Antennas and Electromagnetics Research via Natural Language Processing

Young-ok Cha

A thesis submitted to Queen Mary University of London in partial fulfilment of the requirements for the degree of

Doctor of Philosophy

School of Electronic Engineering and Computer Science
Queen Mary University of London
London, United Kingdom

April 2023

Declaration

I, Young-ok Cha, confirm that the research included within this thesis is my own work or that where it has been carried out in collaboration with, or supported by others, that this is duly acknowledged below, and my contribution indicated. Previously published material is also acknowledged below.

I attest that I have exercised reasonable care to ensure that the work is original and does not to the best of my knowledge break any UK law, infringe any third party's copyright or other Intellectual Property Right, or contain any confidential material.

I accept that the College has the right to use plagiarism detection software to check the electronic version of the thesis.

I confirm that this thesis has not been previously submitted for the award of a degree by this or any other university. The copyright of this thesis rests with the author and no quotation from it or information derived from it may be published without the prior written consent of the author.

Young-ok Cha
April 2023

Acknowledgements

During my PhD journey, I am deeply grateful for the opportunity to have met such exceptional people at Queen Mary who have inspired and enriched my personal and academic growth.

First and foremost, I wish to express my deepest gratitude to my primary supervisor, Prof. Yang Hao, for providing me with unwavering support and exceptional guidance during my research period. His extensive expertise, insights, and constructive feedback have been keeping me motivated throughout the challenges I have faced while exploring my novel research topic. Thanks to his guidance, I have been able to overcome limitations and reach my full potential as a researcher. It is an immense privilege to have had the opportunity to work under the supervision of such a highly respected scholar.

I would also like to thank my second supervisor, Prof. Massimo Poesio for his support in reading my stage reports and providing with valuable comments and suggestions on my work. I am utterly grateful to my independent assessor, Dr. James Kelly for his support in reviewing my work and sharing his fabulous insights with me.

I sincerely thank my colleagues including Dr. Henry Giddens, Dr. Ahsan Noor Khan, Dr. Achintha Avin Ihalage, Dr. Yihan Ma, Dr. Hanchi Ruan, Ms. Parvathy Chittur, Mr. Yibing Guo, Mr. Orestis Christogeorgos, Mr. Jonas Kolb and Ms. Mojan Omidvar from our research group for helping me in various aspects and sharing their expertise with me.

Last but not least, I want to express my gratitude to my husband, Paul and son, Inwoo for supporting me and being by my side throughout the years. I want to thank my parents and sister, Suyeon for always encouraging and inspiring me to give my all. Without the unfailing love of my family, achieving my goals would have undoubtedly been more challenging.

Abstract

Advanced techniques for performing natural language processing (NLP) are being utilised to devise a pioneering methodology for collecting and analysing data derived from scientific literature. Despite significant advancements in automated database generation and analysis within the domains of material chemistry and physics, the implementation of NLP techniques in the realms of metamaterial discovery, antenna design, and wireless communications remains at its early stages.

This thesis proposes several novel approaches to advance research in material science. Firstly, an NLP method has been developed to automatically extract keywords from large-scale unstructured texts in the area of metamaterial research. This enables the uncovering of trends and relationships between keywords, facilitating the establishment of future research directions. Additionally, a trained neural network model based on the encoder-decoder Long Short-Term Memory (LSTM) architecture has been developed to predict future research directions and provide insights into the influence of metamaterials research. This model lays the groundwork for developing a research roadmap of metamaterials. Furthermore, a novel weighting system has been designed to evaluate article attributes in antenna and propagation research, enabling more accurate assessments of impact of each scientific publication. This approach goes beyond conventional numeric metrics to produce more meaningful predictions.

Secondly, a framework has been proposed to leverage text summarisation, one of the primary NLP tasks, to enhance the quality of scientific reviews. It has been applied to review recent development of antennas and propagation for body-centric wireless communications,

and the validation has been made available for comparison with well-referenced datasets for text summarisation.

Lastly, the effectiveness of automated database building in the domain of tunable materials and their properties has been presented. The collected database will use as an input for training a surrogate machine learning model in an iterative active learning cycle. This model will be utilised to facilitate high-throughput material processing, with the ultimate goal of discovering novel materials exhibiting high tunability. The approaches proposed in this thesis will help to accelerate the discovery of new materials and enhance their applications in antennas, which has the potential to transform electromagnetic material research.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background and Motivation

Artificial intelligence (AI) has made remarkable achievements both theoretically and practically. Many researchers in material science have started to recognise the importance of AI techniques in discovering unknown properties of existing or new materials. Utilising such techniques not only saves time and resources but also eliminates the biased insight of human for obtaining the results effectively. The success of these AI-based methods however lies on the quality of input data. It is well known that the acquisition of accurate input data in material science is of a paramount importance in both experimental and computational methods. Besides, to learn effective models using AI techniques, a large amount of input data is required. To solve various problems in material science using AI tools, we need to build

reliable and sufficient database that contains greater variety with increasing volumes and velocity (*i.e.,* big data).

