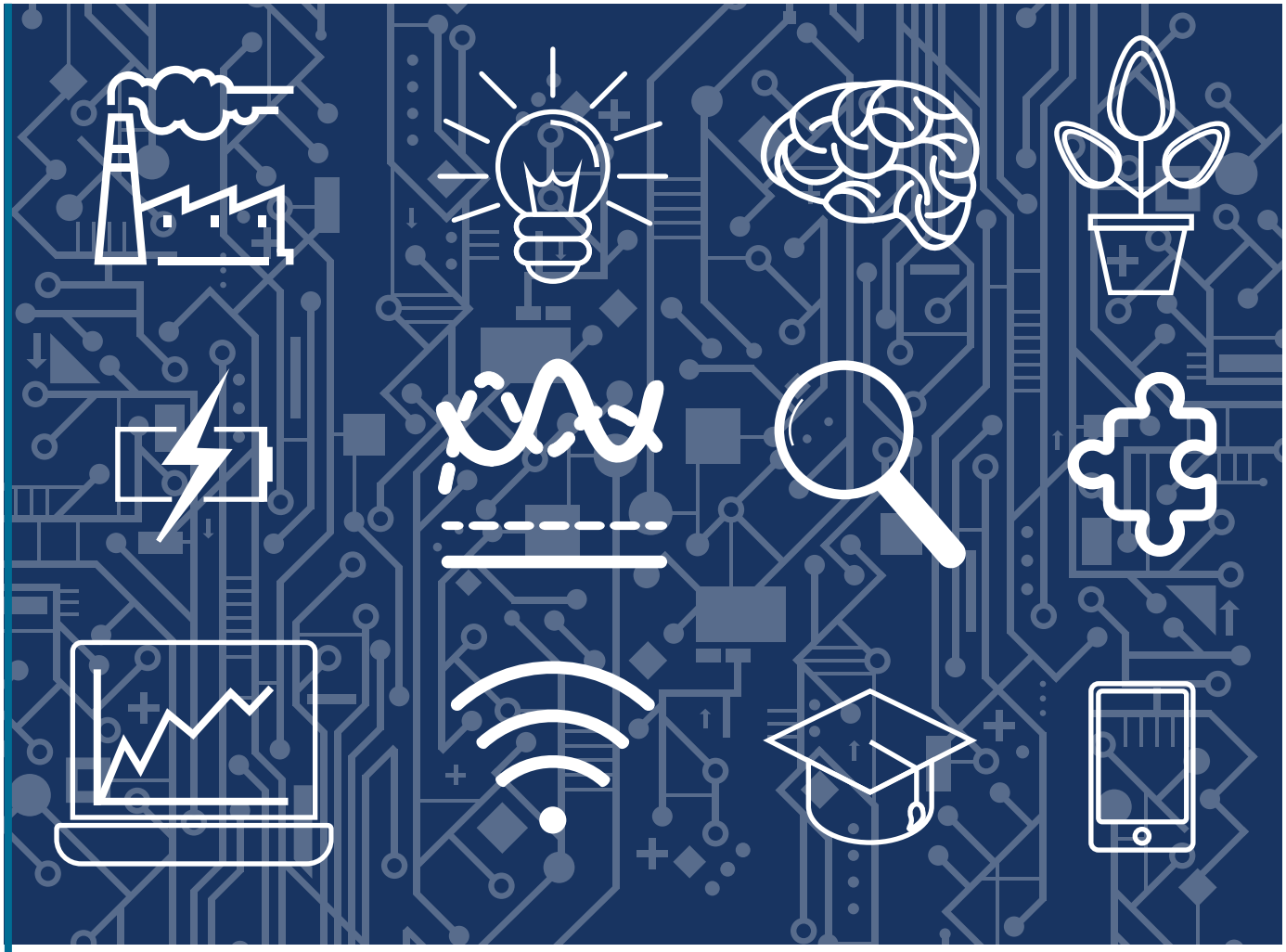
Intellectual Property Office

CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

AI-assisted patent prior art searching - feasibility study



Research commissioned by Intellectual Property Office (IPO) and carried out by Cardiff University with funding from the BEIS Regulators' Pioneer Fund.

IPO | Making life better through IP

Intellectual Property Office is an operating name of the Patent Office

Cardiff University Core Research Team:

Rossi Setchi is Professor in High-Value Manufacturing at Cardiff University. She joined the School of Engineering in 2000 after completing her PhD in Intelligent Systems. She was promoted to Personal Chair in 2011. She has a distinguished track record of research in a range of areas including AI, robotics, systems engineering, manufacturing, industrial sustainability, Cyber-Physical Systems and Industry 4.0, and has built international reputation for excellence in knowledge-based systems, computational semantics and human-machine systems. Professor Setchi leads the Research Centre in AI, Robotics and Human-Machine Systems at Cardiff. She is Fellow of IMechE, IET and BCS, and Senior Member of IEEE.

Irena Spasić received a PhD degree in computer science from the University of Salford, UK in 2004. Following posts at the Universities of Belgrade, Salford and Manchester, she joined Cardiff School of Computer Science & Informatics in 2010, and became full professor in 2016. Her research interests include text mining, knowledge representation, machine learning and information management with applications in healthcare, life sciences and social sciences. She is a Director of the Data Innovation Research Institute at Cardiff University and is a co–founder of the UK Healthcare Text Analytics Research Network (HealTex).

Cardiff University would like to give special thanks to Jeffrey Morgan and Fernando Loizides (Cardiff University School of Computer Science and Informatics) as well as the IPO project board team (Chris Harrison, Rich Corken, Maurice Blount, Peter Thomas-Keefe, Peter Evans, Kingsley Robinson, Stephen Otter, James Selway, Julia Leighton), and the IPO patent examiners (Kunal Saujani, Terence Newhouse, Alessandro Potenza, Chris Bennett, David Kirwin, Tom Simmonds, Caroline Bird, Manolis Rovilos) who were engaged with the core research team in examiner interviews, experimental testing and project evaluation.

ISBN 978-1-910790-80-9

AI-assisted patent prior art searching – feasibility study

Published by Intellectual Property Office
April 2020

1 2 3 4 5 6 7 8 9 10

Any enquiries regarding this publication should be sent to:

Intellectual Property Office
Concept House
Cardiff Road
Newport
NP10 8QQ

Tel: 0300 300 2000 Fax: 01633 817 777

e-mail: research@ipo.gov.uk

This publication is available from our website at www.gov.uk/ipo

Contents

Executive summary 1

1 Introduction 3

1.1 Context 3

1.2 Aim and objectives 3

1.3 Literature review 4

1.4 User-centred approach and scope 5

2 Observations and interviews 5

2.1 Prior art searching as a process 5

2.2 User requirements 6

2.3 Key challenges 6

2.4 Technical requirements 7

2.5 Indicators and measures 7

3 Proof-of-concept 8

3.1 AI techniques considered 8

3.2 Concept model 9

3.3 Description of system functionalities 9

3.3.1 Feature extraction 9

3.3.2 Query expansion 12

3.3.3 Document classification 13

3.3.4 Topic modelling 14

3.3.5 Document similarity 15

4 Evaluation 15

4.1 Experimental protocol 15

4.1.1 Quantitative testing 17

4.1.2 Qualitative testing 17
4.1.3 Focus group 17
4.2 Results and discussion 18
4.2.1 Classification 18
4.2.2 Topic modelling 18
4.2.3 Information retrieval 25
4.2.4 Usability 31
5 Conclusions 32
References 33
Appendices 35
Appendix 1: Software libraries 35
Appendix 2: Validation domains 36
Appendix 3: Data format 36
Appendix 4: Multi-word terms extracted 37
Appendix 5: Multi-word term dendrograms 47
Appendix 6: Domain-specific uses of the word "driver" 50
Appendix 7: Nearest neighbours of the word "driver" in the word embeddings space 53
Appendix 8: Representation of the word "driver" in WordNet 55
Appendix 9: Local installation of Elasticsearch service 56
Appendix 10: The results of cross-validation classification experiments 59

Executive summary

The Intellectual Property Office (IPO) commissioned Cardiff University to conduct a research study to understand the feasibility, technical complexities and effectiveness of how artificial intelligence (AI) solutions could benefit IPO during prior art searching of patent applications. In particular, IPO is interested in a proof-of-concept for an AI-powered prior art search/due-diligence check that could form part of the online patent filing and patent examiner prior art searching processes.

Patent searching is a highly interactive and complex process often requiring multiple searches, diverse search strategies and search management. From an AI point of view, the key linguistic and semantic challenges are legal wording, long sentences, acronyms, and the technical nature of patent claims.

The specific objectives of the study are to evaluate the viability of different AI technologies for patent prior art searching, test different approaches to identity the most effective algorithms, and fully evaluate an optimal solution. A wide range of state-of-the-art supervised and unsupervised machine learning approaches are considered that could support the tasks of feature extraction, query expansion, document classification, document clustering and topic modelling. These include:

- Natural language processing: text segmentation, normalisation, lemmatisation, stemming, co-occurrences, multi-word terms;
- Supervised machine learning: support vector machine, naive Bayesian learning, decision tree induction, random forest;
- Unsupervised machine learning: word embeddings, distributional semantics, neural networks, deep learning;
- Semantic technologies: use of lexico-semantic knowledge, latent Dirichlet allocation.

The research concludes that it is not feasible with current AI tools to provide a fully automated solution as part of the patent application filing process. Patents are manually classified into technology areas by examiners but this research found that an automated classification task produces very high classification accuracy, which shows potential to embed this function in the online patent pre-filing process to allow customers thinking of applying for a patent to more easily undertake due diligence checks.

The viability of the different AI technologies for patent prior art searching is considered and the research finds clear evidence that none of the available AI algorithms on their own can support every aspect of the prior art search process (e.g. classification, forming a search query, retrieval, ranking, identifying similarities and topic visualisation). The intention however is not to design a fully functional information retrieval system but to develop a proof-of-concept that enables experimental comparisons between different approaches. The study shows that different state-of-the-art AI algorithms can be used to retrieve the closest documents, rank relevant documents, suggest synonyms, suggest classifications, cluster and visualise the retrieved documents/concepts.

The developed concept model follows a user-centred design by considering the needs, wants and limitations of patent examiners throughout the prior art searching process. As a result, this human-in-the-loop approach aims to maximise performance by combining AI and human intervention and is designed to supplement, not substitute, human expertise and judgment. The chosen AI algorithms support the user in navigating through large volumes of patent data by suggesting the most plausible search terms and categorisations of patents into easily interpretable topics. In this scenario, the user keeps the role of the key decision maker, whereas the AI provides intelligent decision support.

To support practical experiments, a system based on a proposed concept model is implemented in the programming language Python. For the purposes of this feasibility study, three domains are used to validate the system experimentally throughout its development:

- civil engineering;
- computer technology; and
- transport.

These three domains are chosen because they are the top three technology fields based on number of filings at IPO over the past 10 years. The developed proof-of-concept is trained on English-language patent data from the PATSTAT bibliographic database of worldwide patents, GB full-text patents, EP full-text patents and US full-text patents. Qualitative testing with patent examiners is then undertaken using a number of 'query' patents in each of the three domains.

These experiments conducted with expert patent examiners strongly suggests that the use of AI techniques to retrieve and rank documents could reduce the time and cost of prior art searches, and especially the process of sifting through the large number of patents retrieved. The experimental results for precision varies between 30% and 50%, which means that the first 10 search results contain between 3 and 5 relevant documents. Patent examiners involved in this feasibility study agree that this was a higher 'hit rate' than they achieve with their current search tools. This proof-of-concept for an AI-powered patent prior art search therefore shows that AI has the potential to assist patent examiners in the future as part of the prior art searching process.

However, an AI-assisted search will require a patent examiner to manually formulate a search statement; there are currently no effective AI algorithms which can automatically process the application and generate a search statement, which is one of the most important and knowledge-intensive parts of the prior art searching process. The construction of the search statement requires clear understanding of the critical subject matter and the potential novelty of the patent application; this should remain a human task to suitably bound the AI search because of the wealth of specialist expertise and experience that a patent examiner has, and is not something to be performed by AI.

The patent examiners involved in testing the proof-of-concept make a number of suggestions about how the system performance could be improved. This includes using flexible search strategies (e.g. using different parts of the patent text at different stages of the search process, selecting the most relevant paragraphs to the crux of the invention to make the retrieval task more focused, changing the weighting of the search parameters), hybrid search strategies (e.g. combining text and picture searches) and knowledge-based search strategies (e.g. enhancing the search with knowledge types such as method, process, methodology, etc.) and using domain-specific ontologies.

Experiments conducted as part of the study highlight significant differences in the search strategies employed by the patent examiners and the need for more innovative tools in the future which support more flexible search strategies. There are opportunities to enhance the current prior art search process by developing new tools for retrieving image-based patents, collecting evidence of due diligence, spotting ambiguity, finding contradictions and visualising relationships among documents.

In conclusion, the study evaluates the viability of different AI technologies for patent prior art searching, including supervised and unsupervised machine learning, and finds clear evidence that none of the available AI algorithms on their own can support every aspect of the prior art search process. The proof-of-concept developed as part of this research uses different state-of-the-art AI algorithms for the different parts of the patent prior art searching process. Experimental results give a higher 'hit rate' than patent examiners achieve with their current search tools which shows that an AI tool has the potential to assist patent examiners in the future as part of the prior art searching process. The study identifies the potential of new approaches combining AI with NLP and computational semantics, and highlights the importance of human-centred decision and performance support tools. There is however a need for further work with larger scale and more rigorous testing with a larger collection of patents and more patent examiners across a wider breadth of different technology areas, as well as more cutting-edge research on new algorithms supporting flexible search strategies and a dynamic, iterative search process.

1 Introduction

1.1 Context

As part of the Government's Better Regulation agenda, BEIS (Department for Business, Energy and Industrial Strategy) are encouraging regulators to look at innovation-friendly frameworks and approaches. The £10m Regulators' Pioneer Fund (RPF) was launched in 2018 to drive forward this innovation-friendly agenda and help unlock the long-term economic opportunities identified in the Government's modern Industrial Strategy. Regulators are being asked to consider new and emerging issues where they might be able to collaborate productively with others and help businesses bring innovative new products to market.

Artificial intelligence (AI) is one of the Industrial Strategy Grand Challenges and an area of emerging technology. Intellectual Property Office (IPO) has a working group investigating how they can embrace AI technologies. This includes modernising internal operational processes and the application process for customers applying for intellectual property rights (IPRs), which includes patents, trade marks and designs. The technical nature of patent specifications, and the legal wording used in patent applications in particular, raises a number of AI challenges.

As part of this innovation agenda, IPO commissioned Cardiff University to conduct a study to understand the feasibility, technical complexities and effectiveness of how AI solutions could benefit IPO customers during the filing and prosecution of patent applications. This study was funded by the £10m Regulators' Pioneer Fund.

In particular, IPO was interested in a proof-of-concept for an AI-powered prior art search/due-diligence check that could form part of the online patent filing and patent examiner prior art searching processes. Prior to filing, this could inform a patent application of the most relevant prior art that exists and that may hinder their patent application being expedited to grant. The results of this search would also be passed on to the patent examiner undertaking the patent prosecution, who could use the results to inform their search with potentially significant time, and therefore cost, savings to be had.

The aim is to reduce the time and cost of patent prior art searches/due-diligence checks and subsequently improve the quality of the patent examination process. Customers applying for IPRs may benefit from faster handling of their patent applications. This aim is in line with the vision to provide automated search tools that complement the patent examiners' knowledge and expertise (Andlauer, 2018).

This research will also help to answer a number of technological challenges currently facing the IP community, including the suitability of AI for patent searching given the technical nature of patent specifications and the terminology used. As acknowledged in a recent paper (Krishna et al., 2016), fully automated prior art retrieval systems are challenged by the technical content of the patents and the subtleties in interpretation of patent laws, which are influenced by recent court decisions.

1.2 Aim and objectives

The specific objectives of the study are to:

- evaluate the viability of different AI technologies for patent prior art searching;
- test different approaches to identity the most effective algorithms;
- fully test and evaluate of an optimal algorithm.

1.3 Literature review

The success of a prior art search relies upon the selection of relevant search queries (Bashir and Rauber, 2010). An important component of a successful search process is the transformation of a human query (search request) into a query representation (Crestani, 2003). This process is influenced by the patent examiner's background experience in the technical field, their knowledge, communication and presentation skills, the reputation for trust and reliability that they have built up and their approach to teamwork (Adams, 2018).

Typically, terms for prior art queries are extracted from the claim fields of query patents. However, selecting relevant terms for queries is a difficult task due to the complex technical structure of patents and the presence of mismatched and vague terms; this often involves further research into the domain of the application.

Furthermore, patents are complex legal documents, even less accessible than the scientific literature. As a text genre, the patent domain is associated with several characteristics: huge differences in length, strictly formalised document structure (both semantic and syntactic), extensive use of standard and non-standard acronyms and domain terminology (Anderson et al., 2017). Patent drafters intentionally try to use entirely different word combinations, not only synonyms but also paraphrasing (Atkinson, 2008). Patentees typically use their own lexicon in describing their inventive details or use abstract or generic terms to maximise the protective scope. Patents often include different data types – typically drawings, mathematical formulas, bio-sequence listings or chemical structures that require specific techniques for effective search and analysis.

In addition to the standard metadata (*e.g.* title, abstract, publication date, applicants, inventors), patent offices typically assign some classification coding to assist in managing (allocating) their examination workload and in searching patents, but these classification codes are not consistently applied or harmonised across different patent offices (Alberts et al., 2017). The diachronic aspect of the patent text genre contributes not only to sparse events but also to changes in terminology, where one term may refer to a technical concept during a certain time period and thereafter may switch to represent another (Anderson et al., 2017, Harris et al., 2017). The existing diachronic nature and vocabulary diversity within part of the patent text genre make it more difficult to sample out data in order to establish a training set for text mining applications (Oostdijk et al., 2017).

The ongoing debate among patent professionals about the relative value of full-text versus controlled indexing (Adams, 2018) reveals open questions about search quality and whether full-text search strategies generate too much irrelevant material (low precision searching) or are more prone to miss relevant answers due to unexpected variation in terminology in the source documents (low recall).

When searching for prior art, patent examiners are currently mainly relying on keyword searches and Boolean logic. However, the consensus in the research literature in the information retrieval and patent domains is that a keyword-based search for prior art, even if done with most professional care, often produces suboptimal results (Helmers et al. 2019). This is particularly important considering the different consequences of false positive and false negative results in the patent domain. While false positives cause additional work for the patent examiner, who has to exclude the irrelevant documents from the report, false negatives may lead to an erroneous grant of a patent, which can have significant legal and financial implications (Trippe et al., 2017).

Several recent studies advocate the development of user-centred information retrieval systems, which assist expert patent examiners in identifying relevant literature and making decisions in prior art. Such systems offer improved interactivity and transparency, which are critical in gaining the trust of the users. For example, a system called Sigma, currently piloted at the United States Patent and Trademark Office (USPTO) (Krishna et al., 2016), not only performs basic keyword searches but also allows the experts to create search strategies that are best suited to examining a particular application. Another study explored the use of word embeddings (Showkatramani et al., 2018) and concluded that no model by itself was sophisticated enough to match an expert's choice of keyword expansion.

1.4 User-centred approach and scope

This study follows a user-centred design by considering the needs, wants and limitations of users (including both applicants and examiners) throughout the process. As a result, this human-in-the-loop approach aims to maximise performance by combining AI and human intervention and is designed to supplement, not substitute, human expertise and judgment. The AI algorithms are used to support the user in navigating through large volumes of patent data by suggesting the most plausible search terms and categorisations of patents into easily interpretable topics. In this scenario, the user keeps the role of the key decision maker, whereas the AI provides intelligent decision support.

The scope of the feasibility study is a wide range of state-of-the-art supervised and unsupervised machine learning approaches that will support the tasks of feature extraction, query expansion, document classification, document clustering and topic modelling. The intention is not to design a fully functional information retrieval system but to develop a proof-of-concept that will enable experimental comparisons between different approaches.

2 Observations and interviews

The interviews with IPO patent examiners specialised in different sectors were held in January and February 2019. They were conducted by the academic researchers from Cardiff University.

2.1 Prior art searching as a process

The purpose of the prior art search is to find the closest prior art that may impact the patentability of an application and the likelihood of getting a patent granted. In its simplistic form, the prior art search involves the following steps:

- examining the *claims* and identifying terms/possible keywords;
- distilling what the defining part of the invention is and forming a *search statement*;
- identifying the most relevant *classifications* based on keywords and examiner's background knowledge;
- *optional background search* to identify the most suitable terms and synonyms;
- forming *search queries,* primarily using EpoqueNet[1], using keywords, classification codes and Boolean functions;
- finding the patents that are *likeliest* to be relevant to the application;
- *sifting through* the retrieved documents in EpoqueNet, using colour coded highlights, drawers and sticky notes, to identify the most relevant patents;
- further *narrowing down* the search results, often using the drawings and manual disambiguation of concepts, to identify close conceptual similarities;
- *optional search* for published research/online materials;
- forming a *conclusion* (judgement) about the novelty and inventiveness of the application.

