Explainable Artificial Intelligence for Highlighting and Searching in Patent Text

Renukswamy Chikkamath[1,*], Rana Fassahat Ali[2], Christoph Hewel[3] and Markus Endres[1]

[1]University of Applied Sciences, Munich, Germany
[2]University of Passau, Passau, Germany
[3]PAUSTIAN & PARTNERS, Munich, Germany

Abstract

The verbose content and redundant information present in patents often add complexity to reading and understanding them. Individual subject matters related to an invention and its decisiveness are scattered throughout patent documents. Moreover, these matters could provide relevant key arguments for an effective examination or critical assessment of an invention. To address these complexities and facilitate patent practitioners' efficient reading and in-page semantic searches of patents, we generated a multiclass dataset representing key arguments of patents on a sentence level. Essentially, these key arguments are the concrete details related to an invention, such as the problem it solves or the technical effects or advantages it achieves. We fine-tuned Transfer Learning models on this novel dataset and developed two Chromium extensions. One extension automatically highlights these key arguments using our fine-tuned model, and the other steers semantic search within any opened patent document in the browser. The data and code related to this work are released to the community via a GIT repository. The empirical test cases and manually labeled gold truth data provide evidence supporting our hypothesis regarding in-page patent search and efficient reading, respectively.

Keywords

Patent analysis, prior art search, patent language model, sentence classification, patent datasets

1. Introduction

1.1. Motivation

A patent is a form of intellectual property that provides the owner with legal rights to prohibit others from producing, using, or selling the invention. However, these rights are granted in exchange for disclosing how the invention works. Before a patent can be granted, it must undergo a rigorous examination process, known as the *prior art search*. This search is typically conducted in two stages: the first stage occurs in the early stages of the patent life cycle when patent attorneys draft the patent application. And the second stage takes place in the later stages of the patent life cycle when patent examiners review the patent application.

Since patent claims define the scope of protection, finding any prior art or other competing art that can be used as evidence for the proposed claims is a crucial step. A patent does not only comprise one or several claims defining the legal scope of protection but also a specification providing one or several specific embodiments of the invention. Patent owners tend to keep the specification as general as possible, which may not only be advantageous for further broadening the scope of protection but may also relieve the patent owners from publishing their developed technology. Therefore, most parts of the specification only repeat the text of the patent claims and add generalized boilerplate text concerning the functioning of an invention. Even if a patent specification may typically be 10 to 30 pages long, there are only a few short text passages that explain the concrete technical effects of the invention.

Therefore, it is often challenging for patent practitioners, including attorneys and examiners, to comprehend the invention's definition in the claims, which problem is addressed by the invention, or which technical effects or benefits are achieved by the invention. However, without understanding the motivation behind the invention, it is difficult to compare it with other inventions when assessing its inventive step over the prior art.

For example, suppose the claimed invention defines *a heating system with three temperature sensors*. In that case, the closest prior art document, such as an older patent, may only disclose *a heating system with two temperature sensors*. In such cases, important questions arise, such as what is the technical effect of the third sensor? why does the prior art suggest only two sensors? In case the motivations behind the two concepts are completely different, the claimed invention might be considered as implying an inventive step over the prior art.

***PatentSemTech'23**: **4th Workshop on Patent Text Mining and Semantic Technologies**, colocated with the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, July 27th, 2023, Taipei, Taiwan.*

*Corresponding author.

renukswamy.chikkamath@hm.edu (R. Chikkamath); ali11@ads.uni-passau.de (R. F. Ali); hewel@paustian.de (C. Hewel); markus.endres@hm.edu (M. Endres)

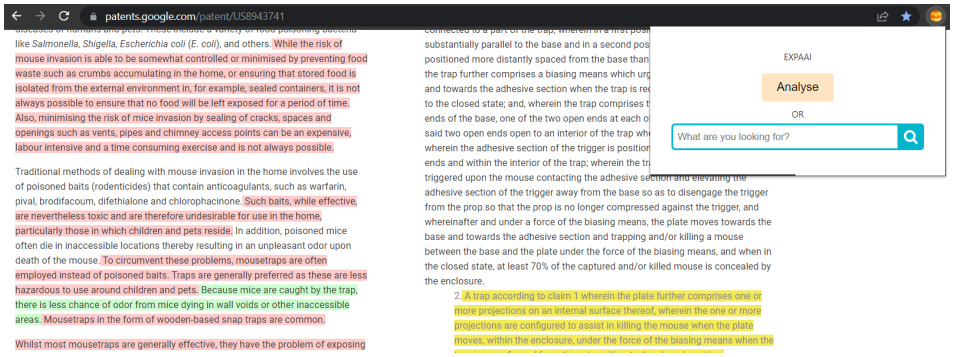

Figure 1: Chromium extension to highlight technical aspects. Analyse button activates this extension, wherein technical problems, solutions, and advantages are colored automatically in red, yellow, and green respectively.

