabic books, the fifth edition of GigaWord, Abu El-Khair Corpus, OSCAR, OSIAN, and the December 2019 dump of Arabic Wikipedia.

5.4 MARBERTv2

MARBERT (Abdul-Mageed et al., 2021) was pre-trained to be best suited for dialectical Arabic using 1B tweets. The dataset makes up 128GB of text —15.6B tokens. For MARBERTv2, they further pre-trained MARBERT on the same dataset of ARBERT in addition to AraNews dataset. New MARBERTv2 dataset makes up 29B of tokens.

5.5 AraBERTv2

AraBERTv2 (Antoun et al., 2020), is a pre-trained BERT model for MSA NLP tasks. It was pre-trained on 77GB or 200M sentences of Arabic content resulting in 8.6B tokens. The corpus consisted of manually scraped Arabic news websites and four publicly available corpora: Arabic Wikipedia dump from 2020/09/01 [10], OSCAR unshuffled and filtered, The 1.5B words Arabic Corpus, and lastly the OSIAN Corpus. The authors noted that words with Latin characters were preserved during the training.

5.6 QARiB

QARiB (Ahmed Abdelali et al., 2021) was pre-trained on a collection of 420M tweets and 180M sentences of text. For the sentences, it was a combination of Arabic GigaWord Fourth Edition, Abu El-Khair Corpus and OpenSubtitles corpus. It resulted in 14B tokens

5.7 Max Voting Ensemble

An ensemble is a collection of models designed to exceed the performance of every single base-model by combining their predictions. Max voting —or majority vote— ensemble is one of the simplest methods of combining predictions. In max voting, each base-model makes a prediction and votes for each instance. In addition to fine-tuning the models, we considered designing an ensemble of the highest performing models. Since weighted sum ensembles presume that some models in the ensemble are more efficient compared to others; we considered applying both weighted and unweighted summing for predictions and reporting the highest.

Models	Acc.	P_{Macro}	R_{Macro}	$F1_{Macro}$
SVM-SG-ARAVEC	60.55	50.33	51.06	50.46
SVM-ArPatent-Word2vec	63.42	50.64	54.37	51.80
SVM-TF-IDF	63.83	54.11	**67.28**	56.62
CAMeL-MSA	68.03	59.10	59.73	59.27
CAMeL-MIX	66.39	57.07	56.72	56.81
ARBERT	70.18	72.60	64.91	**66.53**
MARBERTv2	62.40	52.37	51.36	51.68
AraBERTv2	63.11	53.01	51.40	51.74
QARiB	67.83	58.26	57.84	57.95
ENSEMBLE-1	**70.49**	**73.36**	63.54	64.52
ENSEMBLE-2	62.91	54.16	51.95	52.46
ENSEMBLE-3	68.24	59.60	57.66	58.10

Table 2: Experimental Results with the following metrics: accuracy (Acc.), macro precision (P), macro recall (R) and macro F1 score (F1).

[10] https://archive.org/details/arwiki-20200901

6 Results and Discussions

Table 5 shows our experimental results. We can see that the BERT-based models that were pre-trained on MSA text only, namely ARBERT and CAMeL-MSA, had a superior performance with an F1 measure of 66.53 and 59.27 respectively. Moreover, among the SVM classifiers, the SVM with TF-IDF word Embeddings achieved a significantly higher performance (F1= 56.62). Although ArPatent-Word2vec embeddings was trained on the ArPatent text, yet it did not improve the classification task, this might be attributed to the models being trained on the title and abstract of a patent rather than the complete patent dataset.

On the other hand, the accuracy of ENSEMBLE-1 results outperformed all other models with 70.49%. ENSEMBLE-1 consists of the three best performing models namely: ARBERT, CAMeL-MSA and QARiB, with weighted sum of 1, 0.5 and 0.5, giving ARBERT the priority in voting. We also combined the SVM models in ENSEMBLE-2 with the same weighted sum, giving the priority to SVM-TF-IDF.

The last ensemble, ENSEMBLE-3, was a combination of all SVMs and ARBERT, giving ARBERT and TF-IDF the twice weight of other SVMs.

From the results, we can observe that the best performing model in terms of F1 score was ARBERT. Although we tried to ensemble different models; but due to little diversity in base-models' predictions; the performance could not get improved, for example, refer to the confusion matrices in Figure 2, Figure 3 and Figure 4 for the models participating in ENSEMBLE-1 and notice the similarity in predictions. Moreover, due to the little number of samples for some classes, the overall performance was low, especially for class D where most of its instances were not classified correctly even with ARBERT.