The definition of the search statement is one of the most important steps in the process. It requires clear understanding of the critical subject matter and the potential novelty of the application.

[1] EpoqueNet is a professional patent search tool for national patent offices that is produced by the European Patent Office (EPO)

Examiners often modify the search statement several times as their understanding of the prior art or the potential patentability of the application develops. The search statement may include words, which do not necessarily appear in the original claims.

The most time-consuming step is sifting through the large number of patents retrieved.

Searching strategy: very systematic due to the structured way patent literature is organised.

Search techniques currently used: keywords, classifiers, Boolean logic, proximity operators, truncation operators (*e.g.* right word truncation), linking to full-text documents and patent families, linking to external and internal depositories, keyword and synonym selection, combining saved search queries appropriately, iterative modification of previously stored search queries in light of newly acquired phrases and terminology, citation search and multilingualism.

Post-search analysis techniques currently used: colour coding/highlighting, drawers and sticky notes in EpoqueNet.

2.2 User requirements

- *Key user requirements:* retrieving the closest documents, ranking relevant documents, suggesting synonyms, suggesting classifications, suggesting highlights, visualising the retrieved documents/concepts and clustering.
- *Additional (desirable) requirements* beyond the scope of this feasibility study*:* retrieving image-based patents, collecting evidence of due diligence, spotting ambiguity, finding contradictions, sense disambiguation, visualising relationships among documents and searching pictures/drawings.
- *Scope:* searching and filtering patents from a number of sectors.

The main user requirement is effective prior art searching and filtering of patent literature (*i.e.* granted patents and published patent applications).

2.3 Key challenges

Patent searching is a highly interactive and complex process often requiring multiple searches, diverse search strategies and search management. The key linguistic and semantic challenges are legal wording, long sentences, acronyms, and the technical nature of patent claims.

The usability of an Information Retrieval (IR) system is a function of three aspects: its effectiveness, efficiency and user satisfaction. This feasibility study mainly focuses on effectiveness—the ability of the system to provide documents according to specified relevance criteria. The gold standard is manually judged results. However, research has shown that human judges tend to vary in what they find relevant. Users agree more with each other when asked questions in the form “Which of these two documents is more relevant to the query?” than when asked to provide absolute judgements (*e.g.* “Is this document relevant to the query?”).

Almost all contemporary search technologies are based on ranked retrieval, and it is accepted by the Information Retrieval (IR) community that ranked retrieval is almost always more effective than Boolean retrieval.

2.4 Technical requirements

The technical requirements (TRs) for this feasibility study were:

- TR1: Automated *query expansion* by suggesting synonyms, meronyms, hyponyms and hypernyms;
- TR2: Automated *document classification* by suggesting additional classification codes;
- TR3: Automated *identification of similar documents* using semantic similarity measures;
- TR4: Automated *ranked list* of relevant documents based on document similarity;
- TR5: Visualisation of the distinguishing characteristics of retrieved documents using *topic modelling*.

2.5 Indicators and measures

Precision P and *recall R* are often used in IR as measures of effectiveness. Precision indicates how many irrelevant documents were retrieved together with the relevant ones, while recall measures how many relevant documents were overlooked. Precision is often seen as a measure of exactness or quality, whereas recall is a measure of completeness or quantity.

Precision = number of relevant items retrieved / number of items retrieved

Recall = number of relevant items retrieved / number of relevant items in the whole collection

Both measures require manual labelling of documents and assessment of their relevance by experts. It is impractical to assess all documents in a large collection, in which case only precision is used. In addition, total recall is not always required in a prior art search as it is only necessary to find one reference which predates the filing of the patent application. In practice, most searchers aim to find more references, but there is no requirement for a total recall.

Since the ranking of documents is one of the most important criteria, this feasibility study will measure *precision at k point* (*precision@k*), where *k* is a cut-off point in the ranked list of retrieved documents. The parameter *k* is not fixed, and a range of potential values is considered, *e.g. k* = 10, 20, ..., 100.

This study also uses *F-measure* to assess the accuracy of the classification algorithm. It is defined as the weighted harmonic mean of the precision and recall. Other measures used to assess the human aspects include agreement to measure the interpretability of topic modelling and a focus group discussion to explore user experience.

3 Proof-of-concept

3.1 AI techniques considered

This feasibility study investigates a combination of technologies including natural language processing (NLP), machine learning (ML) and semantic technologies. Different AI algorithms will be considered in terms of their suitability to address the main technical requirements (TR1-TR5).

Table 1: AI and NLP algorithms considered

AI and NLP Algorithms	TR1: Query expansion	TR2: Document classification	TR3: Document similarity	TR4: Ranking	TR5: Topic modelling
Natural language processing: text segmentation, normalisation, lemmatisation, stemming, co-occurrences, multi-word terms	X	X	X	X	X
Unsupervised machine learning: word embeddings, distributional semantics	X	X	X		X
Supervised machine learning: support vector machine, naive Bayesian learning, decision tree induction, random forest		X			
Unsupervised machine learning: neural networks, deep learning		X			
Similarity measures: Jaccard similarity, Euclidean distance, cosine similarity			X		
Semantic technologies: use of lexico-semantic knowledge, latent Dirichlet allocation (LDA)	X		X	X	X

3.2 Concept model

Figure 1 shows a conceptual diagram of the main processes involved in a prior art search and the filtering of patent information. The concept model was developed as a methodological tool for systematic experimentation with different algorithms. The proposed model is based on the following assumptions:

- the examiner reads an application and defines a search statement and a search query;
- the system classifies the application into one or more classes;
- the system extracts the most relevant keywords (including multi-word terms) from the application;
- the system suggests expanding the query with other related words;
- the examiner curates the search query;
- the system launches a search to retrieve documents from the relevant classes;
- the system assorts the retrieved documents into topics, each described by a set of keywords;
- the examiner selects the topic(s) deemed most relevant to the application;
- documents from the relevant topic(s) are ranked based on their similarity to the application;
- the content of each document is colour-coded to highlight its relevance to the application.

Figure 1: Concept model of a prior art search and the filtering of patent information

To support practical experiments, a system based on the proposed concept model has been implemented in the programming language Python. Dependencies on external software libraries are described in Appendix 1.

For the purposes of this feasibility study, three domains were chosen to validate the system experimentally throughout its development: civil engineering, computer technology and transport. These three domains were chosen because they are the top three technology fields[2] based on number of filings at IPO over the past 10 years. Each domain was formally defined as the union of relevant inventions areas identified by their codes in the International Patent Classification (IPC) scheme (World Intellectual Property Organisation, 2019). The chosen IPC codes are listed in Appendix 2. The corresponding validation datasets were created by retrieving patents with these IPC classes/subclasses from sources identified by IPO. The original data were formatted in XML according to a schema provided in Appendix 3. The data were stored in an XML database for easy querying by metadata.

3.3 Description of system functionalities

3.3.1 Feature extraction

3.3.1.1 Single-word features

The purpose of this task is to automate the extraction of lexico-semantic features that will later be utilised by methods described in Sections 3.3.2-3.3.4. Our document representation is based on the bag-of-words (BoW) model, where each document is represented as the bag of its words. Although this simple representation completely ignores the grammar and word order, it has proven successful in applications such as information retrieval and document classification mainly due to the multiplicity of words, which allows their local relevance to be easily quantifiable using measures such as term frequency-inverse

[2] Of the 35 WIPO technology fields – see IPC concordance table at https://www.wipo.int/ipstats/en/

document frequency (TF-IDF) (Spärck Jones, 1972; Salton and McGill, 1986). The success of using individual words as key features also depends on the ability of the system to unify different surface forms. Basic linguistic pre-processing was used to neutralise insignificant orthographic differences between otherwise identical words such as letter casing (*Bayesian learning* vs. *bayesian learning*), non-ASCII characters (*e.g.* naïve Bayes vs. naive Bayes), spelling variations (*nearest neighbour* vs. *nearest neighbor*), spelling mistakes, etc. Further normalisation involves lemmatisation and stemming to support features that focus more closely on the underlying meaning of the words (*e.g.* transportation, transported and transporter are all mapped to transport as their common root).

3.3.1.2 Multi-word features

To model relationships between individual words, additional features based on word co-occurrence were considered. Two approaches were used here: one using a fixed-sized text window called n-grams and the other focusing on domain-specific multi-word terms. By definition, n-grams preserve the local context of individual words. N-grams can simply be added to a BoW to enrich document representation with contextual features. This allows for a finer-grained comparison of the respective documents.

N-grams divide text physically into blocks without any regard for the logical relations between words, either syntactic or semantic. Consequently, this may lead to the loss of important conceptual information. Consider for instance these two documents: *'... the way of doing things on the Internet has evolved...'* and *'...five ways the Internet of Things is transforming businesses...'*. Their BoW representations are similar as they both mention the words *Internet* and *things*. However, only the latter makes reference to the *Internet of Things* as a standard term used to refer to the interconnection via the Internet of computing devices embedded in everyday objects, enabling them to send and receive data. Bi-grams will fail to capture this information. Tri-grams will manage to represent the *Internet of Things* as a single feature in the extended BoW model. However, longer terms such as *Internet Small Computer System Interface* will again fail to be featured. Therefore, a more flexible approach is required to systematically capture important phrases regardless of their length.

Multi-word terms are commonly used as linguistic representations of domain-specific concepts, *e.g.* Internet of Things, Internet Small Computer System Interface, etc. These logical units of text that convey scientific and technical information tend to get lost when text is physically divided into n-grams. Locally developed software FlexiTerm was used to extract multi-word terms from text on the fly (Spasić et al., 2013; Spasić et al., 2018; Spasić, 2018). Additional advantage of using this particular software is its ability to link acronyms to their multi-word term representatives, *e.g.* Internet of Things (IoT), Internet Small Computer System Interface (ISCSI), etc. Unpacking acronyms to their full forms allows for their content (*i.e.* individual words) to become searchable and used as features for further document analysis. Appendix 4 provides examples of multi-word terms extracted from the validation data. Note that different term variants are grouped together. For instance, example ID 14 from Table 23 in Appendix 4 represents a simple case of orthographic variation (*e.g.* bottom hole assembly vs. bottomhole assembly) and links both variants to their acronym *BHA*. This grouping allows for all variants to be represented using a single feature. Example 39 from Table 24 in Appendix 4 shows syntactic variation, where the order of words varies (*e.g.* network functions virtualization vs. virtual network function), resulting in two acronyms, NFV and VNF. Note that the words with the same root, *virtual* and *virtualization*, are matched by way of stemming, which facilitates an interpretation of words based on their core meaning. To facilitate terminology browsing, terms can be automatically arranged into dendrograms based on their types (see Appendix 5).

3.3.1.3 Word embeddings

The above approaches represent words (or terms) as discrete variables, which cannot be easily compared with respect to their similarity and other semantic relationships. Therefore, a representation to compare the meaning of words was required. The semantics of words can be partly inferred from text based on their contextual usage. This bottom-up approach is known as distributional semantics. Its main idea is summarised by the distributional hypothesis, which states that words with similar distributions have similar meanings. Word embeddings represent words in the form of real-valued vectors of low dimensionality, which are learnt from text using approaches such as neural networks or dimensionality reduction. By

capturing (or generalising) the context of a word, word embeddings tend to preserve similarity and other relationships between words by way of distances and directions in the corresponding vector space. Word embeddings also effectively bypass the curse of dimensionality, which is known to reduce the performance of machine learning algorithms (Hughes, 1968).

The study uses state-of-the-art word embedding algorithms – word2vec – (Mikolov et al., 2013) to train word embeddings on each domain separately and obtain domain-specific word representations. Consider, for example, domain-specific uses of the word *driver* given in Appendix 6. This word typically refers to a physical object, software or a person in civil engineering, computer technology or transport respectively. Different meanings of the word have been captured by domain-specific word embeddings through their relationships to similar or otherwise related words (see Appendix 7). For example, the word *driver* in transport is close to its domain-specific synonyms (*e.g.* vehicle-operator), hyponyms (*e.g.* cyclist) and related words (*e.g.* passenger), whereas, in computer technology, it is close to its domain-specific synonyms (*e.g.* controller) and related words (*e.g.* I/O).

3.3.2 Query expansion

A prior art search involves investigating whether a similar idea has already been described in a previously published patent. A thorough prior art search involves creating a search query involving different combinations of relevant search terms. The purpose of this task is to facilitate reformulation of a search query to improve retrieval performance by adding search terms that can identify additional relevant documents. Given an initial list of search terms, the goal of query expansion is to improve retrieval performance by also searching for their lexically related terms (*i.e.* synonyms, hyponyms, hypernyms and meronyms), semantically related terms and surface variants (Azad and Deepak, 2019). For example, given the original query (*e.g.* automobile), it can be expanded by including synonyms (*e.g.* car), meronyms (*e.g.* engine), hyponyms (*e.g.* minivan) and hypernyms (*e.g.* vehicle).

3.3.2.1 Lexical relationships

A classic approach to query expansion involves the use of a thesaurus, which organises words according to the aforementioned relationships of synonymy, meronymy and hyponymy. WordNet is the largest lexical database of English, in which content words (*i.e.* nouns, verbs, adjectives and adverbs) are grouped into synsets (*i.e.* sets of synonyms), each corresponding to a distinct concept in the semantic space (Miller, 1995; Fellbaum, 1998). Synsets are further interlinked by means of lexical relations including hyponymy (*i.e.* 'is a kind of') and meronymy (*i.e.* 'is a part of'). Given a word, these relations can be explored to find related words that can then be presented to the user as plausible candidates for query expansion. Appendix 8 illustrates this concept using WordNet's web interface. To access WordNet programmatically from our system and its own interface, the NLTK WordNet API was used.

3.3.2.2 Semantic relationships

While easy and straightforward to use, WordNet has been designed as a general resource, therefore, its coverage may vary across different domains. For instance, it recognises a *tablet* as a dose of medicine in the form of a small pellet but not as a mobile device. When trained on a domain-specific corpus, word embeddings can capture domain-specific meaning, as illustrated in Appendix 7 using the word *driver* as an example. Given a word, the vector space of word embeddings can be explored using simple arithmetic operations to retrieve related words as its nearest neighbours. They can then be presented to the user as plausible candidates for query expansion. Note that the related words might as well include lexically related words. For example, the neighbourhood of the word *driver* in transport (see Figure 14 in Appendix 7) includes its synonyms (*e.g.* vehicle-operator) and hyponyms (*e.g.* cyclist and motorman). However, when using word embeddings to retrieve related words, it cannot currently differentiate between specific relationships. However, distance can be used to measure the 'strength' of individual relationships and varying this parameter can control the number of alternative search terms suggested to the user.

Using the two approaches described above, a search query can be iteratively tuned to develop an optimal search strategy. The role of the user in this process shifts from the 'art and craft' of recalling search terms from memory to curating those automatically suggested by the system. This would not only improve the efficiency of query formulation but would also improve the consistency across users, thereby supporting the reproducibility of search results. To give a user more control over the query expansion process, an interface has been created to allow them to (de)select additional search terms suggested automatically by the system.

3.3.2.3 Local search engine

Given a query, the actual search is performed using Elasticsearch (Elasticsearch, 2018), the most popular search and analytics engine. For added search flexibility and robustness, the query can be further expanded using built-in suggesters: term, phrase and completion suggesters. The term suggester provides word alternatives on a per-token basis within a certain edit distance, which can be used to account for spelling mistakes, spelling variations and other types of surface variations. The phrase suggester adds additional logic on top of the term suggester to provide entire corrected phrases instead of individual tokens based on an n-gram model. This suggester can help a patent examiner make better decisions about which search terms to select based on word distribution. Both suggesters support did-

you-mean functionality. Finally, the completion suggester provides auto-complete or search-as-you-type functionality. The completion suggester uses data structures that enable fast lookups to provide instant feedback to the user as they type. This navigational feature can be used to guide a patent examiner to relevant documents as they are typing, thus improving search precision. In addition, Elasticsearch can be boosted with plug-ins, *e.g.* the International Components for Unicode (ICU) plugin for better analysis of Asian languages, Unicode normalisation, Unicode-aware case folding, collation support and transliteration.

By default, Elasticsearch is used via REST API, but Python binding can be used to fully integrate Elasticsearch into the proposed system. For the purpose of this feasibility study, a local Elasticsearch server was installed and stored the validation data (see Appendix 9) independently of other system components.

3.3.3 Document classification

The International Patent Classification (IPC) is a hierarchical system of approximately 650 subclasses used to classify patents in a uniform manner (Makarov, 2004). Each patent is assigned at least one classification code, which indicates the main subject to which the invention defined in the patent application relates. Additional codes may be appended to further refine the classification of the patent. For a given patent application, a patent examiner assigns the classification code manually following the classification guidelines. The fact that filed patents are already classified provides a perfect opportunity to explore supervised machine learning algorithms to automate the task of classifying patent applications. Supervised learning uses a large set of training data, where each document is assigned a class label, to generalise the relationships between different features and classes into a classification model (*e.g.* function, decision tree, probability, etc.). Given a new document, the classification model is applied to predict its class label. In our scenario, the new document represents a patent application, which will be indirectly compared to the filed patents whose generalisable properties will be captured by the classification model.

Having identified a training set, the next step involves the choice of a specific supervised learning algorithm. According to the no-free-lunch theorem, any two learning algorithms are equivalent when their performance is averaged across all possible problems (Wolpert, 1996). In other words, there is no universally best learning algorithm, which suggests that the choice of an appropriate algorithm should be based on its performance for the particular problem at hand and the properties of data that characterise the problem. Cross-validation experiments can be used to estimate the performance of machine learning algorithms on unseen data in a less biased/optimistic manner. This is important, as more-complex and data-hungry algorithms such as deep learning may overfit the training data. To that end, 10-fold cross-validation experiments were used to systematically evaluate the performance of a wide range of supervised learning algorithms, including:

- support vector machines (SVMs) with radial basis function (RBF) kernel;
- decision tree induction;
- random forest;
- AdaBoost;
- nearest neighbours;
- multilayer perceptron (MLP);
- Gaussian naïve Bayesian (NB) learning;
- Bernoulli NB learning.

In the proposed system, a binary classifier would be trained for each IPC subclass, using its patents as positive examples and those from all other subclasses as negative examples. To sufficiently challenge a classifier during cross-validation, two relatively similar IPC subclasses from each validation domain were

selected (see Table 2). As the size of IPC subclasses can vary considerably, subclasses of different sizes were chosen to measure the extent to which the size of the training dataset affects the classification performance. Two data representation models were used based on BoW and word embeddings, respectively. Appendix 10 provides a summary of cross-validation results.

Table 2: Classes used in the cross-validation experiments

Domain	Subclass 1	Subclass 2	Subclass size
Civil engineering	E03D (water-closets or urinals with flushing devices; flushing valves therefor)	E03F (sewers; cesspools)	900 (small)
Computer technology	G06K (recognition of data; presentation of data; record carriers; handling record carriers)	G06T (image data processing or generation, in general)	20K (large)
Transport	B62J (cycle saddles or seats; accessories peculiar to cycles and not otherwise provided for, *e.g.* article carriers or cycle protectors)	B62K (cycles; cycle frames; cycle steering devices; rider-operated terminal controls specially adapted for cycles; cycle axle suspensions; cycle sidecars, forecars, or the like)	3K (medium)

3.3.4 Topic modelling

There are two primary paradigms of navigation through large volumes of text data—searching and browsing—which fulfil different purposes. Users who browse are looking to discover new information, whereas users who search are looking to find specific information. Therefore, browsing can support opportunistic exploration of prior art when search terms cannot be easily defined. Browsing requires categorisation of documents into major topics. One such categorisation is IPC, which was mentioned in Section 3.3.3, but each IPC category is broad, and hence its manual inspection is not feasible. To support fine-grained browsing within IPC subclasses, latent Dirichlet allocation (LDA) can be used to discover abstract topics within a collection of documents (Blei et al., 2003). Each topic is characterised by a number of keywords that best discriminate it against other topics. These keywords support the interpretability of topics, therefore allowing the user to quickly assess the relevance of documents associated with that topic. In addition, each document can be associated with multiple topics, which is useful for simultaneously exploring multiple aspects of a patented invention. Different parameters of LDA, such as the number of topics, keywords, iterations, minimum probability, etc., will have different implication on the utility of results and, therefore, require systematic experimentation to find optimal settings for each IPC category. To tune these parameters, a series of topic modelling experiments were performed using the validation data.