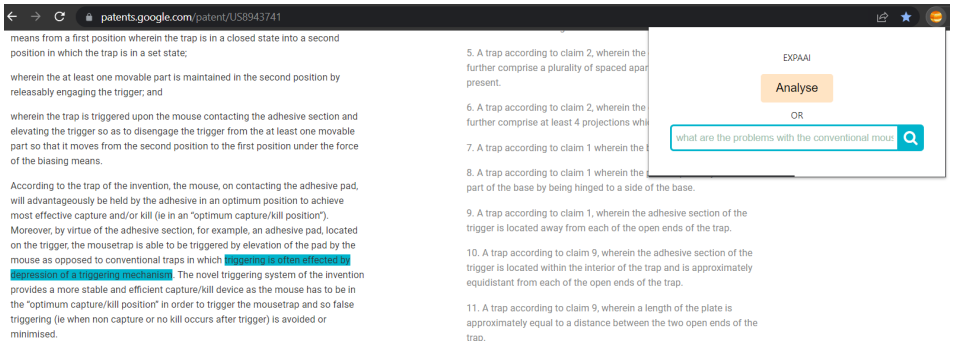

Figure 2: Chromium extension for in-page semantic search. A search bar can be used to ask a question and the answer is highlighted within opened web page or patent.

Consequently, patent analysis often requires retrieving those few text passages in a patent that can reveal the motivation behind the claimed invention. In this work, we aim to address the aforementioned difficulties and *ease the prior art search*. Specifically, we focus on *automatically highlighting* these text passages with the help of the Chrome extension (*Analyse*), as shown in Figure 1. The *Analyse* extension is supported by an Artificial Intelligence (AI) model that is fine-tuned on a *novel dataset* developed in this work. Also, we present a Chrome extension (search text box) to facilitate *cross-questioning* during patent analysis, as depicted in Figure 2.

1.2. Highlight and Search in Patent Text

The quality of a patent prior art search is greatly influenced by the readability and understandability of patents. In prior art search or patent analysis in general, the most important parts of patents are considered to be the claims and technical description which disclose and describe an invention respectively. Since the claims are written in legal terminology, they are often difficult to understand just by reading them alone. Detailed descriptions of patents represent key arguments of any invention, such as advantages, solutions, problems, and justifications for claim features. Understanding and differentiating the above mentioned points in a timely manner aids examiners and attorneys in critical assessments and effective analysis in the light of prior art. The focus of this work is to ease the readability and understandability of patents, unlike investigations of information retrieval or prior art search approaches.

The AI-based assistance presented in this work is in greater demand when individual patents are considered for analysis, and this assistance has two main benefits. Firstly, it provides *ease of readability* by automatically highlighting technical aspects related to the invention on the sentence level. Secondly, it offers *deeper understanding* by allowing readers to ask various cross-questions. For example, the question *What are the problems with conventional mouse catchers?* in the patent *Mousetrap*[1] can be searched, as shown in Figure 2. Such a tool can enhance the user experience by providing the opportunity to explore the documents in greater detail and work

[1] https://patents.google.com/patent/US8943741

with the semantics and context of the patent text, unlike keyword-matching in-page searches (Ctrl+F based search).

Highlighting at the sentence level is more interesting and important than at the keyword or paragraph level. This is because keywords in patents can be succinct but do not provide any evidence to understand the context in which key arguments are used. On the other hand, paragraphs can be informative but can contain mixed opinions. For example, individual sentences explaining different arguments of inventions (advantageous effects, problems, solutions) can be visible in one paragraph. Therefore, in this work, we focus on identifying and highlighting key arguments only at the sentence level.

In this paper, we present a *sentence-level patent dataset* designed to highlight key arguments for any invention at the sentence level. This is a multi-class dataset that was utilized to finetune Bert-for-Patents [1]. We developed two Chromium extensions: one for *automatically highlighting arguments*, facilitated by the internally finetuned Bert-for-Patents model; and another for *in-page semantic search* based on SQuAD[2] models. A free-flow natural language query can be used to search within the opened document on the web. Both extensions can work well on text present in any web page or document. However, to this end experiments are limited to Google patents[3], for example, a patent opened in Google Patents as shown in Figures 1 and 2.

The remainder of this work is organized as follows: Section 2 describes related work. Section 3 explains the methodologies used to develop the data, with a detailed multi-stage flowchart to describe the models developed in this work. Section 4 outlines the browser extension communication architecture. In Section 5, we discuss the results achieved in this work, including a sample test case. In the end, in Section 6, we conclude our work and suggest possible future directions.

2. Related Work

The research aspects of this work are related to the intersection of tasks such as text highlighting, sentence classification, and question answering in the field of Natural Language Processing (NLP).