7 Conclusion

Classification of patents is a crucial part of the patent system, as it allows for the efficient and effective management of patent information. In this paper, we constructed the first Arabic patent dataset called ArPatent and experimented with twelve different classification approaches to develop a baseline for Arabic patent classification.

Our results show that the ARBERT model had the best performance in classifying patents. We can

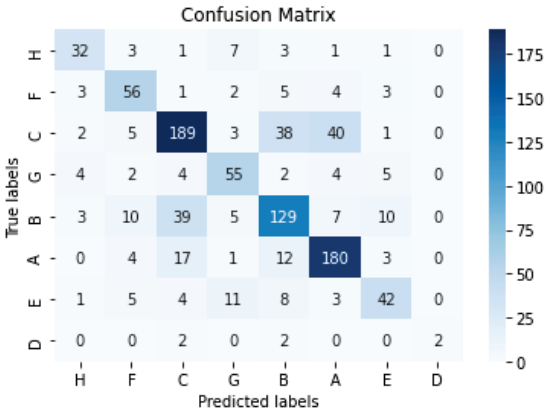

Figure 2: Confusion matrix of the best performing model ARBERT.

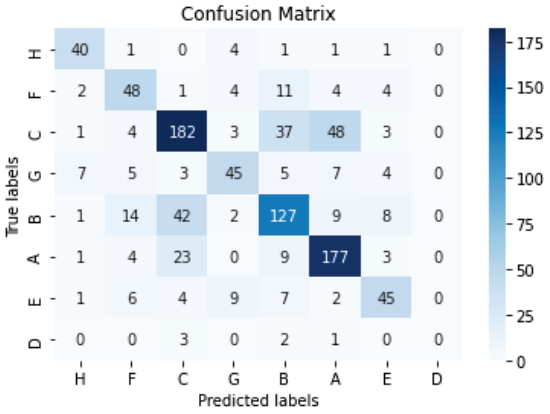

Figure 3: Confusion matrix of CAMeL-MSA.

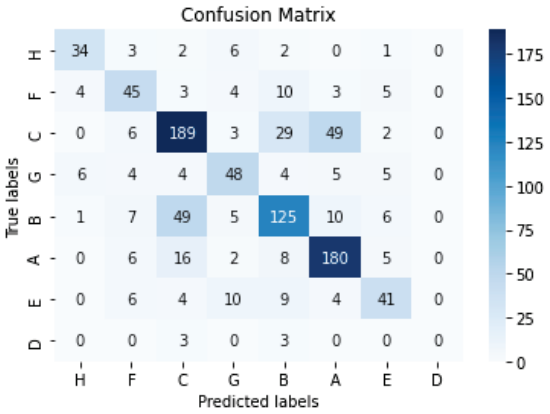

Figure 4: Confusion matrix of QARiB.

say that our results are comparable to those found in the literature, even with our small sized dataset.

One limitation of this work resides in the imbalanced dataset, this affected the performance of patent classification. Also, using only the patent's title and abstract for the purpose of classification did not yield good results. Therefore, as a future plan we intend to repeat the experiments while considering the whole text for classification

also we need to increase the size of the dataset to obtain more successful results.

Finally, we believe that our paper has produced some preliminary knowledge and useful results that will help support the task of Arabic patent classification.

References

Abdelgawad, L., Kluegl, P., Genc, E., Falkner, S., & Hutter, F. (2020). c In U. Brefeld, E. Fromont, A. Hotho, A. Knobbe, M. Maathuis, & C. Robardet (Eds.), Machine Learning and Knowledge Discovery in Databases (pp. 688–703). Springer International Publishing. https://doi.org/10.1007/978-3-030-46133-1_41

Abdul-Mageed, M., Elmadany, A., & Nagoudi, E. M. B. (2021). ARBERT & MARBERT: Deep Bidirectional Transformers for Arabic. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 7088–7105. https://doi.org/10.18653/v1/2021.acl-long.551

Ahmed Abdelali, Sabit Hassan, Hamdy Mubarak, Kareem Darwish, & Younes Samih. (2021). Pre-Training BERT on Arabic Tweets: Practical Considerations. https://arxiv.org/abs/2102.10684

Antoun, W., Baly, F., & Hajj, H. (2020). AraBERT: Transformer-based Model for Arabic Language Understanding. Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection, 9–15. https://aclanthology.org/2020.osact-1.2