As an unsupervised approach, topic modelling is notoriously difficult to evaluate. Topic coherence measures have been used to remedy the problem that topic models give no guarantee on their interpretability (Röder et al., 2015). While topic coherence was measured, a method of measuring interpretability was also proposed, as it is of utmost importance in the context of triaging filed patents.

3.3.5 Document similarity

This task builds upon a traditional information retrieval approach to prior art searching. This approach relies upon a user to map the invention idea onto a set of appropriate search terms. Assuming that the search terms are known, the actual retrieval from the database can be performed efficiently. Given that the invention idea is already described in a patent application, it can be compared directly against filed patents to retrieve the most similar patents. This is traditionally done using a vector space model, in which each document is represented by a vector whose coordinates correspond to individual words weighed by the frequency of their distribution within the document and across all documents, which is known as term frequency-inverse document frequency (TF-IDF) (Spärck Jones, 1972). Two documents can then be compared by measuring the distance (or similarity) between their vectors. For this purpose, experiments with Jaccard similarity and Euclidean distance were conducted, but the best results were achieved, as expected, using cosine similarity because it represents a measurement of orientation and not magnitude (Jurafsky & Martin, 2008).

All of the above approaches can be applied at different levels, ranging from a whole document to individual sections, paragraphs and sentences. Such granularity is of particular importance, as most inventions represent improvements upon existing solutions. Therefore, it is important to identify paragraphs or sentences that refer to ideas already described in other patents. For shorter text snippets such as titles and sentences, the vector space model (even with dimensionality reduction) will result in sparse feature vectors, which would exhibit weak discrimination in the face of high dimensionality (Houle et al., 2010). Alternatively, word embeddings can be reused to encode or compare the meaning of individual sentences. For example, a sentence can be represented by the centroid of its word embeddings and thereby measure the distance between two sentences. A more fine-grained measure of distance between two sentences would be the word mover's distance (WMD) (Kusner et al., 2015), which represents the minimal cumulative distance that the words of one sentence need to travel in the word embeddings space to reach the words of the other sentence. For example, after removing the stop words, the distance between ‘extendible umbrella handle’ and ‘parasol foot with retractable point’ would be the sum of the distances between the closest pairs of words, *i.e.* umbrella and parasol, extendible and retractable and handle and foot.

4 Evaluation

4.1 Experimental protocol

Table 3 outlines multiple system functionalities that were evaluated and the mode of their assessment. Using a dataset of 162,154 published patent applications, the system was evaluated using the experimental protocol outlined in Table 4; specific AI algorithms used to support different functionalities of the system are listed in the right-most column. The associated technical requirements (TRs) are indicated in the first column. The corresponding evaluation experiments and their outcomes are provided in the subsequent sections.

Table 3: Evaluation experiments

Functionality	**Aspect**	**Assessment**
Classification	Accuracy	F-measure
Topic modelling	Interpretability	Agreement
Information retrieval	Accuracy	Precision@k
Usability	User experience	Focus group

Table 4: Experimental protocol

TR	Step	Action	Rationale	Algorithm
TR2	1	The system classifies the application into one of three domains: 1. civil engineering 2. computer technology 3. transport	Constraining the search to a specific domain reduces the number of false positives when homonyms (*i.e.* words that are spelled the same way but have different meanings) are used as search terms. For example, the word bus means 'a large motor vehicle carrying passengers by road' in transport and 'a distinct set of conductors carrying data and control signals' in computing.	Linear support vector machine (SVM) classifier with stochastic gradient descent (SGD) training
TR5	2	The system maps the application to the most relevant topics within the domain, each described by a set of keywords.	Constraining the search to a specific topic reduces the number of false positives as ambiguity persists within a domain. For example, the word 'code' in computing can be used in multiple contexts, *e.g.* software, access control, digital encoding, etc.	Latent Dirichlet allocation (LDA)
	3	The system extracts the most relevant keywords from the application.	Focusing the search to the most relevant keywords supports identification of related patents. More importantly, it reduces the user's total cognitive load, here defined as the amount of mental processing needed to define a search query, to maximise usability of the system.	Term frequency-inverse document frequency (TF-IDF)
TR1	4	The system suggests expanding the query with other related words, which were identified using: 1. general purpose thesaurus 2. domain-specific word embeddings 3. topic modelling	Expanding the search query with other related words increases the recall, *i.e.* identifies a larger set of relevant patents.	1. WordNet 2. word2vec 3. LDA
	5	The user curates the search query.	Manual curation of the query is expected to improve both the recall and the precision of the search results.	*n/a*
TR3, TR4	6	The system launches a search to retrieve and rank at most 30 patents from the relevant domain and topics within.	The retrieved patents are expected to be ranked by their relevance to the application, thereby making their identification more efficient.	Elasticsearch
	7	The retrieved patents are mixed with a set of 30 patents selected randomly from the same domain and then shuffled.	This step was added to the system to reduce the bias in evaluation. In a blinded experiment, information that may influence the participants is masked (or blinded) until after the experiment is complete.	*n/a*
	8	The system cross-references the query against each patent to colour-code its content.	By highlighting parts of the patent that match the search query, the user can assess its relevance to the application faster.	*n/a*
	9	The user assesses the relevance of each patent on a 3-point Likert scale: 1. irrelevant 2. somewhat relevant 3. relevant	Annotating the relevance of each patent creates a gold standard against which the overall search performance can be evaluated.	*n/a*

The AI-assisted prior art searching algorithms were trained on data provided by IPO with publication dates on or before 31 December 2018. Data provided includes the PATSTAT bibliographic database of worldwide patents (Autumn 2018 edition), GB full-text patents (1979-2018), EP full-text patents (1978-2018) and US full-text patents (1976-2018). For data security reasons, IPO was unable to supply the accompanying patent examiner search statements for each training document as these are not published.

The evaluation includes quantitative and qualitative experiments as outlined below.

4.1.1 Quantitative testing

The IPO testing on the algorithms was undertaken in November 2019 on patents published since 1 January 2019 in each of the three test sectors (civil engineering, computer technology and transport) using ten 'query' patents, which reflect a range of different technological complexities in each of the test domains. Results (up to 60 documents for each 'query' patent—split 30/30 from the Cardiff University 'long list' of results to deliberately provide some 'control' results) were sent to IPO for assessment.

4.1.2 Qualitative testing

Two patent examiners from each of the three test domains assisted with the evaluation process. For each of the 10 'query' patents from their domain, each examiner was presented with an EpoqueNet working list pre-populated with up to 60 documents, with the 30/30 split put in a random order (and a different random order for each examiner). Once the examiner had grasped the subject matter of the 'query' patent in question, they went through the result documents in the EpoqueNet working list and added documents to the first drawer that were of any potential relevance to the subject matter of the 'query' patent (*i.e.* the only documents not added to the first drawer were those that are completely irrelevant). Examiners then went through the first drawer and added documents to the second drawer if they were considered to be worth more detailed consideration, in the same way that examiners consider the results of a normal prior art search. A supervisor from the IPO project board was in the room to provide a quick overview of the testing process and to answer any questions throughout the day. Each examiner was expected to complete the evaluation process of the 10 'query' patents in their domain within one day.

4.1.3 Focus group

The qualitative testing was followed by a focus group discussion on usability aspects. The meeting was attended by all patent examiners who have taken part in the evaluation testing and one of the Primary Investigators from Cardiff University.

4.2 Results and discussion

4.2.1 Classification

Cross-validation experiments were used to assess the performance of machine learning algorithms on the training data. The best performing algorithm was chosen to be built into the system's classification module. The model was re-trained on all available training data and finally evaluated with holdout testing using the entire test dataset. The classification performance is summarised in the confusion matrix shown in Table 5. These values were used to evaluate the classification performance in terms of precision, recall and F-measure (see Table 6).

Table 5: Confusion matrix

		Predicted		
		Civil engineering	Computer technology	Transport
Actual	Civil engineering	8,115	0	0
	Computer technology	0	12,422	0
	Transport	0	0	12,560

Table 6: Classification performance

	Precision	**Recall**	**F-measure**	**Support**
Civil engineering	100%	100%	100%	8,115
Computer technology	100%	100%	100%	12,422
Transport	100%	100%	100%	12,560
Micro-average	*100%*	*100%*	*100%*	*33,097*
Macro-average	*100%*	*100%*	*100%*	*33,097*

4.2.2 Topic modelling

The IPC classification system is designed to facilitate prior art searches by organising patents into indexed, manageable structures for easy retrieval. The role of classification described in the previous section is to assign a new application to an appropriate IPC code. Nonetheless, the number of patents across IPC codes varies significantly, with some being very broad and heterogeneous in nature. Topic modelling is a method to organise, understand and summarise large collections of textual information. Within the system, the role of topic modelling is to assort patents within each code into homogeneous clusters. Each cluster corresponds to a topic, which is described by a set of keywords that differentiates it from other topics. The system automatically maps a new application to its most likely topics. However, the user is given an opportunity to validate the proposed mappings or override them, with the immediate goal of enabling a more focused search with fewer false positive for the user to sift through. The secondary goal of such user intervention is to provide feedback to the system so that it can learn to auto-correct itself through its usage. For a user to make an informed decision about the validity of topics, they need to be easily interpretable. Interpretability can also help improve the user's trust in AI as well as diagnose the underlying

issues in the machine learning model and/or training data. However, interpretability is a cognitive concept that is not immediately quantifiable.

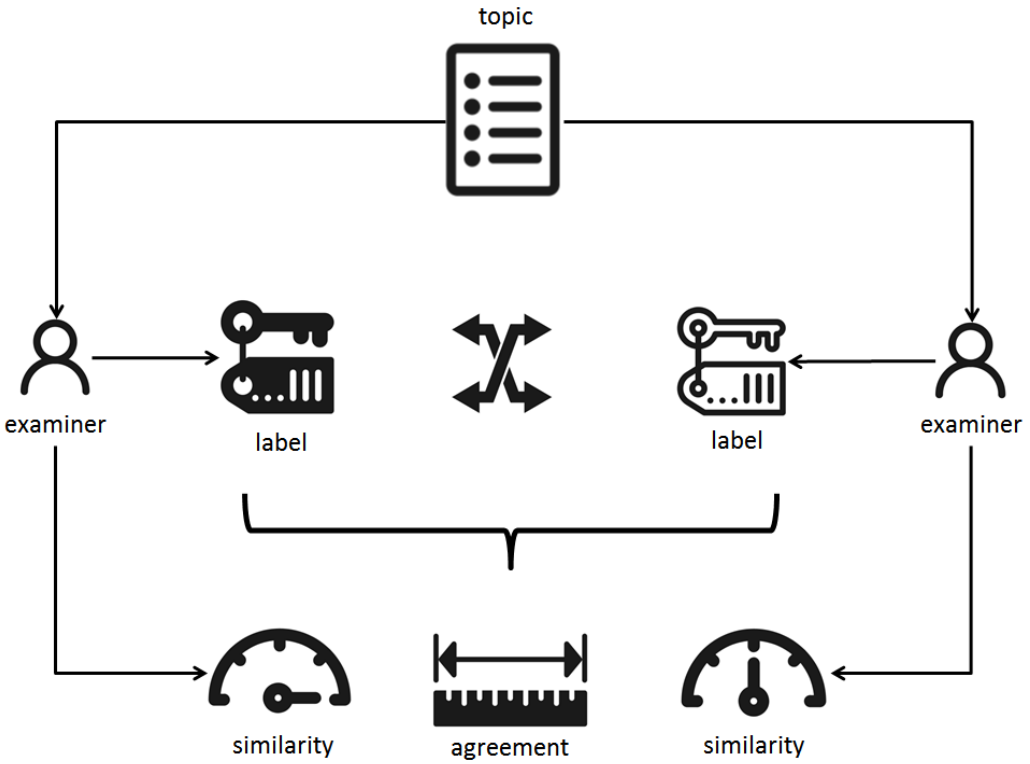


Figure 2: Experimental protocol for measuring topic interpretability

To measure interpretability of topics, experiments were designed using a protocol illustrated in Figure 2. In this scenario, two patent examiners were paired. A topic, described by its keywords, was presented to each examiner. Each examiner was asked to name the topic independently using a phrase that generalises the collective meaning of the keywords. No restrictions were imposed onto the choice of vocabulary or phrase format used by the examiners, but they were allowed to refer to official terminology if they believed it could help them identify a suitable phrase. Similarly, they were also allowed to search the Internet using the topic's keywords. The examiners were asked to estimate the confidence in their final choice on a 5-point Likert scale (see Table 7).

Table 7: Confidence Likert scale

Scale	Description
0	Not confident at all
1	Slightly confident
2	Somewhat confident
3	Moderately confident
4	Very confident

Table 8: Similarity Likert scale

Scale	Description
-3	Very dissimilar
-2	Moderately dissimilar
-1	Slightly dissimilar
1	Slightly similar
2	Moderately similar

In the second phase, both examiners gained access to the other examiner's choice of a topic's name. They were then asked to independently estimate the similarity of the two names on a 6-point Likert scale (see Table 8). The average similarity was used to estimate the interpretability of topics under the hypothesis that high similarity implies high interpretability and vice versa. The experimental data were collected for 10 topics in each domain, each described by a total of 15 keywords (see Table 9 to Table 11). The average confidence was found to be 2.95, 3.00 and 3.10 in civil engineering, computer technology and transport respectively. Therefore, the confidence was consistently found to be moderate (see Table 7 for interpretation of the corresponding Likert scale). The average similarity was found to be 1.40 (slightly similar), 2.35 (moderately similar) and 2.65 (very similar) in civil engineering, computer technology and transport respectively (see Table 8 for interpretation of the corresponding Likert scale). In addition, inter-annotator agreement for both confidence and similarity were calculated to check whether the examiners were consistently finding some topics more difficult to interpret than others. For this purpose, weighted Cohen's kappa coefficient (Cohen, 1960; Cohen, 1968; Fleiss et al., 1969; Fleiss and Cohen, 1973) was used.

The results are given in Table 12 to Table 15. Although the confidence was found to be moderately high overall, it varied significantly across the topics in computer technology (see Table 12 and Table 13). On the other hand, the judgement of similarity was found to be very consistent across all domains (see Table 14 and Table 15), albeit the similarity was found to be low in civil engineering. The high similarity and high agreement obtained for transport illustrate the potential of using topic modelling to support prior art searches. The preliminary results were obtained using a fixed number of topics and their keywords. Further experiments are needed to optimise the parameters of topic modelling for individual domains, as these can vary considerably in terms of their breadth and depth, as illustrated by the preliminary topic modelling results.

Table 9: Topic interpretability experiment results - civil engineering

ID	Keywords	Name	Confidence	Similarity
1	fluid drilling wellbore tool string valve downhole flow gas tubular oil injection sealing bore annular	well boring	very confident	very similar
		oil drilling, particularly bore linings and maintenance	moderately confident	very similar
2	sensor detection data power light unit signal electric information transmitted vehicle display electronic communication receiving	real time traffic signs	somewhat confident	slightly similar
		automated vehicles infrastructure	somewhat confident	slightly similar
3	tower platform post barry ladder vehicle anchor concrete rail road frame track ground member cable	elevator	slightly confident	very dissimilar
		construction of transport infrastructure	slightly confident	very dissimilar
4	water drain air flow toilet valve outlet pipe pool cleaning tank inlet filter flush waste	flushing mechanisms	moderately confident	very similar
		domestic plumbing, toilets in particular	very confident	slightly similar
5	layer inside composition heat sheet panel polymer fiber coating glass resin fibre adhered water particular	insulation for buildings	moderately confident	very similar
		manufacturer of building materials for construction, insulated building panels in particular	moderately confident	very similar
6	panel flow member profile plate edge roof frame building beam tile cover concrete sheet reinforcing	roof drainage or guttering	somewhat confident	moderately similar
		materials for roof structures	moderately confident	slightly similar
7	window rail frame roller sash guide screen member cord slats blind profile sliding panel door	windows for buildings	very confident	very similar
		window/doors and coverings thereof	very confident	very similar
8	drilling bit cutting tubular pipe pile tool member sealing blade string body axis tubular ring	cutting device for well boring	very confident	very similar
		oil drilling, particularly design of the drilling equipment itself	moderately confident	very similar
9	hydraulic engine boom machine valve pump work motor drive cylinder bucket excavator arm speed vehicle	augur or land moving	moderately confident	slightly similar
		civil (not domestic) waste system construction *e.g.* sewers, treatment plants	somewhat confident	slightly similar
10	door lock hinge member latch pin body sliding handle plate lever spring pivot key arm	hinges for doors	very confident	slightly similar
		locks and locking mechanisms	very confident	moderately dissimilar

Table 10: Topic interpretability experiment results – computer technology

ID	Keywords	Name	Confidence	Similarity
1	search database web file document control item query page text network language model code test	information retrieval	slightly confident	very similar
		databases, data retrieval, parsing, code testing, virtual code deployment G06F16, G06F17/20, G06F11, G06F8	moderately confident	moderately similar
2	instruction cache node virtual network host address request bus write logic resource disk machine task	memory addressing/allocation	somewhat confident	very similar
		virtual machine, hypervisor, resource allocation, scheduling, RAID, distributed storage, cloud storage system, virtual address space, memory interconnect G06F3/06, G06F12/08, G06F12/02, G06F9/50, G06F13/16, G06F15/16	moderately confident	moderately dissimilar
3	power circuit voltage cell clock bit line gate transistor supply write switch charge array semiconductor	power supply (PSU)	moderately confident	moderately similar
		power control circuit for system with battery, power save management, clock domains, semiconductor memory, G06F1/28, G06F1/32	moderately confident	moderately similar
4	print job sheet scanning document recording label driver copy page color CPU peripheral installed panel	printers/printing	very confident	very similar
		printers, printing job scheduling, printer control, printer drivers, G06F3/12, H04N1, G06F9	very confident	very similar
5	pixel camera region vehicle model color captured measurement light sensor calculated target analysis frame motion	image recognition for vehicle systems	moderately confident	very similar
		road/speed camera, on-vehicle camera, image processing, image analysis for vehicle recognition, sensor based image processing - not G06F	somewhat confident	moderately similar
6	audio speech encoding code frequency decoding sound noise frame band channel filter voice sample bit	speech processing	very confident	moderately similar
		transmission of speech data, noise filter, not G06F	somewhat confident	slightly similar
7	touch electrode panel sensor light layer surface conductive capacity substrate fingerprint transparent film emitting finger	touchscreens	very confident	very similar
		touchscreen, capacity based touchscreen, security G06F3, G06F21	very confident	very similar
8	tag RFID antenna card magnetic member surface housing layer electronic body circuit board sides contact	RFID tags (record carriers)	moderately confident	very similar
		barcodes readers, not G06F	moderately confident	moderately similar
9	authentication network second client control mobile terminal key message encryption wireless request file card software	user/client authentication	moderately confident	very similar
		security, mobile access control, authentication, G06F21	moderately confident	very similar
10	touch screen terminal mobile electronic sensor key broadcast gesture moving icon button wireless menu control	touchscreen user interfaces	somewhat confident	very similar
		gesture based input to touchscreen, G06F3	very confident	very similar

Table 11: Topic interpretability experiment results – transport

ID	Keywords	Name	Confidence	Similarity
1	frame rear seat bicycle member arm suspended left right pivot motorcycle rider axle cover link	two wheel vehicle suspension	moderately confident	moderately similar
		rider propelled vehicles, cycles	moderately confident	moderately similar
2	light image display data information sensor detection signal camera lamp communication mirror reflected process emitting	vehicle control and driver interaction	moderately confident	very similar
		vehicle control systems	slightly confident	very similar
3	power battery electric charging voltage supply circuit current switch converter cell storage coil inverter energy	electric vehicles	very confident	very similar
		electric vehicles	very confident	very similar
4	gear engine transmitted clutch shaft power speed torque electric output hybrid input machine shift combustion	hybrid vehicles	very confident	very similar
		hybrid vehicles	moderately confident	very similar
5	tire rubber tread layer composition groove bead polymer cord pneumatic group circumferential resin compound fiber	car tires	very confident	very similar
		tyres	very confident	very similar
6	aircraft wing blade vessel lift track trailer assembly platform actuator said propeller landing load flight	aircraft	very confident	very similar
		aircraft	moderately confident	very similar
7	steering brake detection sensor speed value torque acceleration angle assist determined signal calculated target estimated	vehicle stability control	very confident	moderately similar
		vehicle control	moderately confident	moderately similar
8	air valve heat pressure tank cooling gas fluid fuel chamber flow engine compressor inlet water	combustion engines	somewhat confident	moderately similar
		gas turbine engines	moderately confident	very similar
9	member lock shaft steering bearing ring assembly hub housing plate hole engine pin spring column	steering columns	moderately confident	slightly similar
		steering arrangements	slightly confident	very similar
10	seat panel member airbag door wall roof inflator material cover frame rail rear belt bag	vehicle seats and seatbelts	moderately confident	very similar
		vehicle seats	moderately confident	very similar

Table 12: Cohen's kappa coefficient with linear weighting on confidence

Domain	Observed kappa	Standard error	Confidence interval	Maximum possible	Proportion of maximum possible
Civil engineering	0.5283	0.2025	0.1314-0.9252	0.9057	0.5833
Computer technology	0.1111	0.2267	0.0000-0.5554	0.7778	0.1428
Transport	0.1667	0.1318	0.0000-0.4250	0.3750	0.4445

Table 13: Cohen's kappa coefficient with quadratic weighting on confidence

Domain	Observed kappa	Standard error	Confidence interval	Maximum possible	Proportion of maximum possible
Civil engineering	0.7368	0.1558	0.4315-1.0000	0.9474	0.7777
Computer technology	0.0141	*n/a*	*n/a*	0.7183	0.0196
Transport	0.3182	*n/a*	*n/a*	0.3182	1.0000

Table 14: Cohen's kappa coefficient with linear weighting on similarity

Domain	Observed kappa	Standard error	Confidence interval	Maximum possible	Proportion of maximum possible
Civil engineering	0.6970	0.1729	0.3581-1.0000	0.6970	1.0000
Computer technology	0.5352	0.2542	0.0371-1.0000	0.5352	1.0000
Transport	0.8024	0.1706	0.4680-1.0000	0.8024	1.0000

Table 15: Cohen's kappa coefficient with quadratic weighting on similarity

Domain	Observed kappa	Standard error	Confidence interval	Maximum possible	Proportion of maximum possible
Civil engineering	0.8172	0.0872	0.6462-0.9882	*n/a*	*n/a*
Computer technology	0.6475	0.2749	0.1087-1.0000	*n/a*	*n/a*
Transport	0.9231	*n/a*	*n/a*	0.9231	1.0000

4.2.3 Information retrieval

To evaluate the performance of information retrieval, the framework shown in Figure 3 was followed. The role of the system in this framework was to facilitate the formulation of the search query by a patent examiner, contextualise the query in terms of relevant domain and topic within and ultimately to retrieve the corresponding patents. To evaluate the performance of information retrieval, the search results were presented back to the examiner who then annotated their relevance on a 3-point Likert scale (Yes, Maybe, No) in line with the concept of the first and second drawer described in Section 4.1. The annotations were then used to calculate precision, which corresponds to the percentage of relevant documents among those retrieved by the system. The examiners were not shown the rank at this point, but this information was preserved nonetheless in order to calculate precision at k.