In recent times, language representation learning, also known as language model development, and research on reading comprehension, such as question-answering models, have grown rapidly in the field of NLP. Notable models that have achieved top performance include Turing NLR-v5 [2] and Turing ULR-v6 [3]. These models outperform other state-of-the-art models in both sentence classification, such as the GLUE[4] benchmarked on the Stanford Sentiment Treebank (SST-2) dataset, and question answering, based on the Stanford Question Answering Dataset (SQuAD). Various other variants[5] of the BERT [4] architecture can also be seen as competitors in various settings. In recent years, Google released a language model pre-trained on patent data called BERT-for-Patents [1]. Since this model is trained on more than 100 million patents, unlike the above-mentioned general-purpose models, we have used it to fine-tune our classification model.

Text highlighting in this context emphasizes the significance of readability and understandability of patent and non-patent text. There is evidence in the literature regarding how patent examiners from the European Patent Office (EPO) initially read patent documents to come to a preliminary understanding of the patent. In particular, there is a greater need for developing tools to assist them in skimming through patents and achieving a deeper understanding of the contents [5]. Moreover, there is a lot of motivation from patent attorneys on the web to assess the parameters for patentability and skim through the document to find individual subject matters[6,7].

Although there is much interest in the readability of patents [6, 7, 8, 9], these approaches are limited to the analysis of claims. However, segmentation and analysis of claims are other segments of research in prior art search. To the best of our knowledge, there are no approaches that focus on patent text at the sentence level to highlight relevant key arguments. Highlighting important aspects of the text in the context of education/learning is not new [10]. In other non-patent domains, generating and providing a quick summary with highlighted text is proposed to emphasize textual elements [11, 12, 13, 14].

Text highlighting in general encourages a thorough understanding of a document [15] and also supports easier subsequent literature study [16]. To ease access, developing browser extensions to highlight text on the web has drawn attention. For instance, highlighting the disputed claims on the web pages and finding the relevant article from the web for facilitating the arguments in claims is proposed by Ennals et al [17]. Other related research also showed that reading comprehension can be attained by text highlighting on the web or any digital text content [18, 19, 14, 20].

In the patent domain, there are few private sectors that have developed solutions for multi-color highlighting of

[2] https://rajpurkar.github.io/SQuAD-explorer/
[3] https://patents.google.com/

[4] https://gluebenchmark.com/leaderboard
[5] https://huggingface.co/models?sort=downloads&search=bert
[6] https://www.heerlaw.com/difference-patentability-assessment-patent-search
[7] https://www.brmpatentattorneys.com.au/intellectual-property-law-melbourne/how-to-read-a-patent/

keywords[8,9]. However, such approaches would not be efficient because patent applications can be written using different terminologies even for the same concept. Furthermore, considering the context in addition to keywords adds domain knowledge that can explain why a particular keyword was highlighted. Additionally, these solutions are paid, and the reader has to manually find and highlight keywords. These solutions are more like digital pens to highlight and keep a record of keywords, which is again a time-consuming task.

To utilize AI models in the process of automatic highlighting in the patent domain, IPGoggles[10] (one of the motivations for this paper) proposes a new-age cloud-based solution. This service highlights keywords or even phrases in patents based on sentiment. Professionals believe that reading and understanding patents becomes challenging, even at an individual document level, given the huge amount of prior art. However, IPGoggles utilizes general-purpose AI models that are not fine-tuned on patent data to identify technical aspects or key arguments. In the patent domain, researchers have developed a dataset (PaSa) to identify the technical aspects of patent documents on a paragraph level [21]. It contains patent paragraphs named under the headings "Technical Problem," "Solution to Problem," and "Advantageous Effects of Invention."

In PaSa, United States Patent and Trademark Office (USPTO)[11] patent grants from 2010 to 2020 were searched to identify the technical aspects mentioned in clear and distinguishable paragraphs. The authors argue that these paragraphs are not common in all patents, but rather reflect a patent drafting style (based on region) that is mostly followed by Asia-specific patents. Moreover, it is even harder to find these specific paragraphs in Asia-specific patents before 2010 (refer to Table 4 [21], which shows a gradual decrease in the number from 2020 to 2010). This provides strong motivation to utilize these important and infrequent paragraphs as the basis of our investigations. However, the PaSa dataset has not been used in any downstream application or tool so far. Therefore, we decided to develop a dataset using PaSa and to use it to further train AI models that can be used in downstream applications, such as a Chrome extension.

In the state of the art, there is either evidence of highlighting technical aspects based on general-purpose AI models or evidence of a dataset to identify technical aspects on the paragraph level. However, to the best of our knowledge, there are no approaches that focus on *identifying and highlighting technical aspects on a sentence level using a domain and task-specific dataset.* Therefore, in this work, we propose and develop a dataset for finding technical aspects on a sentence level (refer to Section 1.2 to know why the sentence level is preferred). Furthermore, we utilize this dataset to fine-tune a patent domain-specific language model. This fine-tuned model is deployed in a Chrome extension service as a prototype. A detailed description of the technique utilized to develop the sentence-level dataset and the variety of models fine-tuned are described in the next Section 3.