As a branch of AI, natural language processing (NLP) has demonstrated notable achievements in diverse areas of computational science such as machine translation, question answering, sentiment analysis and information retrieval [1],[2]. The recent developments in NLP have made a variety of tools accessible for scientists for high-quality information extraction from massive amount of publicly available unstructured data (*e.g.* texts) [3]. At the same time, publishers of scientific journals have digitised their collections and resources to a computer-readable html/xml format while providing APIs for developers and other users [4].

The motivation for this thesis revolves around harnessing the capabilities of contemporary AI and NLP techniques to enhance the efficiency of the material discovery process. Our objective is to streamline this process, minimising the consumption of time and resources, while also removing the influence of human biases, leading to a fully automated system. Material science is anchored on the foundation of robust databases; these databases play a pivotal role in enabling new material discoveries. The advantages of a well-constructed database in this domain span across several critical aspects. It facilitates faster research outcomes, empowers informed decision-making, ensures the possibility of replication and verification, and most crucially, stands as a foundation work for the success of AI-driven approaches to problem-solving.

In recent times, the strides made in NLP across various computational and scientific disciplines have been commendable. There is a growing realisation of NLP's ability to mine quality information from vast amount of unstructured data. The peer-reviewed scientific literatures are one of the most reliable data sources, however, their unstructured and heterogeneous nature poses a significant obstacle to large-scale analysis of the information contained within [5]. To encapsulate, this thesis is poised to address the complexities

associated with automated document analysis and information extraction. The ultimate goal is to construct a database that simplifies the discovery of new materials in material science by leveraging cutting-edge NLP techniques.

The traditional approach to the new material discovery process, particularly in research forecasting, predominantly relies on human expertise [71]. However, several inherent limitations plague this method. To begin with, there is the issue of subjectivity. Human experts, being products of their own unique experiences, biases, and perspectives, can inadvertently skew their predictions. These predictions may not always align with broader trends or untapped possibilities. Moreover, humans tend to adopt a linear mindset, which can fall short in accurately predicting the often exponential or non-linear progression of technological advancements, potentially resulting in underestimations. Furthermore, human experts might be unduly influenced by current events or fresh breakthroughs, thus sidelining the actual prospective potential.

The existing efforts to automate database construction in the field of material science [43], [45], [46], [53], [55], [56] face several critical challenges that cannot be overlooked. One of the primary issues lies in the inherent ambiguity of chemical descriptions. When chemical information of materials is presented, it is often laden with nuances and complexities that may not be immediately apparent. In the absence of the right contextual cues, automated algorithms can easily misinterpret this information, leading to inaccuracies in the database. A specific example illustrating this challenge is the ChemDataExtractor (CDE) [42] , an automated tool designed to pull out specific entities, such as chemical names, from text. While CDE is proficient in recognising and extracting these singular entities, it grapples with understanding the intricate relationships that might exist between these identified elements. For instance, if two chemicals interact in a particular way under specific conditions, capturing this complexity is beyond the basic capabilities of the tool. Moreover,

the effectiveness of automated database construction algorithms, like that employed by CDE, is intrinsically tied to the quality and breadth of their training data and the list of predefined chemical entities they are programmed to recognise. This introduces another layer of limitation. If a scientific text mentions a new or uncommon chemical compound that is not part of the algorithm's training set or predefined list, the system may fail to recognise and document it.

In conclusion, the domain of comprehensive data-driven forecasting within scientific research largely remains an untapped avenue. Concurrently, although automated solutions for the construction of material databases have made commendable progress, they have substantial limitations. These constraints include the ability to accurately interpret ambiguous chemical descriptions, grasp intricate relationships between chemical entities of materials, and identify novel or less common compounds. Given these challenges, the focus of this thesis will be to employ cutting-edge NLP methodologies to address the complexities associated with automated document analysis and information extraction. The ultimate objective is to contribute to the development of a more robust and versatile database geared towards the discovery of new materials.

1.2 Objectives

The aim of this study is to design and develop a novel automated NLP framework that can read scientific papers from relevant scientific e-repositories, extract features and analyse their contents, and learn models to generate new knowledge or information. The new data created in this study will be used for new material design and discoveries in collaboration with other colleagues in the research group. By pursuing this, the new methodologies proposed in this study are to be validated by the ground truth datasets extracted from the literature.