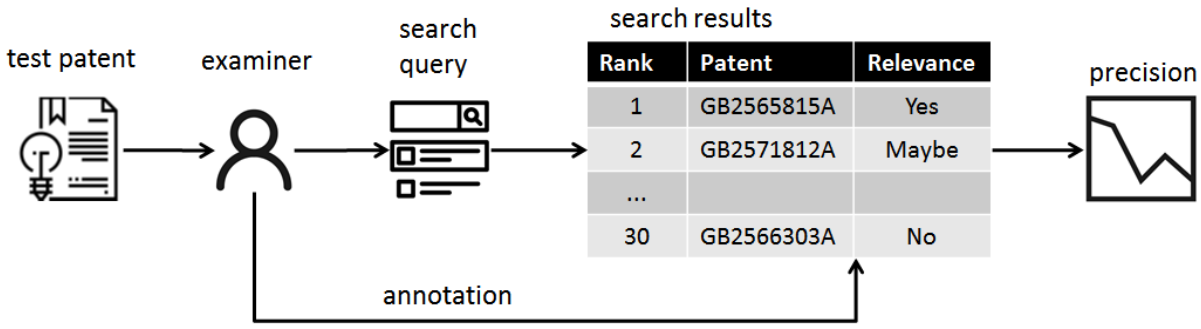


Figure 3: Evaluation framework for information retrieval

Figure 4 provides the distribution of relevance annotations for each test 'query' and each patent examiner separately.

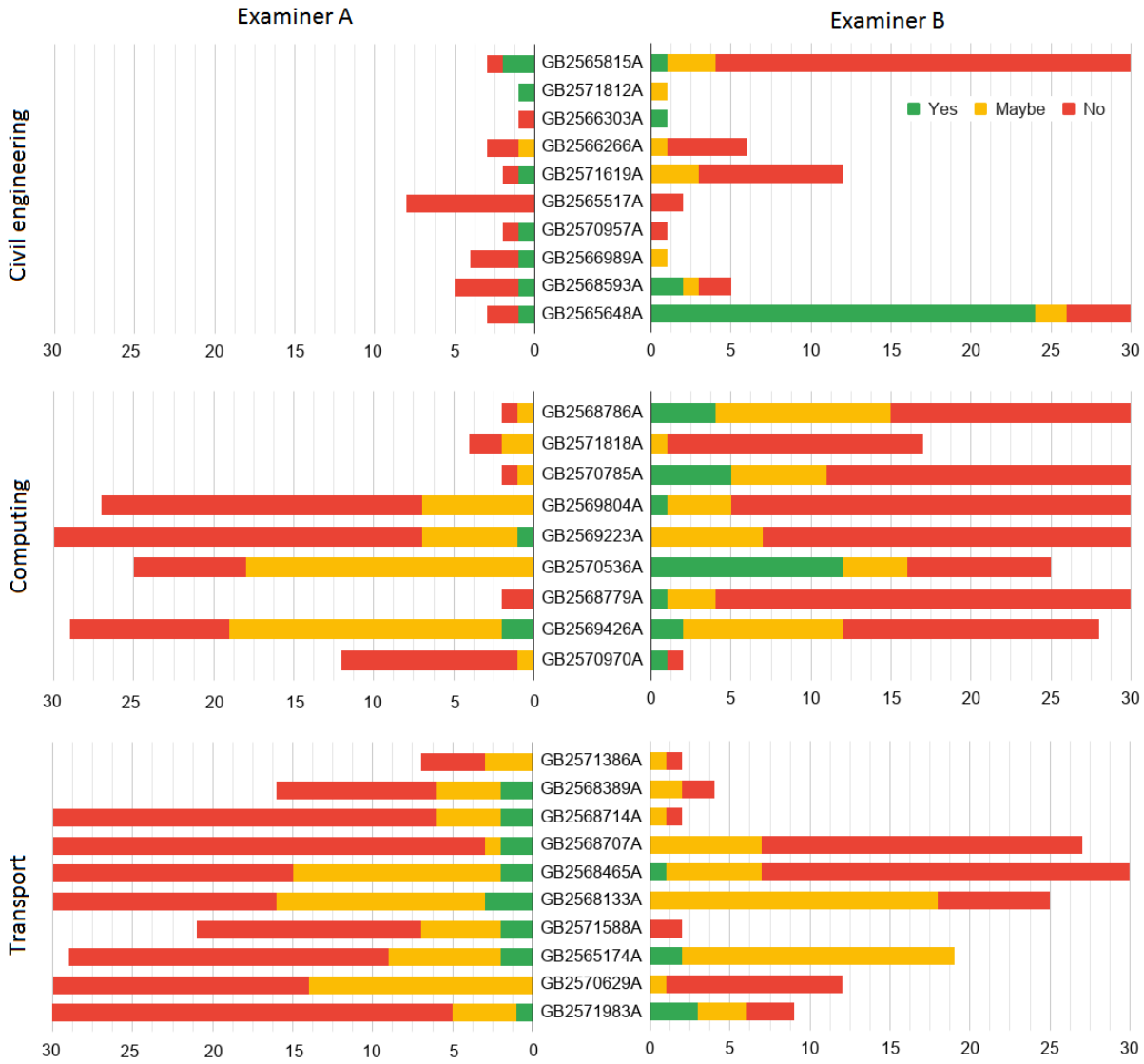


Figure 4: Distribution of annotated results

As the examiners formulated their search queries independently, the search results differed accordingly, hence the variation in the number of retrieved documents and their relevance. Table 16 illustrates the degree of variation in the way search queries were formulated, with some taking full advantage of the search syntax (*e.g.* see Examiner B in computer technology) and others using the search syntax incorrectly (*e.g.* Examiner A in civil engineering used AND to link synonyms such as *water*, *fluid* and *liquid* instead of OR). Despite the *Help* information being provided with the system together with the built-in functionality to semi-automate query formulation, not all examiners seemed to have taken advantage of these features. Ideally, in any future experiments, they should receive prior training and be given a few days to familiarise themselves with the system in order to test the performance of the system rather than the proficiency of the user with respect to system usage.

Table 16: Search queries (note the use of Boolean logic; | indicates an OR operator, multiple keywords suggest an AND operator)

Domain	Patent	Examiner A	Examiner B
Civil engineering	GB2565815A	vacuum vip insulate heat thermal	insulation thermal (panel | board) building construction (vacuum| void) air cavity
	GB2571812A	tent dome geodesic shelter	frame (tent | collapsible shelter) |camp) (pole | rod) junction connect join hub socket
	GB2566303A	panel support mount anchor rail balustrade barrier hand rail	panel balcony (balustrade | handrail) clamp wedge
	GB2566266A	break bend snap tension pipe tube tubular umbilical	bend curve (restrict | limit) (tubular | tube-shaped) prevent wellbore pipe
	GB2571619A	panel plate board water fluid liquid prevent impervious	(panel | board) building water exterior cavity (channel | groove) drainage wall
	GB2565517A	water rain storage tank vessel	water storage reservoir tank collapse portable bladder
	GB2570957A	water fluid layer oil gas interface antenna transmit radio microwave	fluid layer interface wellbore downhole oil (detection | sensing) electromagnetic microwave
	GB2566989A	brick mould cast block build	(panel | board) brick masonry cement mould (imitate | copy)
	GB2568593A	body fluid control drilling hydraulic abandon end of life plug seal string bore wellbore	drilling (sealing | seal off) bore bit cutting sealed oil well plug mill abandon
	GB2565648A	bit sinter cutting tip drill tool	((drilling | boring) | drill) bit hard tungsten sintered (earth | ground) carbide
Computer technology	GB2568786A	view plant gui configure theme	gui* | "user interface*" theme* | color* | colour* | dimension* | size* | font* | display* chang* | adjust* | modif* | adapt* | differ* measur* | control* | sens* | detect* | param*
	GB2571818A	encoding neural network select interpolation	encode encoding encoded "neural network" "machine learning" choose (choice | (pick | selection)) option
	GB2570785A	floorplan robot image	(robot* | automat* | autonom*) + (floor* | plan* | map*)
	GB2569804A	authentication device service two second factor credential registered	authenticat* + (lan | "local area network") + (multiple | second* | devices | plural*) + (register* | subscrib* | join* | registrat*)
	GB2569223A	feed paper printer display	(print* | paper* | sheet*) + (manag* | config* | control*)
	GB2570536A	wearable ecg authentication temperature	(biometric* | heart* | ecg | pulse*) + (authenticat* | authori* | secur*) + (wearabl* | watch* | cloth*)
	GB2568779A	compare specie database	("imag* | object* | scene* | species | visual* recogn*") + (confidence* |

			threshold*) compar* | match* propert* | dimension* | attribute* | shape* | size* | characteristic* | parameter*
	GB2569426A	segmentation roi neural second	"medical imag*" | "medical diagnos*" | cade | cadx roi | loi | "region of interest" | locat* | posit* | area* | region*
	GB2570970A	sharp blur exposure virtual select region	("long-exposure" | "long exposure") virtual photography image (aggregate | combine | flatten | composite)
Transport	GB2571386A	vehicle control autonomous training learning	steer sensor park autonomous
	GB2568389A	aircraft seat passenger light lamp sign display information	aircraft airplane display information sign
	GB2568714A	vehicle car pedal accelerator throttle lock	(pedal | foot pedal) prevent
	GB2568707A	vehicle car load floor spare wheel	floor (raise | lift) (clip | hold | retain)
	GB2568465A	electric battery vehicle car charge control	electric charge range predict
	GB2568133A	child seat vehicle car	(child | baby) seat harness lock
	GB2571588A	sensor detect object target identify classify vehicle car	adaptive cruise camera image coefficient
	GB2565174A	gear change shift foot pedal motorcycle	speed gear (motorcycle | motorbike) (shift | downshift | upshift) (foot|feet|boot|shoe)
	GB2570629A	rear view mirror camera control gesture	rear camera gesture
	GB2571983A	vehicle car driver camera monitor image	camera driver angle

To investigate the impact of different search queries, the corresponding search results between the two examiners was compared. Table 17 shows the total number of patents retrieved by the examiner A but not the examiner B (see column A – B) and vice versa (see column B – A). On the overlapping set of patents (see column A ∩ B), *i.e.* those retrieved (and annotated) by both examiners, inter-annotator agreement using Cohen's kappa coefficient (Cohen, 1960) was calculated. Strict agreement was applied using the original annotations (Yes, Maybe and No). For lenient agreement, the three labels were conflated into two, Relevant (Yes or Maybe) versus Irrelevant (No). Fair agreement was observed in civil engineering and computer technology but was found to be unexpectedly low in transport, which invalidates the evaluation results in this domain. Ideally, in any future experiments, a third independent examiner should resolve any disagreements in order to establish ground truth.

Table 17: Differences in the search results and their interpretation

Domain	A – B	B – A	A ∩ B	Strict agreement	Lenient agreement
Civil engineering	119	206	53	0.4135	0.6710
Computer technology	183	226	78	0.3221	0.5636
Transport	31	80	34	0.1990	0.2446

Finally, using the two labels Relevant versus Irrelevant, the precision was calculated using all annotated patents. The results are given in Table 18. On average, the overall precision varied between 34% and 50% across the six examiners, with the overall average being 38%. Taking the ranking into account, these results were stratified across top 10, 20 and 30 documents (see Figure 5). Upon closer inspection, it was observed that precision at k = 10 varied between 30% and 50%. This means that the first page of search results contained between 3 and 5 relevant documents.

Table 18: Overall precision of information retrieval

	Civil engineering		Computer technology		Transport	
Patent	**A**	**B**	**A**	**B**	**A**	**B**
1	67%	13%	50%	50%	33%	43%
2	100%	100%	50%	6%	50%	38%
3	0%	100%	50%	37%	3%	20%
4	33%	17%	26%	17%	18%	10%
5	50%	25%	23%	23%	30%	50%
6	0%	0%	72%	64%	97%	53%
7	50%	0%	0%	13%	23%	33%
8	25%	100%	66%	43%	37%	31%
9	20%	60%	8%	50%	53%	47%
10	33%	87%			20%	16%
Average	38%	50%	38%	34%	36%	34%

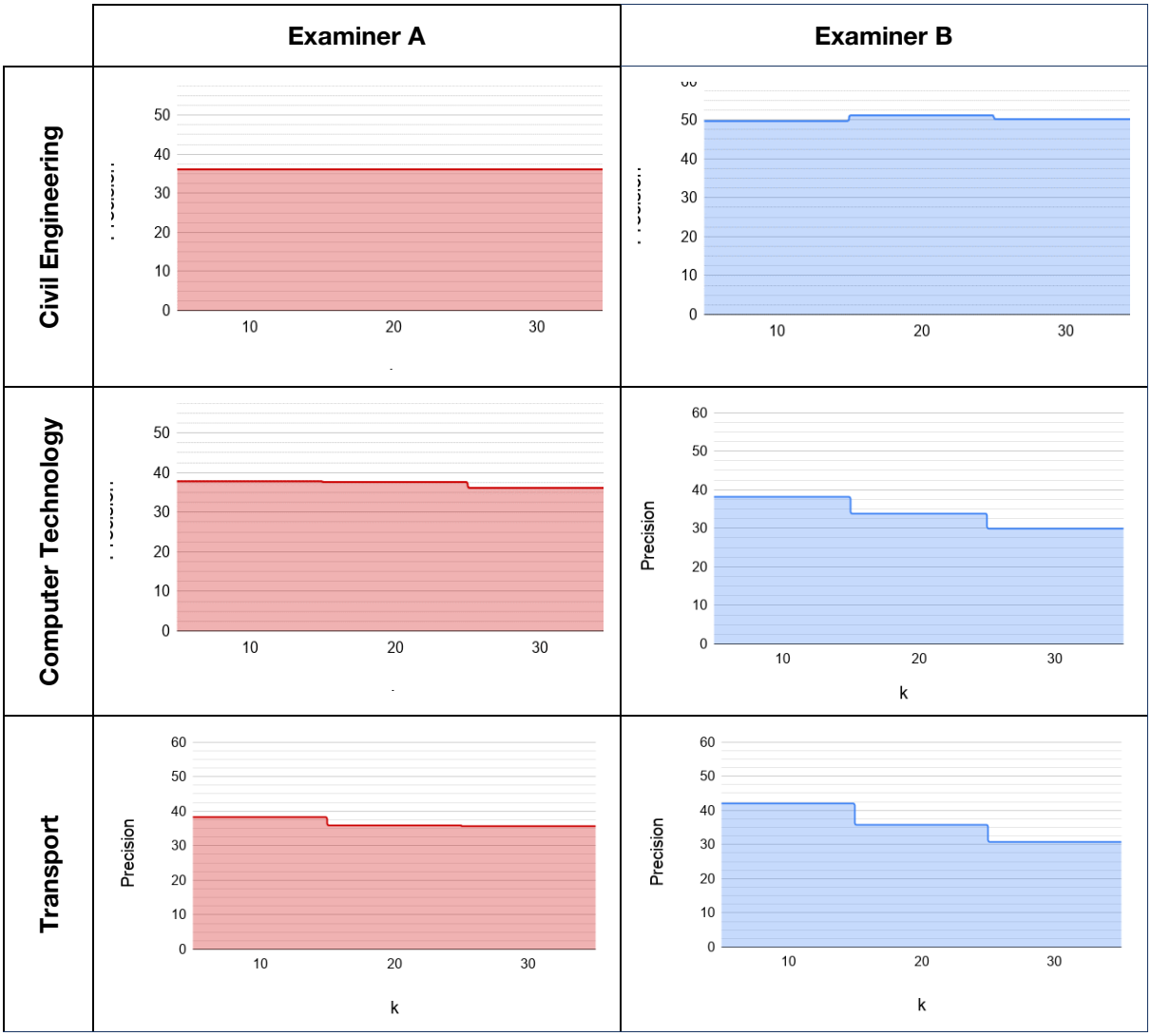


Figure 5: Precision at the top k retrieved documents (k = 10, 20, 30)

4.2.4 Usability

The focus group discussion mainly focused on effectiveness, the ability of the system to retrieve the closest documents and their ranking. The group discussed relevance in the context of prior art searches and the different search strategies patent examiners employ.

In general, the patent examiners were disappointed by the large number of irrelevant items on their list (note that the retrieved 'results' deliberately included a large number of irrelevant patents. The examiners were not told about the 30/30 split at the time of the testing; they were under the impression that the purpose of the study was to generate search queries. The 30/30 split to remove positive bias may have inadvertently led to introducing negative bias.

In most cases, the examiners did not find the suggested keywords very helpful. They thought that topic modelling and visualisation could be potentially very useful. Ranking was in their opinion the most interesting aspect (note that the similarity scores were removed from the interface and the results were presented in random order). The examiners had different views about full-text search strategies (most felt that the full-text was full of misinformation) and which part of the patent provides the best starting point for their searches. The examiners were interested in the potential to discover new classifications and commented that incremental inventions are described using the existing taxonomy, but emerging disruptive technology and radically new inventions require evolving classifications.

The examiners made a number of suggestions about how the system performance could be improved. This includes using flexible search strategies (*e.g.* using different parts of the patent text at different stages of the search process, selecting the most relevant paragraphs to the crux of the invention to make the retrieval task more focused, changing the weighting of the search parameters), hybrid search strategies (*e.g.* combining text and picture searches) and knowledge-based search strategies (*e.g.* enhancing the search with knowledge types such as method, process, methodology, etc.) and using domain-specific ontologies. The usability of the graphical user interface (GUI) and the impact of scrolling, especially on search term/synonym selection were also discussed. The focus group agreed that the best search tool should be one that supports a dynamic, iterative search process.

5 Conclusions

This study aimed to develop a proof-of-concept for an AI-powered patent prior art search/due-diligence check that could form part of the online patent filing and patent examiner prior art searching processes. The proof-of-concept was used as a platform for experimental comparisons between different AI techniques. A wide range of state-of-the-art supervised and unsupervised machine learning approaches were considered that could support the tasks of feature extraction, query expansion, document classification, document clustering and topic modelling.