[8] https://help.patsnap.com/hc/en-us/articles/115005478629-What-Can-I-Do-When-I-View-A-Patent-

[9] https://patseer.com

[10] https://ipgoggles.com/

[11] https://developer.uspto.gov/product/patent-grant-full-text-dataxml

3. Data and Models

To the best of our knowledge, there is no dataset available in the literature that identifies technical aspects at the sentence level. Therefore, we proposed to generate a sentence-level dataset based on a paragraph-level dataset called PaSa [21]. The patent paragraphs of PaSa (shown in the top left of Figure 3) represent essential key arguments that are crucial for effective patent reading. They also facilitate critical assessment of the boundaries of an invention. To aid patent practitioners in making decisions during report writing or formal hearings in examinations, AI models trained on such a dataset are necessary. However, it is not always true that all sentences in a specific paragraph represent the heading.

The following excerpt from the patent "US10834907B2" shows that there are sentences reflecting both problems and advantages under the same heading "Technical Problem".

For e.g., *"In summer, when rock oysters come in season, sea areas are highly contaminated... which causes inhibition of distribution...Accordingly, an object of the present invention is to provide ... enables the production of virus-free oysters having no experience of being exposed to a sea area.... present invention solves the above-mentioned problems."*.

Therefore, in this work, we utilized the PaSa dataset to develop sentence-level data for identifying the key technical aspects present in patents. We also used the "sentiments" naming convention for the three classes in our dataset, which are solutions-neutral, advantages-positive, and problems-negative.

The dataset generation and model training in this work can be seen in three stages, as shown in Figure 3. In Stage-I, as a straightforward approach, we used the NLTK tokenizer[12] to convert a paragraph into sentences based on full stops. Further, preprocessing was carried out to remove smaller sentences containing fewer than 20 characters, which are mostly small phrases or sentences oriented toward special symbols. After preprocessing, PaSa_Sentence-Baseline contains 940,000 sentences, and Figure 3 displays samples from each class. It is clear that

[12] https://www.nltk.org/

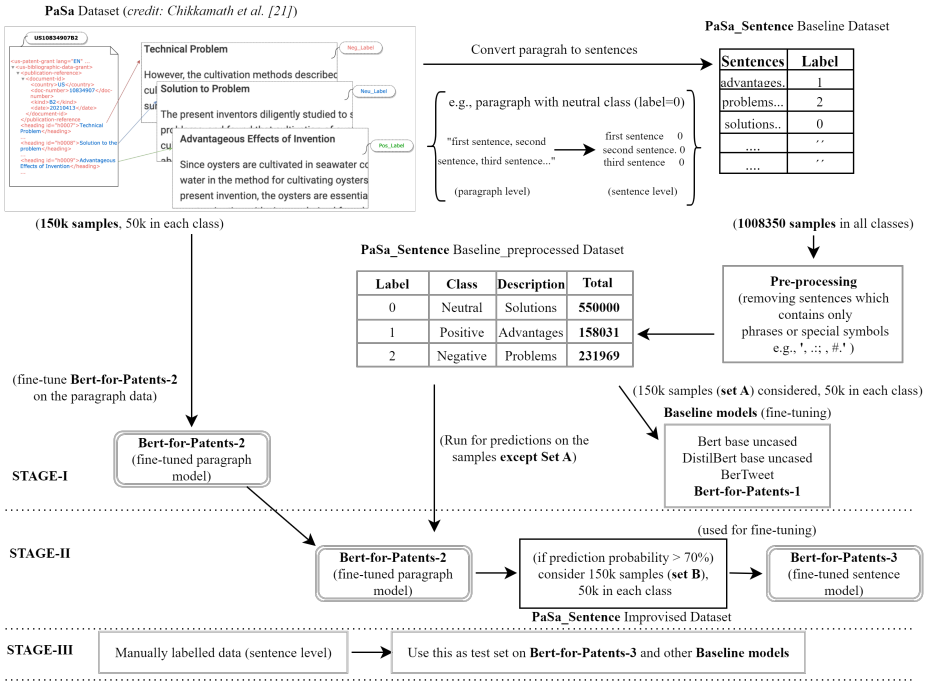

Figure 3: PaSa sentence level dataset generation and models including types and statistics of datasets in different settings.

the dataset is unbalanced as we have fewer samples in the positive and negative classes.

To maintain standard experimental settings, as in PaSa, and to avoid class imbalance problems, we chose only 150k samples (set A) to train the baseline models in Stage-I. The remaining samples were used for other experiments such as "except set A" which was used in Stage-II, and 650 samples for manual labeling of the data in Stage-III. In Stage-I, we also used the original PaSa paragraph dataset to train transformer models, as the PaSa paper focused only on machine learning models. In Stage-II, we generated an improvised version (set B) of the PaSa_Sentence Baseline data to address errors and shortcomings identified in using PaSa_Sentence Baseline (refer to Section 5.2 for error analysis). The data samples used for various purposes (set A, set B, manually labeled data) were kept completely non-identical to avoid bias in learning the models.