Aristodemou, L., & Tietze, F. (2018). The state-of-the-art on Intellectual Property Analytics (IPA): A literature review on artificial intelligence, machine learning and deep learning methods for analysing intellectual property (IP) data. World Patent Information, 55, 37–51. https://doi.org/10.1016/j.wpi.2018.07.002

Eleni Kamateri, Vasileios Stamatis, Konstantinos Diamantaras, & Michail Salampasis. (2022). Automated Single-Label Patent Classification using Ensemble Classifiers. https://arxiv.org/abs/2203.03552

Fall, C. J., Törcsvári, A., Benzineb, K., & Karetka, G. (2003). Automated categorization in the international patent classification. ACM SIGIR Forum, 37(1), 10–25. https://doi.org/10.1145/945546.945547

Grawe, M. F., Martins, C. A., & Bonfante, A. G. (2017). Automated Patent Classification Using Word Embedding. 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA), 408–411. https://doi.org/10.1109/ICMLA.2017.0-127

Gomez, J. C., & Moens, M.-F. (2014). A Survey of Automated Hierarchical Classification of Patents. Professional Search in the Modern World, 215–249. https://doi.org/10.1007/978-3-319-12511-4_11

Harris, C. G., Arens, R., & Srinivasan, P. (2010). Comparison of IPC and USPC classification systems in patent prior art searches. Proceedings of the 3rd International Workshop on Patent Information Retrieval, 27–32. https://doi.org/10.1145/1871888.1871894

Inoue, G., Alhafni, B., Baimukan, N., Bouamor, H., & Habash, N. (2021). The Interplay of Variant, Size, and Task Type in Arabic Pre-trained Language Models. Proceedings of the Sixth Arabic Natural Language Processing Workshop, 92–104. https://aclanthology.org/2021.wanlp-1.10

International Patent Classification (IPC). (n.d.). Retrieved September 9, 2022, from https://www.wipo.int/classifications/ipc/en/index.html

Lee, J.-S., & Hsiang, J. (2020). Patent classification by fine-tuning BERT language model. World Patent Information, 61, 101965. https://doi.org/10.1016/j.wpi.2020.101965

Li, S., Hu, J., Cui, Y., & Hu, J. (2018). DeepPatent: Patent classification with convolutional neural networks and word embedding. Scientometrics, 117(2), 721–744. https://doi.org/10.1007/s11192-018-2905-5

Lim, S., & Kwon, Y. (2016). IPC Multi-label Classification Based on the Field Functionality of Patent Documents. In J. Li, X. Li, S. Wang, J. Li, & Q. Z. Sheng (Eds.), Advanced Data Mining and Applications (pp. 677–691). Springer International Publishing. https://doi.org/10.1007/978-3-319-49586-6_48

Office, E. P. (n.d.). Cooperative Patent Classification (CPC). Retrieved September 9, 2022, from https://www.epo.org/searching-for-patents/helpful-resources/first-time-here/classification/cpc.html#:~:textThe%20Cooperative%20Patent%20Classification%20(CPC,%2C%20groups%20and%20sub%2Dgroups

Risch, J., & Krestel, R. (2019). Domain-specific word embeddings for patent classification. Data Technologies and Applications, 53(1), 108–122. https://doi.org/10.1108/DTA-01-2019-0002

Sofean, M. (2021). Deep learning based pipeline with multichannel inputs for patent classification. World Patent Information, 66, 102060. https://doi.org/10.1016/j.wpi.2021.102060

Soliman, A. B., Eissa, K., & El-Beltagy, S. R. (2017). AraVec: A set of Arabic Word Embedding Models for use in Arabic NLP. Procedia Computer Science, 117, 256–265. https://doi.org/10.1016/j.procs.2017.10.117

Tikk, D., Biro, G., & Törcsvári, A. (2008). A Hierarchical Online Classifier for Patent Categorization [Chapter]. Emerging Technologies of Text Mining: Techniques and Applications; IGI Global. https://doi.org/10.4018/978-1-59904-373-9.ch012

Xiao, L., Wang, G., & Zuo, Y. (2018). Research on Patent Text Classification Based on Word2Vec and LSTM. 2018 11th International Symposium on Computational Intelligence and Design (ISCID), 01, 71–74. https://doi.org/10.1109/ISCID.2018.00023

Zhu, H., He, C., Fang, Y., Ge, B., Xing, M., & Xiao, W. (2020). Patent Automatic Classification Based on Symmetric Hierarchical Convolution Neural Network. Symmetry, 12(2), 186; https://doi.org/10.3390/sym12020186