The study concluded that it was not feasible with current AI tools to provide a fully automated solution as part of the application filing process. Nevertheless, the classification task produced very high classification accuracy, which shows potential to embed this function in the online patent pre-filing process to allow customers thinking of applying for a patent to more easily undertake due diligence checks. The developed proof-of-concept for an AI-powered patent prior art search showed that AI has the potential to assist patent examiners in the future as part of the prior art searching process. Different state-of-the-art AI algorithms can be used to retrieve the closest documents, rank relevant documents, suggest synonyms, suggest classifications, cluster and visualise the retrieved documents/concepts.

The study strongly suggests that the use of AI techniques to retrieve and rank documents could reduce the time and cost of prior art searches, and especially the process of sifting through the large number of patents retrieved. The experimental results for precision varied between 30% and 50%, which means that the first 10 search results contained between 3 and 5 relevant documents. However, AI is less effective in selecting relevant search queries. This was expected as the drafting of the search statement is one of the most important and knowledge-intensive parts of the process. It requires clear understanding of the critical subject matter and the potential novelty of the application. Patent examiners often modify the search statement several times and often use words which do not necessarily appear in the original claims. Drafting of the search statement should remain a human task to suitably bound the AI search because of the wealth of specialist expertise and experience that an examiner has, and should not be something to be performed by AI. Therefore, it could be feasible to provide examiners with a tool to aid searching but an AI-assisted search would require an examiner to formulate a search statement; there are currently no effective AI algorithms which can process the application and generate a search statement.

Another useful function could be topic modelling, *i.e.* the categorisation of patents into easily interpretable topics, each described by a set of keywords. It could be used by both applicants and patent examiners to visualise a domain but could be also utilised by data analysts to discover abstract topics, new terminology and trends in different domains emerging in parts of the world.

The evaluation of the AI algorithms has clearly been challenging without separating the two aspects (search and retrieval). A better approach would have been to use the search statements formed by the patent examiners and focus on the retrieval and ranking aspects of the task only, although this was unfortunately out of the scope of this study because of IPO data sharing restrictions on the unpublished examiner search statements.

The study highlighted significant differences in the search strategies employed by the examiners and the need for innovative tools which support more flexible search strategies. There are opportunities to enhance the current search process by developing new tools for retrieving image-based patents, collecting evidence of due diligence, spotting ambiguity, finding contradictions and visualising relationships among documents.

In conclusion, the study evaluated the viability of different AI technologies for patent prior art searching, including supervised and unsupervised machine learning, and found clear evidence that none of the available AI algorithms on their own can support every aspect of the prior art search process. The study identified the potential of new approaches combining AI with NLP and computational semantics, and highlighted the importance of human-centred decision and performance support tools. There is a need for a larger scale and more rigorous testing with more patents and examiners and more cutting-edge research on new algorithms supporting flexible search strategies and a dynamic, iterative search process.

References

Adams, S. (2018). Is the Full Text the Full Answer? – Considerations of Database Quality. World Patent Information, vol. 54, pp. S66-S77.

Alberts D., Yang C.B, Fobare-DePonio D., Koubek K., Robins S., Rodgets M., Simmons E., DeMarco D. (2017). Introduction to Patent Searching Practical Experience and Requirements for Searching the Patent Space. In: Lupu M, Mayer K, Tait J, Trippe AJ (eds) Current challenges in patent information retrieval. The information retrieval series, vol 29. Springer, Berlin/Heidelberg, pp. 287–304.

Anderson L., Hanbury A., Rauber A. (2017). The Portability of Three Types of Text Mining Techniques into the Patent Text Genre. In: Lupu M, Mayer K, Tait J, Trippe AJ (eds) Current challenges in patent information retrieval. The information retrieval series, vol 29. Springer, Berlin/Heidelberg, pp. 287–304.

Andlauer, D. (2018). Automatic Pre-Search: An overview. World Patent Information, 54, pp. 559-565.

Atkinson KH (2008). Toward a more rational patent search paradigm. In: Proceedings of the 1st ACM workshop on patent information retrieval, PaIR '08. ACM, New York, pp. 37–40.

Azad H.K., Deepak, A. (2019). Query expansion techniques for information retrieval: A survey. Information Processing & Management, Vol. 56, No. 5, pp. 1698-1735.

Bashir S., Rauber A. (2010). Improving Retrievability of Patents in Prior-Art Search. In: Gurrin C. et al. (eds) Advances in Information Retrieval. ECIR 2010. Lecture Notes in Computer Science, vol 5993. Springer, Berlin, Heidelberg.

Blei, D.M., et al. (2003). Latent Dirichlet allocation. Journal of Machine Learning Research, Vol. 3, No. 4-5, pp. 993-1022.

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales, Educational and Psychological Measurement, vol. 20, pp. 37-46.

Cohen, J. (1968). Weighted Kappa: Nominal Scale Agreement with Provision for Scaled Disagreement or Partial Credit. Psychological Bulletin, 70, pp. 213-220.

Crestani, F. (2003). Combination of Similarity Measures for Effective Spoken Document Retrieval. Journal of Information Science, vol. 29(2), pp. 87-96.

Dumais, S.T. (2005) Latent semantic analysis. Annual Review of Information Science and Technology, Vol. 38, pp. 188-230

Elasticsearch (2018) https://www.elastic.co/

Fellbaum, C. (1998, ed.) WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press

Fleiss, J.L., Cohen, J., and Everitt, B.S. (1969). Large-Sample Standard Errors of Kappa and Weighted Kappa. Psychological Bulletin, 72, pp. 323-327.

Fleiss, J.L. and Cohen, J. (1973). The Equivalence of Weighted Kappa and the Intraclass Correlation Coefficient as Measures of Reliability. Educational and Psychological Measurement, 33, pp. 613-619.

Harris CG, Arens R, Srinivasan P (2017) Using classification code hierarchies for patent prior art searches. In: Lupu M, Mayer K, Tait J, Trippe AJ (eds) Current challenges in patent information retrieval. The information retrieval series, vol 29. Springer, Berlin/Heidelberg, pp. 287–304.

Helmers, L., Horn, F., Biegler, F., Oppermann, T., Muller, K.-R. (2019). Automating the Search for a Patent's Prior Art with a Full Text Similarity Search, PLOS ONE, 14(3): e0212103.

Houle, M.E. et al. (2010) Can shared-neighbour distances defeat the curse of dimensionality? In Proceedings of the 22nd International Conference on Scientific and Statistical Database Management, Heidelberg, Germany, pp. 482-500

Hughes, G.F. (1968) On the mean accuracy of statistical pattern recognizers. IEEE Transactions on Information Theory. Vol. 14, No. 1, pp. 55-63.

World Intellectual Property Organisation (2019) International Patent Classification. Available from: https://www.wipo.int/classifications/ipc/en/

Jurafsky, D., Martin J.H. (2008) Speech and Language Processing, 2nd ed. Englewood Cliffs, NJ, USA: Prentice-Hall

Krishna, A., Feldman, B., Wolf, J., Gabel, G., Beliveau, S., Beach, T. (2016) Examiner Assisted Automated Patents Search. AAAI Fall Symposium Series: Cognitive Assistance in Government and Public sector Applications.

Kusner, M.J. et al. (2015) From word embeddings to document distances. In Proceedings of the 32nd International Conference on International Conference on Machine Learning, Lille, France, pp. 957-966

Makarov, M. (2004) The process of reforming the International Patent Classification. World Patent Information, Vol. 26, No. 2, pp. 137-141.

Mikolov, T. et al. (2013) Efficient estimation of word representations in vector space. arXiv, 1301.3781

Miller, G.A. (1995) WordNet: A lexical database for English. Communications of the ACM, Vol. 38, No. 11, pp. 39-41

Oostdijk N, D'hondt E, van Halteren H, Verberne S (2010) Genre and domain in patent texts. In: Proceedings of the 3rd international workshop on patent information retrieval, PaIR '10. ACM, New York, pp. 39–46.

Pennington, J. et al. (2014) Glove: Global vectors for word representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, pp. 1532-1543

Röder, M., Both, A., Hinneburg, A. (2015) Exploring the space of topic coherence measures. In Proceedings of the eighth ACM international conference on Web search and data mining, Shanghai, China, pp. 399-408.

Salton, G., McGill, M. J. (1986) Introduction to modern information retrieval. McGraw-Hill.

Showkatramani G., Krishna A., Jin Y., Pepe A., Nula N., Gabel G. (2018) User Interface for Managing and Refining Related Patent Terms. In: Stephanidis C. (eds) HCI International 2018 – Posters' Extended Abstracts. HCI 2018. Communications in Computer and Information Science, vol 850. Springer, Cham.

Spärck Jones, K. (1972) A statistical interpretation of term specificity and its application in retrieval. Journal of Documentation, Vol. 28, No. 1, pp. 11-21

Spasić, I. (2018) Acronyms as an integral part of multi-word term recognition - A token of appreciation. IEEE Access, Vol. 6, p. 8351-8363

Spasić, I. et al. (2013) FlexiTerm: A flexible term recognition method. Journal of Biomedical Semantics, Vol. 4, 27

Spasić, I. et al. (2018) Head to head: Semantic similarity of multi-word terms. IEEE Access, in press

von Ahn, L. (2006) Games with a purpose. IEEE Computer Magazine, Vol. 39, No. 6, pp. 92-94.

Wolpert, D.H. (1996) The lack of a priori distinctions between learning algorithms. Neural Computation, Vol. 8, No. 7, pp. 1341-1390.

Trippe A, Ruthven I. Evaluating Real Patent Retrieval Effectiveness. In: Lupu M, Mayer K, Kando N,

Trippe AJ, editors. Current Challenges in Patent Information Retrieval. Berlin, Heidelberg: Springer

Berlin Heidelberg; 2017. p. 143–162.

Appendices

Appendix 1: Software libraries

Table 19: Software libraries used to support implementation

ID	Functionality	Library
S1	Linguistic pre-processing - Tokenisation - Lemmatisation - Stemming - WordNet interface	Natural Language Toolkit (NLTK) https://www.nltk.org/
S2	Search engine - Tokenisation - Lemmatisation - Stemming - Unicode normalisation - Indexing - Document similarity	Elasticsearch https://www.elastic.co/
S3	Term extraction - Multi-word terms - Acronyms	FlexiTerm https://github.com/ispasic/FlexiTerm
S4	Word embeddings	word2vec https://code.google.com/archive/p/word2vec/
S5	Machine learning - Vectorization - Classification - Similarity measures	scikit-learn https://scikit-learn.org/stable/
S6	Topic modelling	gensim https://radimrehurek.com/gensim/

Appendix 2: Validation domains

Table 20: Validation domains - civil engineering

Code	Heading
E01	Construction of roads, railways, or bridges
E02	Hydraulic engineering; foundations; soil-shifting
E03	Water supply; sewerage
E04	Building
E05	Locks; keys; window or door fittings; safes
E06	Doors, windows, shutters, or roller blinds, in general; ladders
E21	Earth or rock drilling; mining
E99	Subject matter not otherwise provided for in this section

Table 21: Validation domains – computer technology

Code	Heading
G06	Computing; calculating; counting
G10L	Speech analysis or synthesis; speech recognition; speech or voice processing; speech or audio coding or decoding
G11C	Static stores

Table 22: Validation domains - transport

Code	Heading
B60	Vehicles in general
B61	Railways
B62	Land vehicles for travelling otherwise than on rails
B63	Ships or other waterborne vessels; related equipment
B64	Aircraft; aviation; cosmonautics

Appendix 3: Data format

XML schema: see https://xmlgrid.net/

Appendix 4: Multi-word terms extracted

Table 23: Multi-word terms extracted automatically by FlexiTerm - civil engineering

ID	Term variants	Score	Rank
1	present invention	47.1340	1
2	present disclosure	14.5561	2
3	lift arm assembly	10.9861	3
4	drilling fluid	10.8131	4
5	patent document cf patent document	9.9351	5
6	ESP electric submersible pump electric submersible pumps ESPs	9.8875	6
7	hydraulic pump hydraulic pumps	9.7041	7
8	drill string drill strings	9.0109	8
9	variable speed limit VSL	7.9649	9
10	wall structure	7.7979	10
11	rock drilling machine rock drilling machines	7.6903	11
12	support wall structure	7.6903	11
13	preamble of claim preamble claim preamble of claims preambles of claims	7.6246	12
14	BHA bottom hole assembly bottomhole assembly	7.6246	12
15	screen device screen devices screening device	7.6246	12
16	elevator car	7.2780	13
17	formation fluid formation fluids	7.0701	14
18	arm directional control valve	6.9315	15
19	vacuum thermal insulator	6.5917	16
20	speed limit	6.4694	17
21	architectural decoration panel dry-hang structure	6.4378	18
22	subterranean formations subterranean formation	5.5452	19
23	boom directional control valve	5.5452	19

24	power machines power machine	5.5452	19
25	hydraulic fluid	5.5452	19
26	transverse skeleton transverse skeletons	5.5452	19
27	fiber optic lines fiber optic line	5.4931	20
28	faaade cleaning apparatus	5.4931	20
29	DFA downhole fluid analysis	5.4931	20
30	schematic view of apparatus	5.4931	20
31	exhaust treatment device exhaust treatment devices	5.4931	20
32	composite thermal insulator	5.4931	20
33	hydraulic system	5.1986	21
34	polycrystalline diamond PCD	5.1986	21
35	waterproof membrane waterproofing membrane	4.8520	22
36	axis of rotation	4.8520	22
37	open position open positions	4.8520	22
38	hydraulic excavator hydraulic excavators	4.8520	22
39	steel joist steel joists	4.8520	22
40	rock material	4.8520	22
41	data center	4.8520	22
42	retail package retail packaging retail packages	4.8520	22
43	outer surface outer surfaces	4.8520	22
44	storage compartment	4.8520	22
45	applicant 's application no	4.7365	23
46	suspension systems suspension system	4.6787	24
47	door frame	4.6210	25
48	downhole tool	4.6210	25
49	electronic control unit	4.3944	26
50	sheet metal frame	4.3944	26
51	hydraulic drive system	4.3944	26
52	thermal insulation performance	4.3944	26
53	architectural decoration panel	4.3944	26

54	door panels	4.3899	27
55	carrier element carrier elements	4.1589	28
56	ski slope snow tiller	4.1589	28
57	waste receptacle waste receptacles	4.1589	28
58	earth-boring tools	4.1589	28
59	vsl signs variable speed limit signs	4.1589	28
60	construction machine	4.1589	28
61	metal frame	4.1589	28
62	closed position	4.1589	28
63	boom cylinder	4.1589	28
64	arm cylinder	4.1589	28
65	opening operation restriction device	4.1589	28
66	sandwich support wall structure	4.1589	28
67	plate-shaped support wall structures	4.1589	28
68	hydraulic cylinder	4.1589	28
69	engagement mechanism	4.1589	28
70	guide rails	4.1589	28
71	elevator shaft	4.1589	28
72	data centres data centre	4.1589	28
73	wireless portable listening devices portable wireless listening device	4.1589	28
74	screen roller screen rollers	4.1589	28
75	door end wall door inner wall	4.1589	28
76	rock bolt rock bolts	3.9856	29
77	drive system	3.9278	30
78	prior art	3.8123	31
79	side regions side region	3.4657	32
80	adhesive portion	3.4657	32
81	carrier sheet	3.4657	32
82	wheel loader wheel loaders	3.4657	32
83	spacer plate	3.4657	32
84	electric motor electrical motor	3.4657	32
85	exhaust gas	3.4657	32

86	window system window systems	3.4657	32
87	frame structure structural frame	3.4657	32
88	wire mesh	3.4657	32
89	frame segments frame segment	3.4657	32
90	sash plane	3.4657	32
91	wet area	3.4657	32
92	flow-chart diagram	3.4657	32
93	coupling assembly	3.4657	32
94	spring packet	3.4657	32
95	swash plate angle	3.2958	33
96	hydraulic drive device	3.2958	33
97	artificial neural network	3.2958	33
98	formation fluid property	3.2958	33
99	formation fluid sample	3.2958	33
100	drilling fluid properties properties of such drilling fluids	3.2958	33

Table 24: Multi-word terms extracted automatically by FlexiTerm – computer technology

ID	Term variants	Score	Rank
1	present invention	33.2711	1
2	electronic device electronic devices	21.9497	2
3	operation mode modes of operation	18.9922	3
4	processing device processing devices	18.0218	4
5	PCI-E peripheral component interconnect express	16.6355	5
6	USB universal serial bus	15.5375	6
7	image data	14.8160	7
8	neural network unit	13.6542	8
9	image processing	12.4766	9
10	computing system computer system	12.1301	10
11	present disclosure	11.7835	11
12	mobile terminal mobile terminals	11.7835	11
13	user interface	11.5855	12

14	patent no	11.4947	13
15	fingerprint recognition	11.4864	14
16	mobile device mobile devices	9.9640	15
17	electronic picture books electronic picture book	9.8875	16
18	security system security systems	9.7041	17
19	computing device computing devices	9.4268	18
20	portable device portable devices	9.3575	19
21	detection unit	9.0109	20
22	neural network unit with output buffer feedback	8.9588	21
23	session timeout period	8.7889	22
24	PLM product lifecycle management	8.7889	22
25	image processing apparatus	8.7889	22
26	display device	8.7337	23
27	rfid tag	8.3178	24
28	communication device communication device 5a communication between devices communication device 5b	8.3178	24
29	electronic system electronic systems	7.6246	25
30	DPI dots per inch	7.4513	26
31	power consumption	6.9315	27
32	computer program computer programs	6.9315	27
33	position indicator position indicators	6.9315	27
34	fingerprint data	6.9315	27
35	system for data	6.7582	28
36	NMSs network management systems network management system	6.5917	29
37	gas turbine engine gas turbine engines	6.5917	29
38	REE rich execution environment rich ree	6.5917	29
39	network functions virtualization NFV	6.5917	29

	virtual network function VNF		
40	SCM source code management	6.5917	29
41	frequency band frequency bands	6.4694	30
42	contact lens virtual fitting method	6.4378	31
43	japanese patent no	6.3170	32
44	touch panel touch panels	6.2383	33
45	user guide user guides	6.2383	33
46	fingerprint sensor fingerprint sensors	6.2383	33
47	control device	6.2383	33
48	operation mode control unit	6.2383	33
49	data transfers transfer of data	5.5452	34
50	patent document	5.5452	34
51	virtual machines virtual machine VMs	5.5452	34
52	execution environment	5.5452	34
53	data card data cards	5.5452	34
54	data connector	5.5452	34
55	usb jack	5.5452	34
56	liquid crystal terminal device	5.5452	34
57	chinese patent application no	5.5452	34
58	electronic card electronic cards	5.5452	34
59	DRAM dynamic random access memory	5.5452	34
60	count unit	5.5452	34
61	electronic files electronic file	5.5452	34
62	portable security device portable security devices	5.4931	35
63	ASR automatic speech recognition	5.4931	35
64	operating system OS	5.3719	36
65	network function	5.3141	37

66	transport layer transport layers	5.1986	38
67	image environment	5.1986	38
68	RAMs random access memories	4.9438	39
69	fingerprint recognition apparatus fingerprint recognition apparatuses	4.9438	39
70	computer system interface	4.9438	39
71	audio signal audio signals	4.8520	40
72	host device host of devices	4.8520	40
73	peripheral devices peripheral device	4.8520	40
74	patent literature patent literatures	4.8520	40
75	head-mounted display head-mounted displays	4.8520	40
76	wireless tag wireless tags	4.8520	40
77	pci-e bus	4.8283	41
78	audio file audio files	4.6210	42
79	mobile electronic device mobile electronic devices	4.3944	43
80	SDK software development kit	4.3944	43
81	position detection sensor	4.3944	43
82	fingerprint recognition pattern	4.3944	43
83	flexible circuit board	4.3944	43
84	display image data	4.3944	43
85	displays images	4.3899	44
86	control unit	4.3322	45
87	communication system	4.1589	46
88	power state power state power	4.1589	46
89	resource manager resource management	4.1589	46
90	usb interface	4.1589	46
91	storage system storage systems	4.1589	46
92	internet small computer system interface ISCSI	4.1589	46
93	contact lenses	4.1589	46

94	motion detection	4.1589	46
95	power supply	3.9278	47
96	imaging device	3.9278	48
97	information processing	3.9278	48
98	speech recognition	3.9278	48
99	type fingerprint recognition	3.8451	49
100	remote tlt remote tlts	3.4657	50