We utilized pre-trained transformer models from the Hugging Face platform[13] to fine-tune our datasets. With the exception of Bert-For-Patents, the remaining three baseline models (refer to Stage-I) were pre-trained on non-patent literature and hosted on Hugging Face. The naming convention (Bert-For-Patent-#) indicates that these models were fine-tuned on different datasets. For example, Bert-for-patent-2 is a completely new pre-trained model that was fine-tuned using PaSa paragraph data in Stage-II. In Stage-III, the same Bert-For-Patents-2 (fine-tuned) was used solely for making predictions on "except set A" (i.e., there was no role of "except set A" in training Bert-For-Patents-2). Thus, all models and datasets used were kept separate. The baseline models shown in Figure 3 were fine-tuned with a sequence length of 512 and batch size of 16, except for Bert-For-Patents-#, which was fine-tuned with a sequence length of 128 and batch size of 8. The reason for this difference is that Bert-For-Patents-# is an extremely large architecture with 24 hidden layers and creates hardware dependencies during fine-tuning, even for an NVIDIA server with an A30 GPU.

And for the in-page patent semantic search, we have used SQuAD-dataset based question-answering models[14] hosted on Hugging Face. The best-performing and most downloaded models are Bert Large (uncased), RoBerta base, and DistilBert based (cased). To the best of our knowledge, no datasets are available in the state-of-the-art with SQuAD format in the patent domain (which opens the door for research in developing a question-answering dataset in the patent domain). SQuAD models

[13] https://huggingface.co/models

[14] https://huggingface.co/models?pipeline_tag=question-answering&sort=downloads

are feasible for in-page searching in this work because natural text queries can be searched within a given context (e.g., patent text in chunks), unlike keyword matches. SQuAD models can be easily hosted and deployed in Chrome extensions. Therefore, we investigated the aforementioned models in our in-page semantic search extension. The components of the Chromium extension are explained in detail with the help of communication architecture in the next Section 4.

4. Browser Extensions

The browser extensions developed in this work are aimed at enhancing the readability and understandability of patents. Readability is more effective when the technical aspects of the considered patents are automatically highlighted. This automation is based on knowledge from domain-specific AI models fine-tuned in this work, and the respective model is deployed in a Chrome extension (refer to Figure 1). The understandability of patents is improved when there is an opportunity to ask cross-questions during patent analysis within a patent document. Such a feature is provided by our other extension developed in this work (refer to Figure 2). Patent practitioners can install and activate these two Chrome extensions in their browsers for effective prior art searches (refer to the GIT repository[15] of this work for installation). More details including the usability of the Chrome extension, request run times, and responsiveness of the interface are also added to the GIT repository.

The browser extensions presented in this paper operate on the browsers such as Google Chrome (Chromium based), with development in two parts: i) Python Flask[16] API for models (acts as backend) and ii) Chromium extension (acts as front end). We used Flask to develop an API for our models, further to get the predictions from our fine-tuned models we utilized Hugging Face transformers pipelines[17]. We hosted our fine-tuned models in the Hugging Face repository to make use of them in pipelines. The API has two POST endpoints one for each of the tasks (classification/sentiment-predict and in-page semantic search). The classification POST endpoint accepts an array of sentences of any opened document in the browser and collects the prediction response from the transformer pipeline with our fine-tuned model (BertForPatents-3). Further, the endpoint will assign classes to the array of sentences. With respect to the semantic search POST endpoint, a context (complete patent text in our case) and question are given as input and passed to question-answering model pipeline (e.g., Bert large trained on SQuAD from Hugging Face[18]). The response will be an answer (start and end positions of text from the context considered) for the question searched.

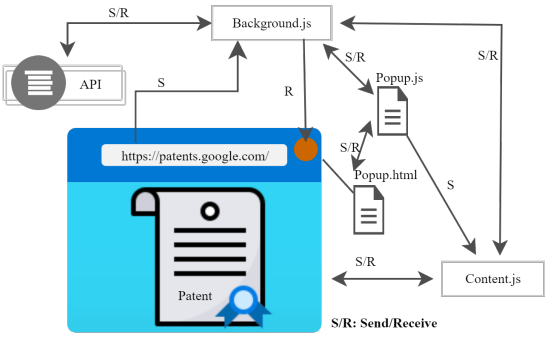

Figure 4: Browser extension communication architecture with its components.

We use **chrome-extension-cli**[19] for developing the Chromium extension. In addition, we used technologies such as Javascript, HTML, and CSS for data handling and styling. The communication architecture of the browser extension with its components is shown in Figure 4. The functionalities of individual components are as follows:

- **Popup:** The component that is visible when we click the browser extension icon, which acts as the only point of contact between the user and the extension. The popup is responsible for providing buttons for both classifications with multicolor highlighting and a search bar. Additionally, the Loader shows the task being performed or stopped. The Popup script communicates with both the "Content" and "Background" components. Text content from the web page will be accessed, analyzed (predictions, answers), and highlighted in the final step.
- **Content:** This component collects the text present in the opened web page and communicates with both the "Background" and "Popup" components. The "Content" component is responsible for receiving a message from the "Popup" script and for sending and receiving messages to and from the "Background" component. In this case, it prepares the content for analysis and highlights the relevant content on the web page based on predictions from the "Background" component. Highlighting the content (sentences and answers) is one of the salient tasks of the "Content" component. This is achieved by using a

[15] https://github.com/Renuk9390/expaai_model
[16] https://flask.palletsprojects.com/en/2.2.x/
[17] https://huggingface.co/docs/transformers/main_classes/pipelines
[18] https://huggingface.co/bert-large-uncased-whole-word-masking-finetuned-squad
[19] https://github.com/dutiyesh/chrome-extension-cli

"div" number or "class" on the HTML page for the respective matched answer or sentence to highlight.

- **Background:** This is the only component communicating with the Flask API backend. When it receives a message from the "Content" component with a payload to perform a task, the API endpoint will be called with inputs. Background listens to two types of messages from Content such as *"Patent_Text"* for highlighting technical aspects based on the type of class it belongs to and *"Patent_Semantic_Search"* to accomplish in-page search. After receiving a response from API, the response will be sent to "Content" for further processing. In addition, Background is also responsible for sending messages *task_started* and *task_stopped* to "Popup" to keep the "Loader" busy or active for taking the next task from the user. More details on the communication of components can be collected via the code base repository of this paper.

5. Findings

In this section, we discuss the results of this work and perform an error analysis to show how the dataset representation problem affects the model performances.

5.1. Scores and Test Cases

Table 1 displays the classification accuracies of the models developed in this work using the PaSa_sentence Baseline_preprocessed dataset. Bert-for-Patents-1 exhibits better performance than the other models, possibly because it was pre-trained by Google on patent literature. As a result, we opted to employ only the Bert-for-Patents pre-trained architecture in Stage-II.

Table 1
Classification scores on sentence level

Data Size	Model	Accuracy
150k	BerTweet	80%
150k	Bert base	83.5%
150k	DistilBert	84%
150k	Bert-for-Patents-1	**86.30%**

"PaSa_Sentence Improvised Dataset" (refer to Stage-II in Figure 3) is used to fine-tune Bert-for-Patents-3. Due to the improvements made in the dataset, this model shows an accuracy of 97.11%. As shown in Figure 3, Bert-for-Patents-2, fine-tuned on a paragraph level with an accuracy of 98.13%, is competent enough to represent the classes. Therefore, we decided to use Bert-for-Patents-2 to obtain improved samples from our "Baseline_preprocessed Dataset". We considered only those samples where the prediction score was greater than 70% when predicted by Bert-for-Patents-2.

With respect to in-page semantic search, we are utilizing models (Bert Large uncased, RoBerta base, and DistilBert based cased) which are fine-tuned on SQuAD data. To our knowledge, there are no SQuAD formatted datasets in the patent domain to address in-page question answering. Therefore in this work, we are not fine-tuning them on any patent data. Instead, we only perform test cases to compare and evaluate them. For the test cases, we considered various contexts (patent text) and questions to compare the answering capability of said models. DistilBert is competitive with Bert Large in some cases. For instance, as depicted in Figure 5, we provided the same context and question to the aforementioned models. Bert Large exhibited superior performance in retrieving the answer; nevertheless, DistilBert also performed reasonably well in retrieving the correct answer. In most cases, Bert Large uncased model performed better in finding accurate answers for longer queries (which are common in patent searches). Therefore, Bert Large is deployed in the in-page semantic search extension.

To test and debug the API endpoints for intended functioning, we used an open-source application called Insomnia[20]. We provided Insomnia test requests to the in-page semantic search API and the classification (aka sentiment_predict) API endpoints. For example, we passed an array of sentences to the *sentiment_predict* API endpoint, and the fine-tuned model returned a response with the label and prediction probability score. Similarly, for *semantic_search*, we passed a sample patent text as a context along with a question, and the retrieved response included the begin and end token numbers of the possible answer text snippet with confidence scores. After confirming the intended functioning of the APIs using Insomnia tests, we deployed the APIs in the Chromium extensions.