Table 25: Multi-word terms extracted automatically by FlexiTerm - transport

ID	Term variants	Score	Rank
1	present invention	48.9824	1
2	electric power	21.7186	2
3	conventional converter	20.3931	3
4	electric vehicle EV electric vehicles	18.7150	4
5	motor vehicle motor vehicles	17.3287	5
6	patent application	14.5561	6
7	UAS unmanned aerial system	14.2820	7
8	pneumatic tire pneumatic tires	13.0543	8
9	shock absorber	12.9387	9
10	power transmission	12.9099	10
11	secondary battery secondary batteries	11.0904	11
12	present disclosure	10.3972	12
13	side wall side walls	10.3972	12
14	torque sensor	9.7041	13
15	door mirror	9.7041	13
16	japanese patent application publication no japanese patent application publications no	9.6566	14
17	vehicle body	9.3575	15
18	hybrid vehicle	9.0109	16
19	road surface	9.0109	16
20	lithium secondary battery lithium secondary batteries	8.7889	17
21	vehicle system vehicle systems	8.6148	18
22	rubber polymer rubber polymers	8.3178	19
23	control unit	8.3178	19

24	rubber composition rubber compositions	8.3178	19
25	vehicle driver	8.3178	19
26	drive wheels wheel drive	8.0405	20
27	power transmission device	7.6903	21
28	patent literature	7.6246	22
29	side sections side section	7.6246	22
30	publication no	7.2780	23
31	wireless power transmission system wireless power transmission systems	6.9315	24
32	transmission shaft support elements	6.9315	24
33	battery pack	6.9315	24
34	gear connection element gear connection elements	6.5917	25
35	patent document patent documents	6.2383	26
36	transmission shaft transmission shafts	6.2383	26
37	emergency vehicles emergency vehicle	5.9611	27
38	wheel hub	5.5452	28
39	blind spots blind spot	5.5452	28
40	pneumatic tyre pneumatic tyres	5.5452	28
41	thrust reverser cowlings	5.4931	29
42	thrust reverser thrust reversers	4.8520	30
43	control apparatus	4.8520	30
44	vehicle for drive	4.8520	30
45	shaft gears shaft gear	4.8520	30
46	japanese patent application laid-open JP-A	4.8520	30
47	cargo compartment cargo compartments	4.8520	30
48	work vehicle	4.8520	30
49	p-polarized light s-polarized light p-polarized light from light	4.8520	30
50	milling machine	4.8520	30
51	emergency vehicle patient transport systems emergency vehicle patient transport system	4.8283	31
52	power transfer unit	4.3944	32
53	wheel suspension arrangement	4.3944	32

54	side rear view	4.3944	32
55	pulse width modulation PWM	4.3944	32
56	rotary connector device	4.3944	32
57	electric drive vehicle vehicle with electric drive	4.3944	32
58	clutch control unit	4.3944	32
59	railway freight car	4.3944	32
60	magnetic field generator	4.3944	32
61	half-latch engagement portion	4.3944	32
62	HEV hybrid electric vehicle	4.3944	32
63	lng storage tank	4.3944	32
64	primary output command	4.3944	32
65	BMS battery management system battery management systems	4.3944	32
66	moulded article moulded articles	4.1589	33
67	electric drive	4.1589	33
68	mixed cathode active material	4.1589	33
69	kick-up frame connection structure	4.1589	33
70	grip performance on road surfaces grip performance on such road surfaces	4.1589	33
71	sudden inattention	4.1589	33
72	outer ring outer rings	4.1589	33
73	magnetic field	4.1589	33
74	hev mode mode of hybrid electric vehicle	4.1589	33
75	conventional transportation scheduling method	4.1589	33
76	natural gas	4.1589	33
77	japanese unexamined patent application	4.1589	33
78	speed change	4.1589	33
79	gas turbine engine	4.0282	34
80	knuckle boom	3.8123	35
81	air system	3.6968	36
82	landing gear	3.4657	37
83	aircraft engine aircraft engines	3.4657	37
84	compressor section	3.4657	37
85	vertical distance	3.4657	37
86	steering torque	3.4657	37
87	steering system steering systems	3.4657	37

88	storage system	3.4657	37
89	control device	3.4657	37
90	power source power sources	3.4657	37
91	rear wheels	3.4657	37
92	transfer unit	3.4657	37
93	control systems control system	3.4657	37
94	tapered rollers	3.4657	37
95	acoustic resonance acoustic resonances	3.4657	37
96	safety arrangement	3.4657	37
97	car sunshades car sunshade	3.4657	37
98	energy source energy sources	3.4657	37
99	air guide	3.4657	37
100	vehicle driveline system	3.2958	38

Appendix 5: Multi-word term dendrograms

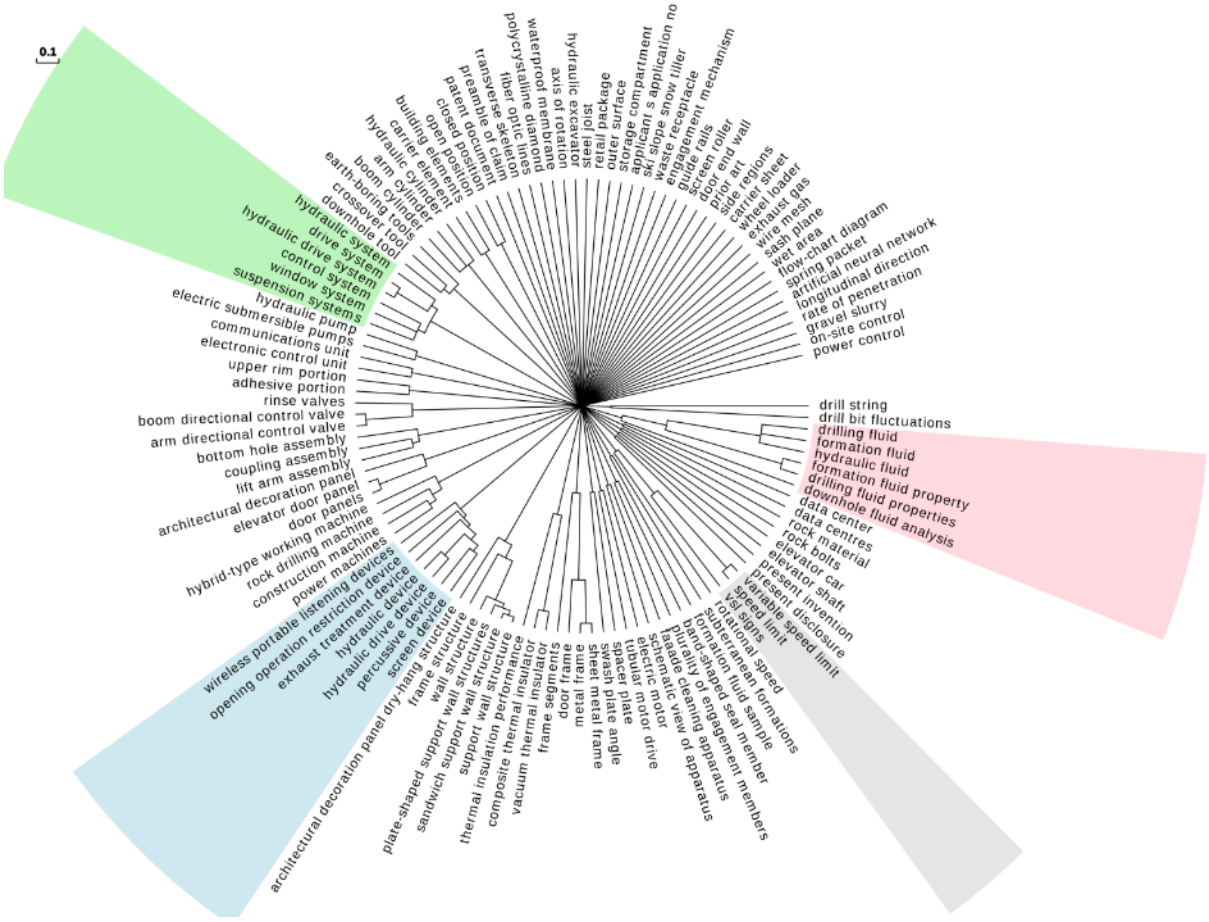


Figure 6: Dendrogram of terms from civil engineering

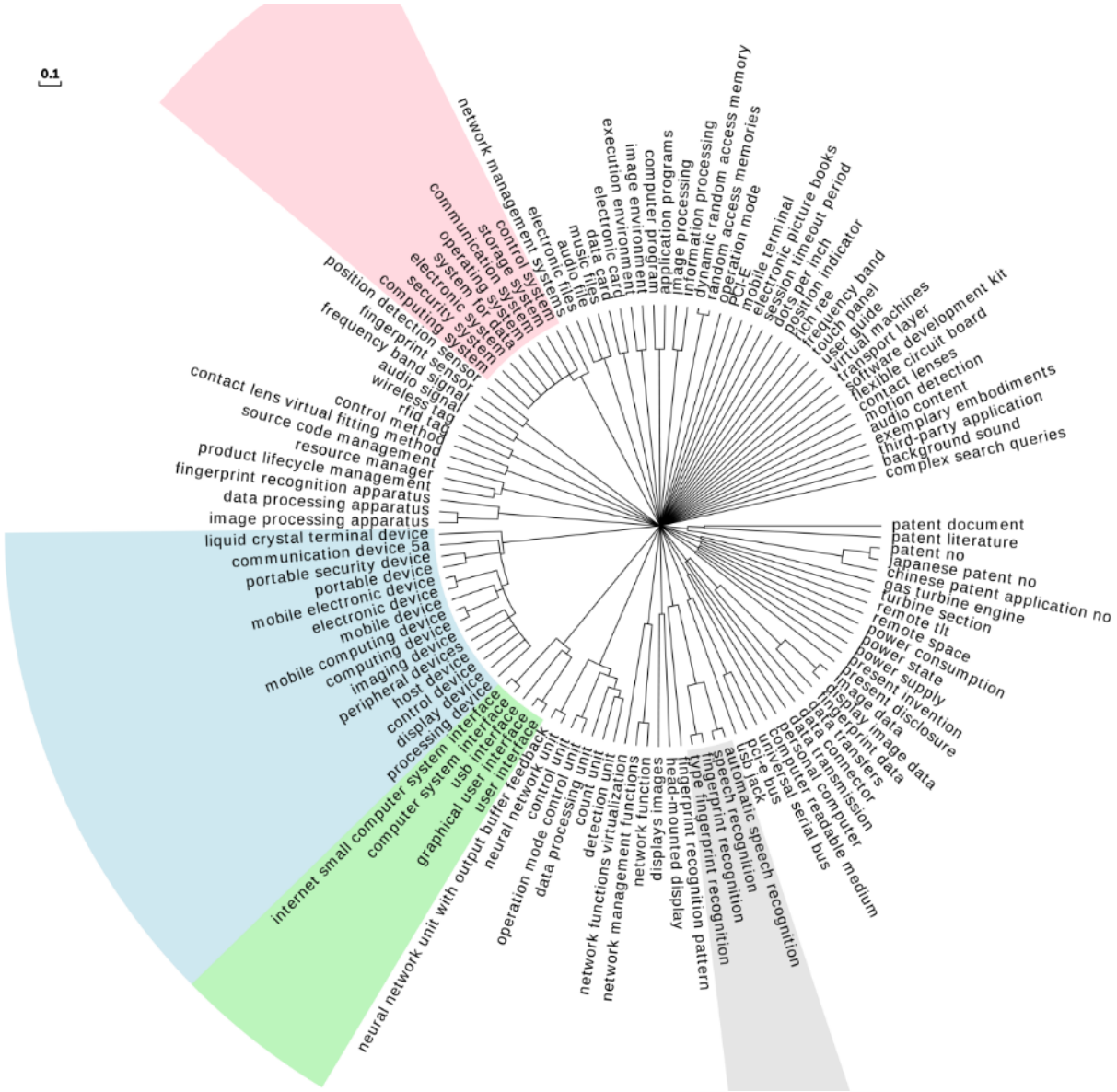


Figure 7: Dendrogram of terms from computer technology

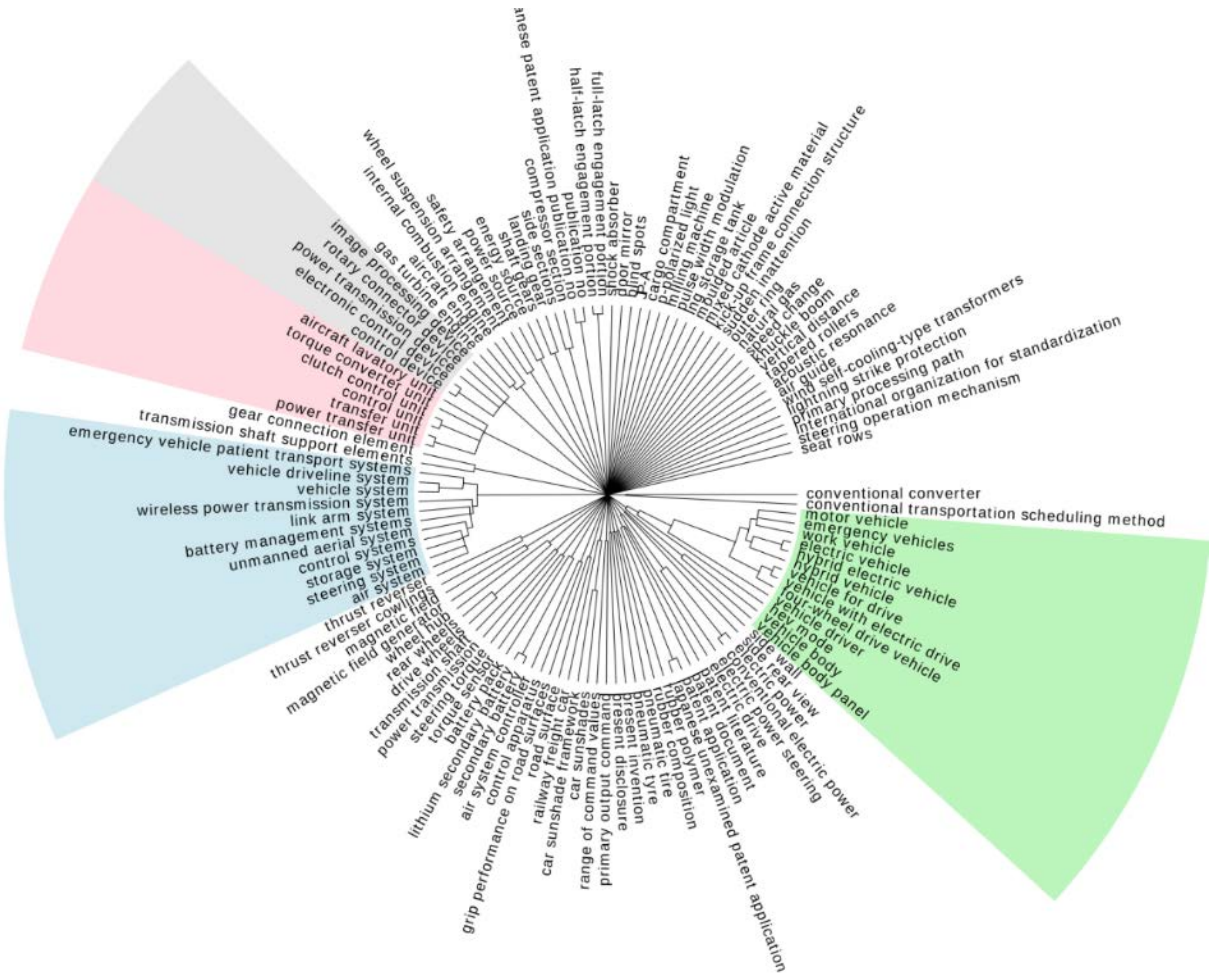


Figure 8: Dendrogram of terms from transport

Appendix 6: Domain-specific uses of the word "driver"

Figure 9 to Figure 11 show an example of domain-specific uses of a particular word, in this instance “driver”, in each of the three technology domains considered.

```
 8663  be/operated with standard spanner, socket, screw driver, ratchet, power tool, specialised tool and/
20006         circuit 1024 may also comprise a graphics driver card./The interface circuit 1024 also
21046   to control slippage in such a way that when the driver of the/mining machine detects slippage, he
21052    because even a long time can be taken/from the driver to detect the slippage and further to start
21055  equipment, that is, equipment/operated without a driver./Furthermore, solutions are known in which
25372   site being known, a display device visible to a driver of the/earthmover shows the difference
38190  port disposed in the non-occluding ear portion; a/driver assembly positioned within the housing
38190    having a front volume disposed in front of the driver/assembly and a back volume disposed behind
38191   /assembly and a back volume disposed behind the driver assembly; and an acoustic insert positioned
38192   insert positioned/within the housing behind the driver assembly and attached to an interior
38249      and second ends; a speaker/assembly having a driver unit and a directional sound port proximate
38250    offset from/the longitudinal axis, wherein the driver unit is aligned to emit sound from the
38258    in the/housing; a speaker assembly including a driver unit and a directional sound port, wherein
38259    unit and a directional sound port, wherein the/driver unit is aligned to emit sound from the
38281       disposed within the/housing and including a driver unit comprising a first magnet, the driver
38281      a driver unit comprising a first magnet, the driver unit aligned to emit/sound from the
38284       can be disposed within the ear portion. The driver unit can/include a diaphragm and a voice
38713  of the/earbud. The speaker assembly can include a driver unit aligned to emit sound from the
38714   emit sound from the directional/sound port. The driver unit can include an electromagnetic voice
38715            voice coil, a speaker diaphragm and a/driver magnet (shown in FIG. 3 as magnet 325)
38716     signals and produce sound. In addition to the driver magnet,/earbuds according to some
39435   designed to direct sound waves from an internal driver (e.g., part of an earbud speaker, not/shown
39475  shown in FIG. 32,/earbuds 3000a, 3000b include a driver 3205, an acoustic insert 3220, a flexible
39476    battery 3335 and an electrical connector 3060. Driver 3205 is located/within ear portion 3010 and
39477      a front acoustic volume 3210 in front of the driver and a rear/acoustic volume 3215 behind the
39478  driver and a rear/acoustic volume 3215 behind the driver. Driver 3205 can include an electromagnetic
39478    a rear/acoustic volume 3215 behind the driver. Driver 3205 can include an electromagnetic voice
39479  3205 can include an electromagnetic voice coil, a/driver magnet and a speaker diaphragm (not shown
39480   . 32). Acoustic insert 3220 is positioned/behind driver 3205 and adhered to housing 3005, as
39682  housing) that can be used to provide venting for driver (e.g., speaker)/in earbud 3000a. More
39689  3025. These apertures can provide venting for the driver, sound for/the user, and can help tune the
39699  5. As shown in/FIG. 45B, acoustic insert 4505 and driver 4570 are disposed within housing 3005 (
39701  acoustic insert/4505 is described in more detail. Driver 4570 can be positioned within cavity 4510
39702  3005,/forming a front volume 4515 in front of the driver and a back volume 4520 behind the driver.
39702   of the driver and a back volume 4520 behind the driver. Driver/4570 can be positioned such that
39702  driver and a back volume 4520 behind the driver. Driver/4570 can be positioned such that front
40271  port disposed in the non-occluding ear portion; a/driver assembly positioned within the housing
40271    having a front volume disposed in front of the driver/assembly and a back volume disposed behind
40272   /assembly and a back volume disposed behind the driver assembly; and an acoustic insert positioned
40273   insert positioned/within the housing behind the driver assembly and attached to an interior
40295  a/stem; a cavity formed within the ear portion; a driver assembly positioned within the cavity and/
40296  /defining a front volume disposed in front of the driver assembly and a back volume disposed behind/
40297    assembly and a back volume disposed behind/the driver assembly; an acoustic insert positioned
40297    insert positioned within the cavity behind the driver assembly and/attached to an interior
40443      and second ends; a speaker/assembly having a driver unit and a directional sound port proximate
40444    offset from/the longitudinal axis, wherein the driver unit is aligned to emit sound from the
40465      and second ends; a speaker assembly having a/driver unit and a directional sound port proximate
40466    offset from the longitudinal/axis, wherein the driver unit is aligned to emit sound from the
40473    in the/housing; a speaker assembly including a driver unit and a directional sound port, wherein
40474    unit and a directional sound port, wherein the/driver unit is aligned to emit sound from the
40525       disposed within the/housing and including a driver unit comprising a first magnet, the driver
40525      a driver unit comprising a first magnet, the driver unit aligned to emit/sound from the
40528       can be disposed within the ear portion. The driver unit can/include a diaphragm and a voice
40535       disposed within the housing and including a driver unit comprising a/first magnet, a diaphragm
40537  diaphragm in response to electrical signals, the driver unit aligned to/emit sound from the
40697        within the enclosed cavity and including a driver unit/comprising a magnet, the driver unit
40698  including a driver unit/comprising a magnet, the driver unit aligned to emit sound from the
40716       disposed/within the housing and including a driver unit comprising a magnet, the driver unit
40716  including a driver unit comprising a magnet, the driver unit aligned to emit/sound from the
```

Figure 9: Concordances from civil engineering

```
 493          for bus size apply, including space and driver constraints/of the physical layout, and the
2606 via plug-and-play or other hardware/detection and driver selection process, and the driver, agent
2606  /detection and driver selection process, and the driver, agent and support software for the
2806     in the host. The support software comprises a driver and an application/programming interface (
2807       the wireless module with the host. Both the driver/and API are based on the standard driver
2808 Both the driver/and API are based on the standard driver for a cellular wireless module, but
2813 OTA) security communications channel, the API and driver are extended to allow/trusted applications
2902          agent, a calling agent, a full function driver agent, a/partial driver agent, a Computrace
2903    agent, a full function driver agent, a/partial driver agent, a Computrace agent or other similar
3009 used. Also installed in within the OS is a module driver 16 for allowing the host agent to/interact
3010  cellular wireless security module 19. The module driver 16 may include a/compressed agent 17 and
3013   enables 34 driver-based persistence. The/module driver 16 comprising the compressed agent 17 may
3014       readable instructions 43 forming the module driver/installer and the necessary driver code and
3015     the module driver/installer and the necessary driver code and files, and a compressed version of
3016 version of the host agent 44./The wireless module driver and compressed agent may also be installed
3017  Microsoft update 45/which includes the necessary driver code 46 and compressed agent 47.
5850    a display area (a sensor area) 21, a display/H driver 22, a display V driver 23, a sensor readout
5850      area) 21, a display/H driver 22, a display V driver 23, a sensor readout H driver 25 and a
5850     22, a display V driver 23, a sensor readout H driver 25 and a sensor V driver 24./The display
5850   23, a sensor readout H driver 25 and a sensor V driver 24./The display area (the sensor area) 21
5858      are arranged in a/matrix form./The display H driver 22, together with the display V driver 23,
5858  display H driver 22, together with the display V driver 23, line-sequentially drives a liquid/
5861    display drive circuit 12./The sensor readout H driver 25, together with the sensor V driver 24,
5861   readout H driver 25, together with the sensor V driver 24, line-sequentially drives a/
5863     will be described later, the sensor readout H driver 25 and the sensor V/driver 24 perform an
5864 , the sensor readout H driver 25 and the sensor V/driver 24 perform an image pickup drive so that
12183  the display unit 130 may further include a panel/driver (not shown) to drive the display panel./The
18260     166, a photodiode array 167, a laser emitting driver 168, and a laser array 169./The amplifier 1
18266      Array,/photodiode array). The laser emitting driver 168 may be a VCSEL Driver (Vertical Cavity
18266   ). The laser emitting driver 168 may be a VCSEL Driver (Vertical Cavity Surface/Emitting Laser
18267    Driver (Vertical Cavity Surface/Emitting Laser Driver, vertical cavity surface emitting laser
18267    Driver, vertical cavity surface emitting laser driver). The laser array 169 may be a/VCSEL Array
18273 21 of the previous embodiment; the laser/emitting driver 168 and the laser array 169 correspond to
18284   the previous embodiment; and/the laser emitting driver 168 and the laser array 169 correspond to
19806 ) to access trusted application services. The TEE driver 28, the/monitor 32, and the TEE core 48 are
20006   REE kernel 22 provides an RTC core 26 and a TEE driver 28, and the TEE kernel 42 provides the/TEE
20012  2, a monitor 32 acts as a bridge between/the TEE driver 28 and TEE core 48, and can facilitate the
20014   80 of Fig. 5)./Referring now to the REE 12, TEE driver 28 provides access to the TEE 14 for client
20023   RTC core corresponds to a standard Linux kernel/driver. Optionally, the user space 20 may also
20024    include a Replay Protected Memory Block (RPMB)/driver 34 and a TEE supplicant 36 (if RPMB
20024    features are utilized). Collectively, the RPMB driver 34/and TEE supplicant 36 are used to store
20068 and write requests to the RPMB 68 (e.g., via/RPMB driver 34 and TEE supplicant 36)./Fig. 5
20115     is only an example, and if excluded the "RMPB Driver" shown in these/figures could instead
20116    in these/figures could instead correspond to a driver for storing/retrieving data from other
20119          , if the TEE core 48 may act as its own/driver for TEE memory (e.g., the EEPROM 78)./
20122  202), and/stores that value in the RPMB via RPMB driver 34 (step 204). Step 204 may be performed in
43504          with the passengers/in general, and the driver in particular. Some aspects of the user
43509   may be designed to provide the user, either/the driver or a passenger, with the current status of
43542  the dashboard may be easily viewed by either the driver or/the front seat passenger, but may be
43554   the user interface is adjacent to the vehicle's driver seat; (ii) a user interface/positioning
43561       in the data entry position/is closer to the driver seat than when the user interface is in the
43569        position selector is settable by a vehicle driver./Preferably when the user interface is in
43572  user interface is preferably located between the driver seat and the/adjacent passenger seat. The
43697     ). Touch screen user interface/105 allows the driver, or a passenger, to interact with the
43704  /management system to provide information to the driver and/or passenger, information such as a/
43712   user interface 105 may also be used to warn/the driver of a vehicle condition (e.g., low battery
43791    allowing access to the interface by either the driver or the/passenger. In some embodiments the
43792          the interface may be angled towards the driver, and/or positioned/closer to the driver's
43794      side of the vehicle, thus providing improved driver access to the interface./It should be
```

Figure 10: Concordances from computer technology

```
1318        .The aircraft system of claim 1 wherein the driver/includes a solenoid.A method for operating
6102          the blind spot of the motor vehicle./The driver of a motor vehicle must recognize that
6103 /respect to the motor vehicle being driven by the driver. As such, the driver must constantly
6103  vehicle being driven by the driver. As such, the driver must constantly review/his or her
6105    , would cause a collision. Tools that assist a driver in reviewing the space surrounding the/
6107   side rear view mirrors. These mirrors allow the/driver to review the surroundings generally
6107    the surroundings generally disposed behind the driver without the driver having to/turn his or
6107  generally disposed behind the driver without the driver having to/turn his or her head more than a
6110      spots are spaces that are not visible to the/driver when the driver is looking in the mirrors
6110       that are not visible to the/driver when the driver is looking in the mirrors and viewing of
6111     and viewing of these blind spots requires the/driver to turn his or her head to look to see if
6113       enter blind spots. These sensors notify the driver that a blind spot is now being/occupied.
6115 is critical as it is in the best interests of the driver to have these located within/the driver's
6116        peripheral vision and in an area where the driver frequently looks./US 2006/0056003 A1
6129        peripheral vision and in an area where the driver frequently/looks./This object is achieved
6164  side rear view mirror assembly 26 that is on the driver side 22 of the/motor vehicle 10. It will be
6166  that the following discussion/with regard to the driver side rear view mirror assembly 26 applies
6169           of the mirror includes a portion of the driver side 22 of the motor vehicle/therein. The
6235        when an object is in the blind spot on the driver side of the motor vehicle 10 or when an/
6431   a vehicle./Traffic accidents often occur due to driver impairment caused by, for example,
6432    , etc. In order to prevent accidents caused by driver impairment, it may/be vital to provide the
6433 driver impairment, it may/be vital to provide the driver with a warning message to re-establish the
6433      message to re-establish the attention of the driver/to the surrounding traffic situation, or in
6434        , or in a critical situation to advice the driver to take a/break or switch to another driver
6435   the driver to take a/break or switch to another driver of the vehicle./Several systems are known
6436     which attempt to predict the behaviour of the driver and provide the/driver with a warning
6437       the behaviour of the driver and provide the/driver with a warning message in the event that he
6440   be vital to provide a warning message/which the driver is capable to assimilate and react to. In
6441          warning messages for different causes of driver impairment. For example, a drowsy driver/
6441       of driver impairment. For example, a drowsy driver/should be given a warning message intended
6443    intended for e.g. an/intoxicated or distracted driver. A warning message intended for e.g. a
6443  A warning message intended for e.g. a distracted driver when the/driver in fact is drowsy, may
6444    intended for e.g. a distracted driver when the/driver in fact is drowsy, may result in that the
6444 /driver in fact is drowsy, may result in that the driver does not assimilate and react to the/
6446   for monitoring the physiological behaviour of a driver. The system/measures, for example, the
6448       , etc. A warning message is provided to the driver of the vehicle when the/system detects one
6452    uses EEG to determine the attention level/of a driver. The attention level is thereafter compared
6456  that measures duration of inattentive state of/a driver. Based on the duration of measured
6457   warning device provides warning/messages to the driver./WO 2007/090 896 discloses a method for
6460         parameters. Depending on the state of the driver,/the vehicle dynamics are adjusted./DE 10 2
6466        a device and method for determining when a driver is not paying enough/attention during
6468    is based on steering wheel behavior as/well as driver behavior such as e.g. eye-lid closure,
6471           well between the actual/causes for the driver impairment, i.e. to specifically determine
6473   improving performance estimation/of the vehicle driver./According to an aspect of the invention,
6641  to be able to more precisely detect the cause of driver impairment. This may provide e.g./warning
6724   calibrated each time an operator 202 enters the driver seat of the/car 100. As the camera system 2
8176     , and offers good/driving comfort to both the driver and to possible passengers on the sledge./
10536       to have a drive train system that provides driver selectable AWD/capability by redistributing
11022       shock feeling, by which an upper body of a driver may bend down forward in a vehicle/travel
11153   pressing level (accelerator opening level) of a driver and a signal/from the engine rotational
11166 if it is not necessary to accelerate a vehicle, a driver recovers the accelerator pedal./When the
11179       24 detects a range position selected when a driver operates a selector./When a D-range is
11300        shock/feeling, by which an upper body of a driver may bend down forward in a vehicle travel
11361 the wheel side portion are locked at/timing t4, a driver may feel a torque shock if a torque change
11364 a lock-up state and an open state at timing t3, a driver/does not feel a shock even when the engine
12326      the on and off states of/the LED, by which a driver may control the LED to turn on/off. When
12438 60. The lamplight control switch may be used by a driver to control the illuminating element/141 to
12440 the illuminating/element 141 is turned off by the driver, and by cooperation between the first shell
12443  the illuminating element 141 is/turned on by the driver, by cooperation between the first shell 110
```

Figure 11: Concordances from transport

Appendix 7: Nearest neighbours of the word "driver" in the word embeddings space

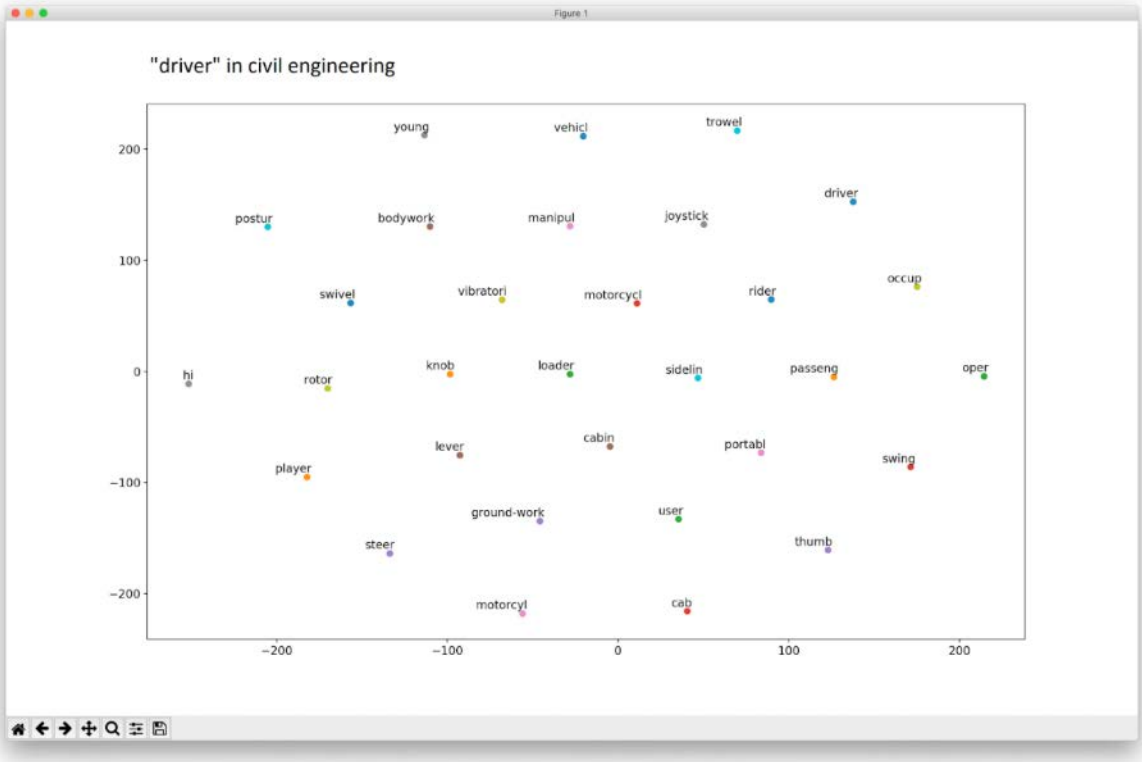


Figure 12: Visualisation of word embeddings from civil engineering

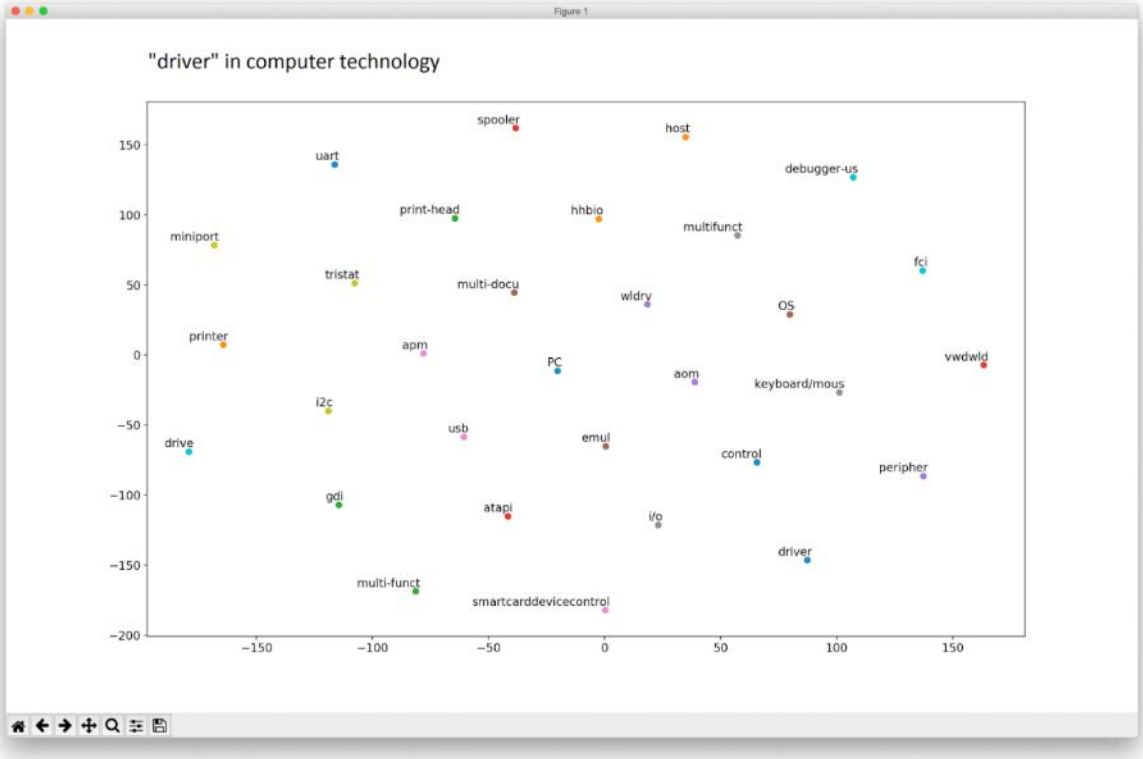


Figure 13: Visualisation of word embeddings from computer technology

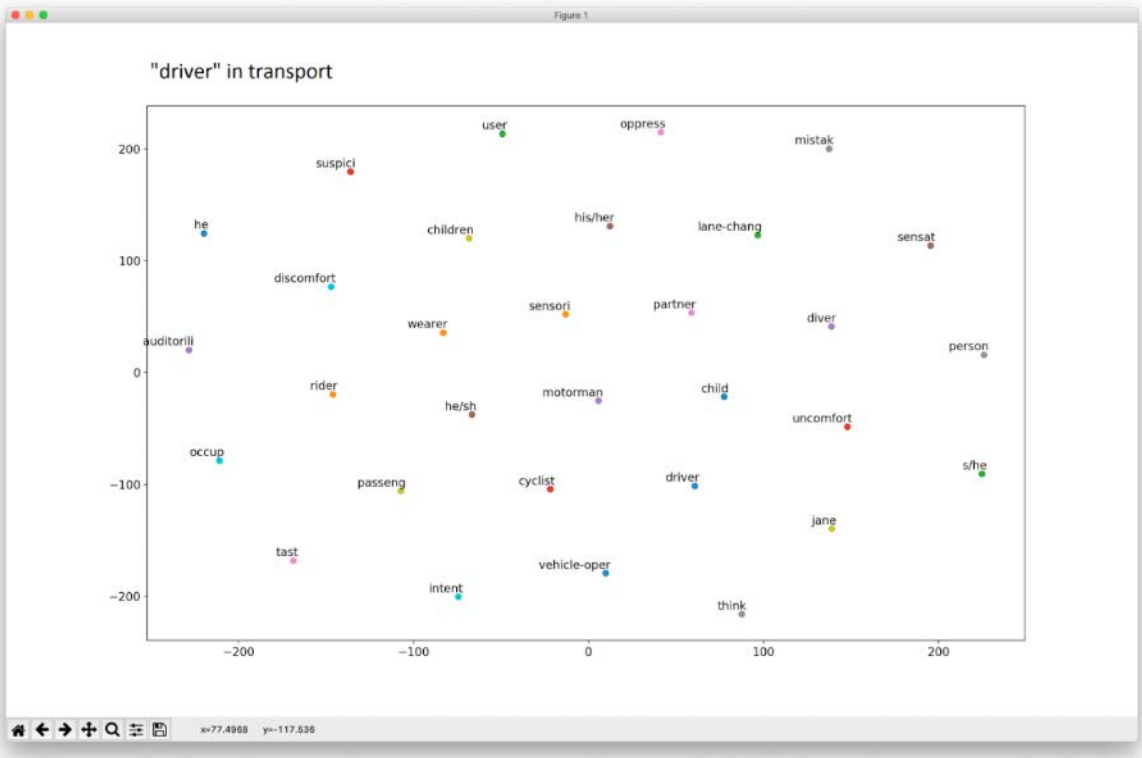


Figure 14: Visualisation of word embeddings from transport

Appendix 8: Representation of the word "driver" in WordNet

Word to search for: driver Search WordNet

Display Options: (Select option to change) Change

Key: "S:" = Show Synset (semantic) relations, "W:" = Show Word (lexical) relations

Display options for sense: (gloss) "an example sentence"