5.2. Error Analysis

There are three different ways in which the labels are assigned to the sentence level dataset of this work. Firstly, automatic labeling is based on the NLTK tokenizer (in STAGE-I). Secondly, labels are given by fine-tuned paragraph model (in STAGE-II). And thirdly, manually assigned labels (in STAGE-III). Although "Baseline models" developed in this work show good performances in terms of accuracy, there are cases where the models' validation loss is less than the training loss at the end of 3rd epoch. The validation data was easier to predict than learning the training data for the models. This signifies a dataset

[20] https://docs.insomnia.rest/insomnia/get-started

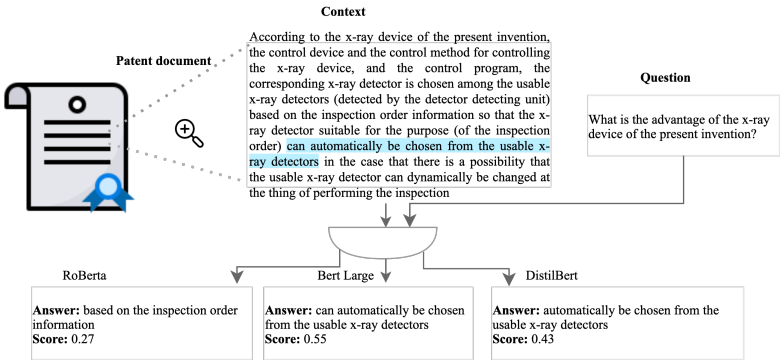

Figure 5: An example SQuAD-based in-page patent semantic search tested on different models

representation problem, i.e., classes are not equally represented by all the samples because of various reasons as shown below. The models finetuned on this poorly represented data induce bias in predicting the validation set. There are various samples in PaSa_Sentence Baseline_preprocessed data which can be examples of substandard training samples.

Example 1: *"In the view of the problem of the background art, it is an object of the present invention to provide a conveyor which estimates the weight of a transport object while it is carried without using devices such as a load cell which directly measures weight."*

Observation 1: The above example is automatically labeled as a negative class during PaSa_Baseline generation, but it is not when we do manual labeling. During patent drafting, mostly in "Technical Problem" paragraphs, attorneys/applicants commonly use underlined phrases to quickly repeat their invention while describing problems with other prior art. If sentences with such underlined phrases are present in the negative class, then such samples can be discarded.

Example 2: *"An embodiment provides a lighting device in which an optical plate is disposed on at least one light source and a light source module including the same."*

Observation 2: The above sentence, as well as others that are similar, are automatically labeled as negative even though they are not. This indicates the presence of mixed opinions at times on the paragraph level, which also appears in some sentences.

There are other samples that are very long (60-70 words); in such cases, smaller sentences are joined using special symbols such as ";,:". Manually checking every such sample in large datasets is laborious. Therefore, we decided to fine-tune a model on the paragraph level so that this model would have a greater understanding of the representativeness of classes on advantages, problems, and solutions in a patent text. Such a fine-tuned model is used to consider the sentences that show at least a 70% probability of representing a class. Further, we have used these improvised samples (PaSa_Sentence Improvised Dataset) to fine-tune a new model (Bert-For-Patents-3), which outperforms other baseline models in terms of both accuracy and class representativeness.

We manually labeled 650 randomly selected samples, which were not used in any of the experiments. The original labels for these samples from PaSa_Baseline_preprocessed were kept separate. To verify the presence of bias and representation problems in the baseline models, we compared the prediction accuracies of manual predictions, baseline models, and Bert-for-Patents-3. The manual and Bert-for-Patents-3 prediction accuracies were 68.59% and 69.05%, respectively. Bert-for-Patents-3 was fine-tuned on the improved dataset, and its prediction performance was closer to the manual labels. However, due to bias, the baseline models showed higher scores with accuracies of 87.80% (DistilBert base uncased), 87.04% (Bert base uncased), and 94.06% (Bert-for-Patents-1). Therefore, Bert-for-Patents-3 is more suitable for use in the Chrome extension for highlighting technical aspects.

Technical aspects in a patent represent advantages over the prior art, proposed solutions, or problems with other prior art. The core objective of this work was to automatically identify and highlight these aspects in patents. Although this objective may resemble a sentiment analysis problem, general sentiment analysis datasets or algorithms are not suitable for this task. Our sentence-level dataset is distinct from other sentiment analysis datasets such as IMDB[21] and Amazon product reviews[22]. These datasets mostly contain sentences expressing people's opinions on products, things, or other social aspects. In contrast, our dataset highlights the key technical arguments in patents that demonstrate the invention's technical capabilities in comparison to the prior

[21] https://www.imdb.com/interfaces/
[22] https://cseweb.ucsd.edu/~jmcauley/datasets.html

art. Most importantly, our dataset is specific to the patent domain and accounts for patent-specific vocabulary and knowledge.

6. Conclusion and Future Work

In this work, we present a multi-class dataset at the sentence level to highlight the technical subject matters of patents, which can serve as important key arguments to determine a patent's novelty. We fine-tuned language models on our new dataset and developed a Chromium extension to automatically highlight key arguments based on predictions, provided the probability exceeds 70%. We also developed another Chromium extension to facilitate in-page semantic search.

We anticipate a growing need for AI-based tools to assist patent practitioners in conducting patent prior art searches. We hope this empirical work serves as preliminary research and motivates researchers and patent practitioners to develop tools that can automate prior art searches. Future work in this area could identify additional technical aspects in patent documents and train new classes for highlighting. For this study, we focused only on advantages, problems, and solutions. Furthermore, sentence-level data could be improved to enhance the representativeness of samples belonging to a particular class. For example, sentences representing "advantages" should not be mixed with sentences related to "problems".

Developing a question-answering dataset in the patent domain is crucial, and such datasets can be used to develop tools to automate in-page semantic searches. We also hope that AI-based tools to assist prior art searches will enhance the interaction of patent analysts with patent documents. For instance, the automatic highlight and semantic search tools prototyped in this work can allow for cross-questioning within any patent document opened in a web browser.

Acknowledgments

This research is part of the project "BigScience", which is funded by the Bavarian State Ministry for Economic Affairs, Regional Development, and Energy under the grant number DIK0259/01.

References

[1] R. Srebrovic, J. Yonamine, Leveraging the BERT algorithm for Patents with TensorFlow and BigQuery, Technical Report, Technical Report. Global Patents, Google https://services. google. com/fh …, 2020.

[2] P. Bajaj, C. Xiong, G. Ke, X. Liu, D. He, S. Tiwary, T.-Y. Liu, P. Bennett, X. Song, J. Gao, Metro: Efficient denoising pretraining of large scale autoencoding language models with model generated signals, arXiv preprint arXiv:2204.06644 (2022).

[3] B. Patra, S. Singhal, S. Huang, Z. Chi, L. Dong, F. Wei, V. Chaudhary, X. Song, Beyond english-centric bitexts for better multilingual language representation learning, arXiv preprint arXiv:2210.14867 (2022).

[4] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).

[5] P. Lahorte, Inside the mind of an epo examiner, World Patent Information 54 (2018) S18–S22.

[6] A. Shinmori, M. Okumura, Y. Marukawa, M. Iwayama, Patent claim processing for readability-structure analysis and term explanation, in: Proceedings of the ACL-2003 workshop on Patent corpus processing, 2003, pp. 56–65.

[7] G. Ferraro, H. Suominen, J. Nualart, Segmentation of patent claims for improving their readability, in: Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR), 2014, pp. 66–73.

[8] A. Shinmori, M. Okumura, Y. Marukawa, Aligning patent claims with detailed descriptions for readability., in: NTCIR, 2004.

[9] S. Sheremetyeva, Natural language analysis of patent claims, in: Proceedings of the ACL-2003 workshop on Patent corpus processing, 2003, pp. 66–73.

[10] L. Rello, H. Saggion, R. Baeza-Yates, Keyword highlighting improves comprehension for people with dyslexia, in: Proceedings of the 3rd workshop on predicting and improving text readability for target reader populations (PITR), 2014, pp. 30–37.

[11] S. Spala, F. Dernoncourt, W. Chang, C. Dockhorn, A web-based framework for collecting and assessing highlighted sentences in a document, in: Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations, 2018, pp. 78–81.

[12] J. J. Li, K. Thadani, A. Stent, The role of discourse units in near-extractive summarization, in: Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue, 2016, pp. 137–147.

[13] K. Woodsend, M. Lapata, Automatic generation of story highlights, in: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, 2010, pp. 565–574.

[14] M. Kaisser, M. A. Hearst, J. B. Lowe, Improving search results quality by customizing summary

lengths, in: Proceedings of ACL-08: HLT, 2008, pp. 701–709.
[15] F. I. Craik, R. S. Lockhart, Levels of processing: A framework for memory research, Journal of verbal learning and verbal behavior 11 (1972) 671–684.
[16] H. W. Faw, T. G. Waller, Mathemagenic behaviours and efficiency in learning from prose materials: Review, critique and recommendations, Review of Educational Research 46 (1976) 691–720.
[17] R. Ennals, B. Trushkowsky, J. M. Agosta, Highlighting disputed claims on the web, in: Proceedings of the 19th international conference on World wide web, 2010, pp. 341–350.
[18] M. Yeari, M. Oudega, P. van den Broek, The effect of highlighting on processing and memory of central and peripheral text information: Evidence from eye movements, Journal of Research in Reading 40 (2017) 365–383.
[19] J. A. Brown, K. Knollman-Porter, K. Hux, S. E. Wallace, C. Deville, Effect of digital highlighting on reading comprehension given text-to-speech technology for people with aphasia, Aphasiology 35 (2021) 200–221.
[20] A. Winchell, A. Lan, M. Mozer, Highlights as an early predictor of student comprehension and interests, Cognitive Science 44 (2020) e12901.
[21] R. Chikkamath, V. R. Parmar, C. Hewel, M. Endres, Patent sentiment analysis to highlight patent paragraphs, arXiv preprint arXiv:2111.09741 (2021).