Noun

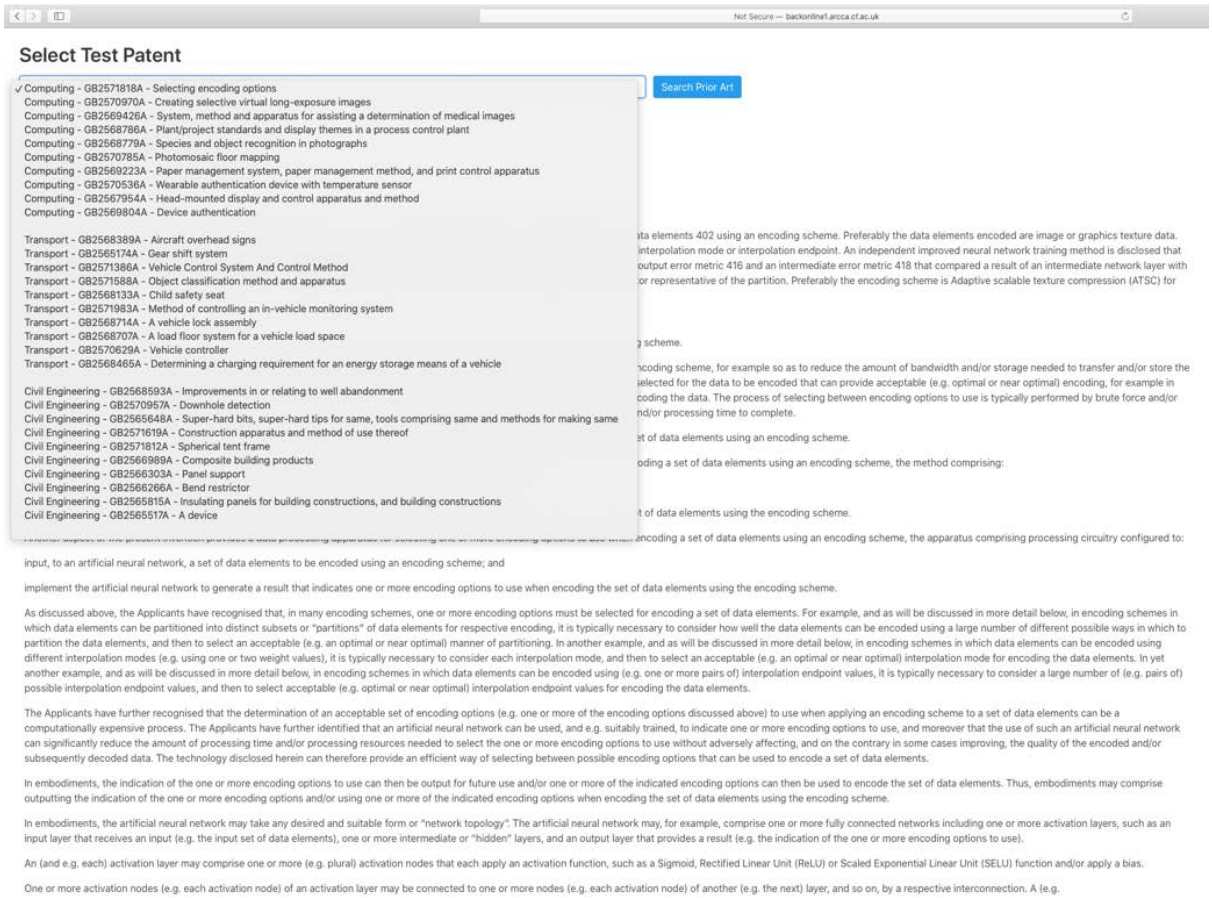
- S: (n) **driver** (the operator of a motor vehicle)
  - *direct hyponym* / *full hyponym*
    - S: (n) busman, bus driver (someone who drives a bus)
    - S: (n) chauffeur (a man paid to drive a privately owned car)
    - S: (n) designated driver (the member of a party who is designated to refrain from alcohol and so is sober when it is time to drive home)
    - S: (n) honker (a driver who causes his car's horn to make a loud honking sound) *"the honker was fined for disturbing the peace"*
    - S: (n) kerb crawler (someone who drives slowly along the curb seeking sex from prostitutes or other women)
    - S: (n) motorist, automobilist (someone who drives (or travels in) an automobile)
    - S: (n) owner-driver (a motorist who owns the car that he/she drives)
    - S: (n) racer, race driver, automobile driver (someone who drives racing cars at high speeds)
    - S: (n) road hog, roadhog (a driver who obstructs others)
    - S: (n) speeder, speed demon (a driver who exceeds the safe speed limit)
    - S: (n) tailgater (a driver who follows too closely behind another motor vehicle)
    - S: (n) taxidriver, taximan, cabdriver, cabman, cabby, hack driver, hack-driver, livery driver (someone who drives a taxi for a living)
    - S: (n) teamster, trucker, truck driver (someone who drives a truck as an occupation)
    - S: (n) test driver (a driver who drives a motor vehicle to evaluate its performance)
  - *direct hypernym* / *inherited hypernym* / *sister term*
    - S: (n) operator, manipulator (an agent that operates some apparatus or machine) *"the operator of the switchboard"*
  - *antonym*
  - *derivationally related form*
- S: (n) **driver** (someone who drives animals that pull a vehicle)
- S: (n) **driver** (a golfer who hits the golf ball with a driver)
- S: (n) **driver**, device driver ((computer science) a program that determines how a computer will communicate with a peripheral device)
  - *domain category*
  - ***direct hypernym*** / *inherited hypernym* / *sister term*
    - S: (n) utility program, utility, service program ((computer science) a program designed for general support of the processes of a computer) *"a computer system provides utility programs to perform the tasks needed by most users"*
  - *derivationally related form*
- S: (n) **driver**, number one wood (a golf club (a wood) with a near vertical face that is used for hitting long shots from the tee)

Figure 15: WordNet's web interface

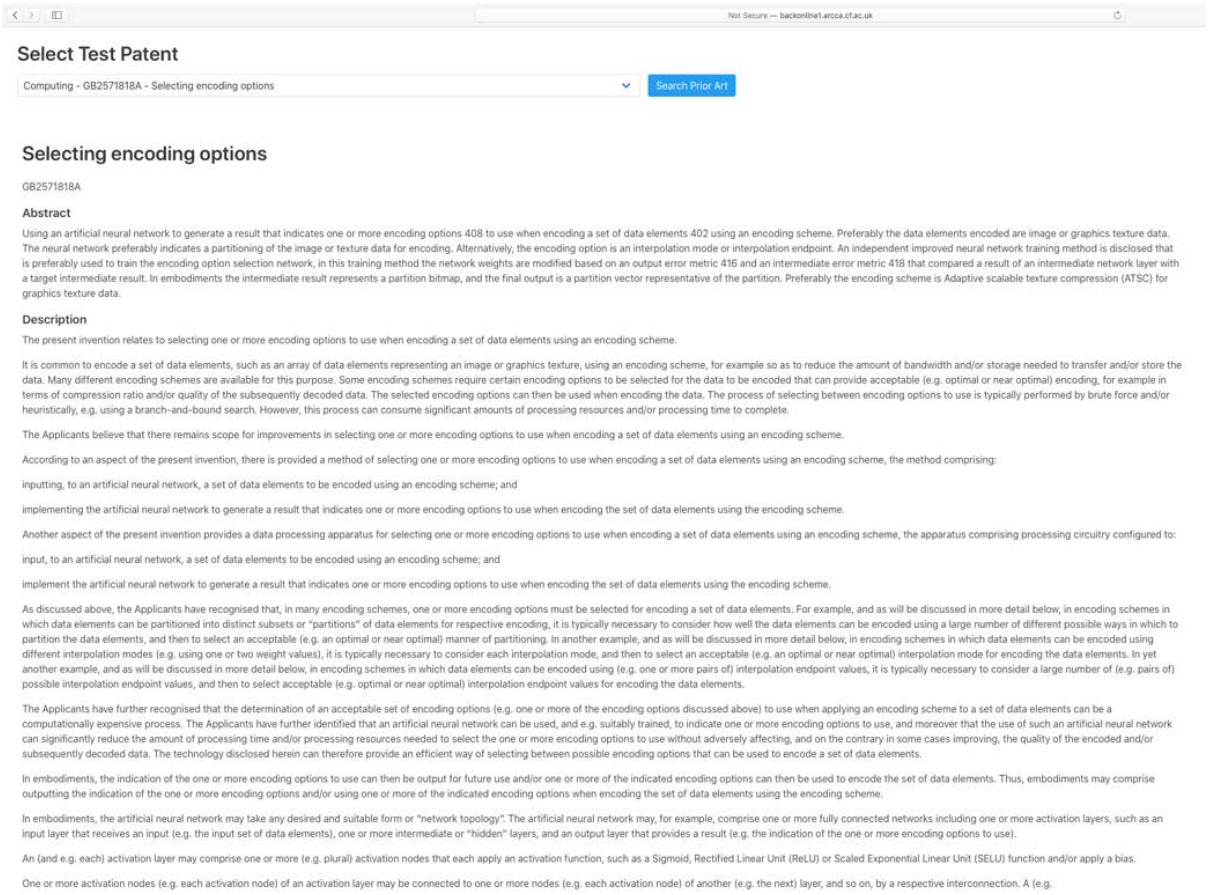
Appendix 9: Local installation of Elasticsearch service

Using the Python bindings to Elasticsearch, a web interface was created that enables patent examiners to search for prior art in the three domains of interest: civil engineering computer technology and transport. Examiners start their search by selecting a 2019 patent from a drop-down menu. This menu provides three groups of 10 patents with 10 per domain.

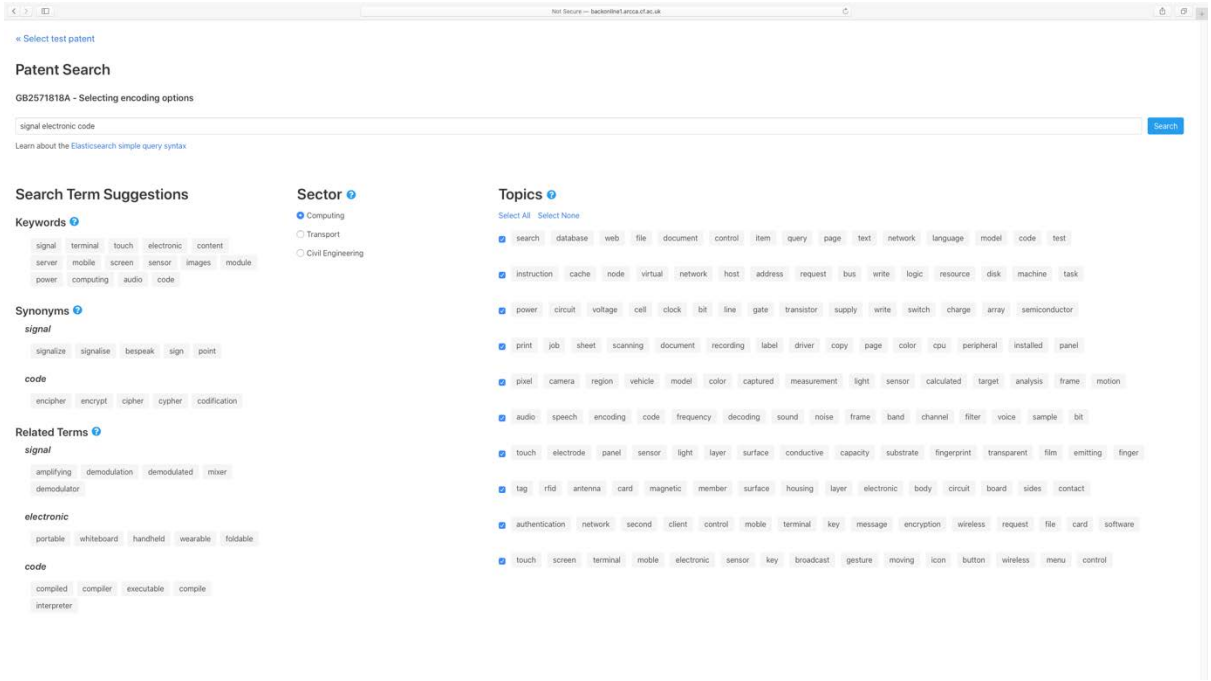
The following screenshot shows the drop-down menu that groups the 10 patents in each domain.



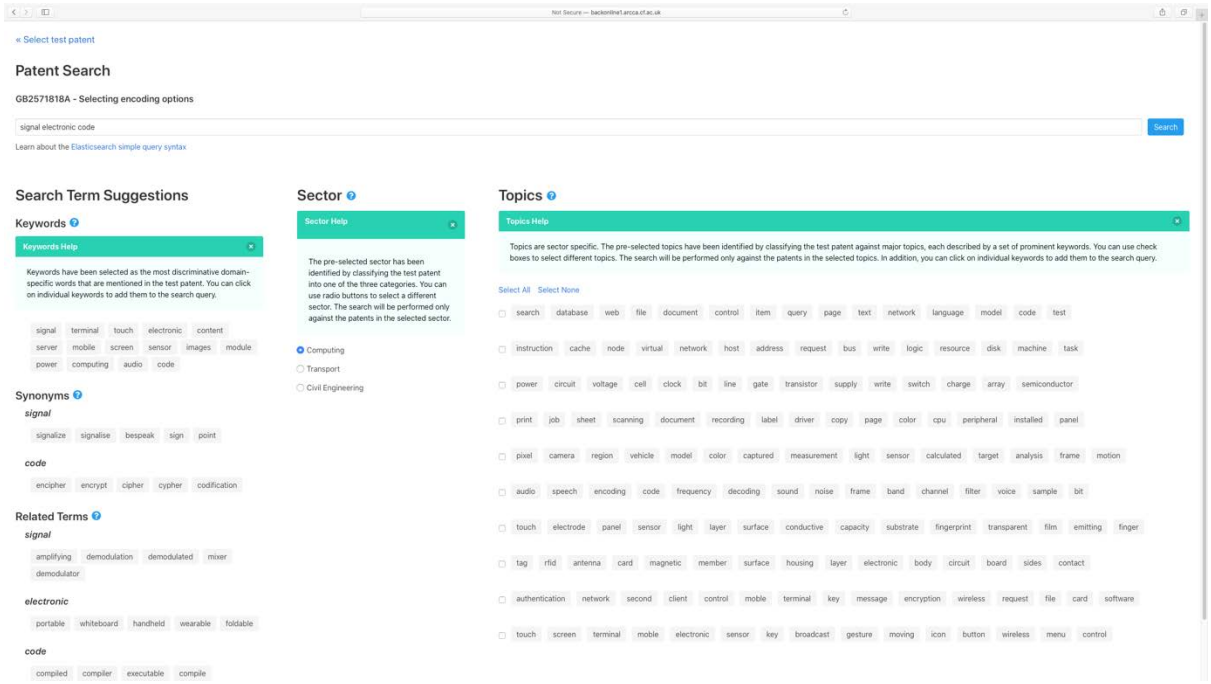
After selecting a patent from the drop-down menu, the interface updates to show the abstract and description of the patent, as shown in the following screenshot.



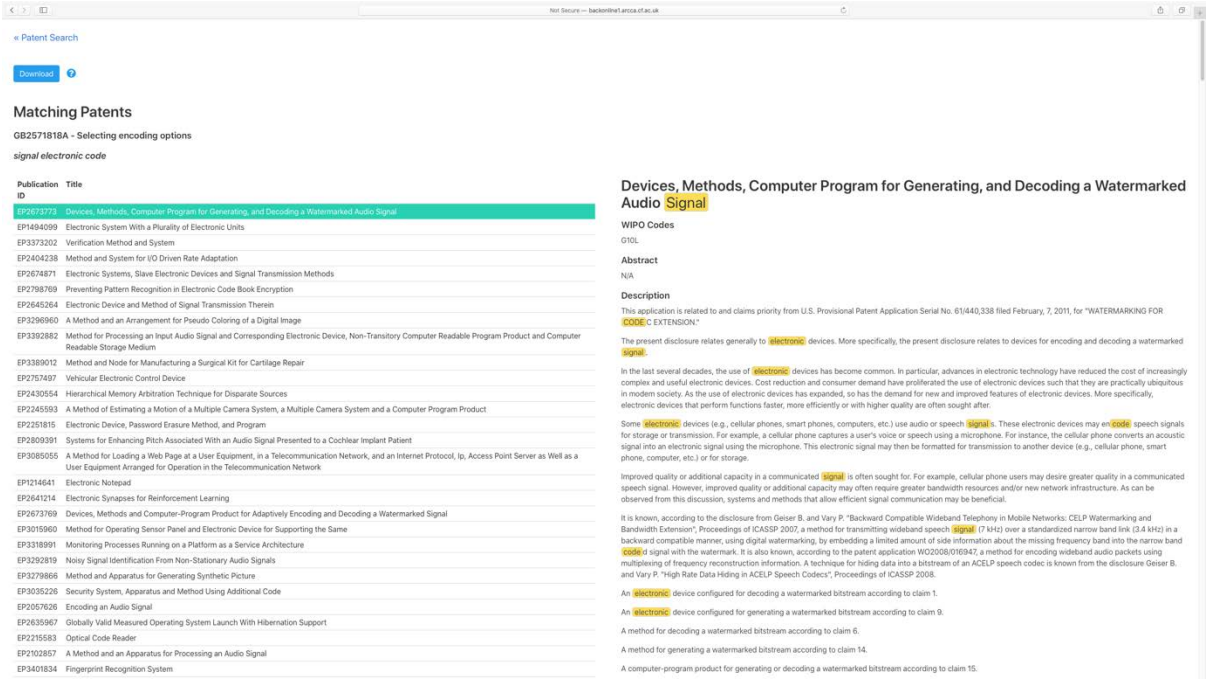
Clicking on the Show Prior Art button presents the patent searching interface shown in the following screenshot. The input box is initially populated with the top-ranking TF/IDF terms in the patent selected on the previous page. Examiners can edit the search terms in the input box with their own terms or with terms chosen from the topic keywords and search terms suggested by the AI algorithms of the system. The Sector control enables the examiner to select search term suggestions and topic keywords from one of the three domains: civil engineering, computer technology or transport. The initial domain is selected by the classifier described in step 1 of Table 4.



The patent search interface provides contextual pop-up help for each set of examiner-selectable data. The help for a data control is displayed by clicking on the question mark icon next to the title of the control and is displayed below the control's title with a green background.



When examiners have finished editing their search query, clicking the Search button next to the input box performs the search with Elasticsearch and presents the results, as shown in the following screenshot. The titles of the matching patents are listed on the left. Clicking on a title presents the patent's abstract, description and claims on the right.



Appendix 10: The results of cross-validation classification experiments

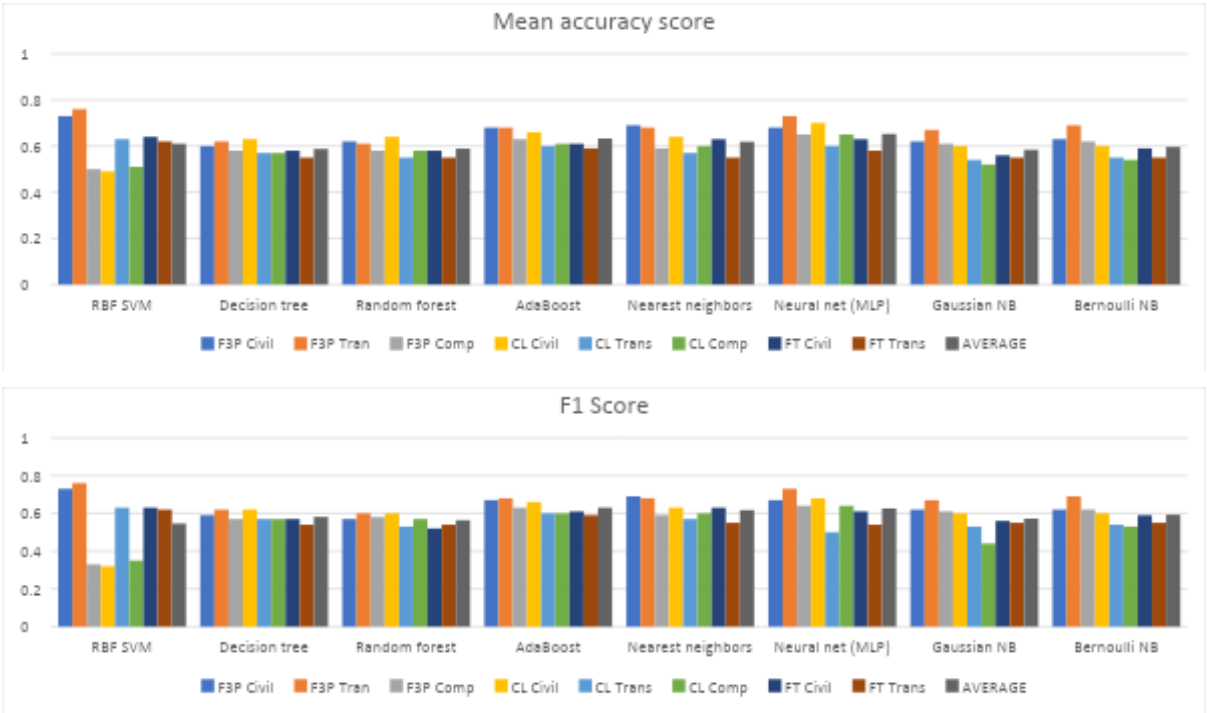


Figure 16: Results of cross-validation classification experiments (F3P = first three paragraphs; CL = claims; FT = full-text)

Concept House
Cardiff Road
Newport
NP10 8QQ

Tel: 0300 300 2000
Fax: 01633 817 777
Email: research@ipo.gov.uk
Web: www.gov.uk/ipo

Facebook: TheIPO.UK
Twitter: @The_IPO
YouTube: ipogovuk
LinkedIn: uk-ipo

For copies in alternative formats please contact our Information Centre.

When you no longer need this booklet, please recycle it.

Published: April 2020