In striving to achieve the overarching aim of this study—to develop automated methods for forecasting scientific research trends in the field of metamaterials—this thesis is guided by several meticulously planned objectives. The first objective involves the creation of an information extraction tool built on the capabilities of NLP. This tool is designed to automate the process of gathering scientific literature, specifically targeting papers accessible online from recognised publishers such as Scopus. The automation serves to efficiently collate the data crucial for analysing trends in metamaterial research. Secondly, the research concentrates on amassing a robust keyword pool. This is attained by employing the Rapid Automatic Keyword Extraction (RAKE) algorithm and by mining author-specified keywords from the papers that have been collected. The union of these keywords into a singular pool establishes a strong foundation for subsequent stages of data analysis. The third objective of the study pivots towards visual analytics. Here, each keyword from the assembled pool is transformed into vector format. These vectors are then subjected to a variety of clustering algorithms and trend analysis techniques. This visual exploration not only clarifies the existing research themes but also elucidates their interrelations, leading to a nuanced understanding of the current state of the field. The fourth objective is squarely aimed at forecasting. Utilising an encode-decoder Long Short-Term Memory (LSTM) model, the research aims to predict future trends and focal points in metamaterial research. This model is trained using the historical keyword frequencies from the accumulated data, enabling it to project likely future research directions. Beyond these four core objectives, the thesis expands the scope of metamaterial research by extracting new features from raw data to enhance prediction accuracy. A novel metric is introduced to assess the scientific influence of each article, offering another layer of depth to the data analysis.

To further augment the comprehensiveness of scientific surveys, an NLP-based summarisation tool has been developed. This tool automatically pinpoints and extracts the

key concepts from relevant literature, simplifying the complex task of literature review. Lastly, the research also endeavours to contribute to the discovery of new phase change materials that are easy to fabricate and non-toxic. Modifications have been made to the ChemDataExtractor (CDE) tool, adding specific functions aimed at extracting data pertinent to phase change materials (PCMs). The model has been updated accordingly, aligning it more closely with the study's objectives.

1.3 Structure of Thesis

To show the utility and advantages of the proposed methodologies, they are adopted to several material research applications. The structure of this thesis is as follows:

- Chapter 2 summarises the recent developments in NLP, followed by a short introduction to contemporary machine translation, auto-summarisation, and information extraction techniques with their potential impact on the field.
- In Chapter 3, a novel metamaterial keywords prediction framework using NLP and machine learning (ML) and the novel methodologies on extracting metamaterial keywords and forecasting the future keywords trend are discussed.
- In Chapter 4, an NLP-assisted review system utilising the state-of-the-art auto-summarisation methods is proposed, validated and its performance is compared with human review.
- In Chapter 5, the proposed prediction framework in Chapter 3 is further developed utilising attention mechanism, new features and weighting method. The novel concepts used in this framework is then applied to antenna and propagation research in order to obtain more realistic prediction results.

- In Chapter 6, a new database has been developed for tunable materials, accompanied by a comprehensive introduction of contemporary methodologies and a detailed analysis of results derived from specific data sets.
- Chapter 7 provides a brief conclusion with suggestions for future research directions.

Chapter 2

Related Work

2.1 History of NLP

Many NLP researchers have endeavoured to make computers understand natural human language to aid the resolution of many real problems in language. Numeric structures added by NLP could also lead to practical applications such as text analytics and speech recognition. Since the first concept of NLP was introduced in the 1950s, scientists have attempted to teach the computer the meaning of words and complex grammar [6]. This however succeeded only partly. When a word has different meanings in different contexts, limitations were shown. For example, the word ‘bank’ has two different meanings, one related to rivers and the other related to money [7]. Without the context of the whole sentence, it was impossible to know the exact meaning of the word.

In the 1990s, researchers started to build a language model using massive language data rather than attempting to teach the computer vocabularies and the rules of human languages [8]. The language models can be defined as a probability distribution which helps predict the next word, phrase and sentence in the sequence of words. For example, applied to the sentence, "I am very happy because my boy passed the ____", the language model calculates the probability of which words are the most reasonable or adequate in the blank. These language models can be applied to create a form of question and answer. There are three main levels in NLP history, namely, syntactic, semantic, and pragmatic. As Erik *et al.* mentioned in Figure 1, this era saw the leap from the Syntactics Curve to the Semantics Curve [9]. A vast increase in computing power, the availability of large linguistic data, higher success rate of machine learning algorithms and a much richer understanding of human language enabled traditional NLP to push into the big data era and led to the notable development of NLP [9], [10]. In this era, NLP's ability to solve real-world problems began to be tested and successfully demonstrated state-of-the-art results in various tasks such as machine translation (MT), information extraction, text summarisation, semantic classification, Q&A, *so forth*.

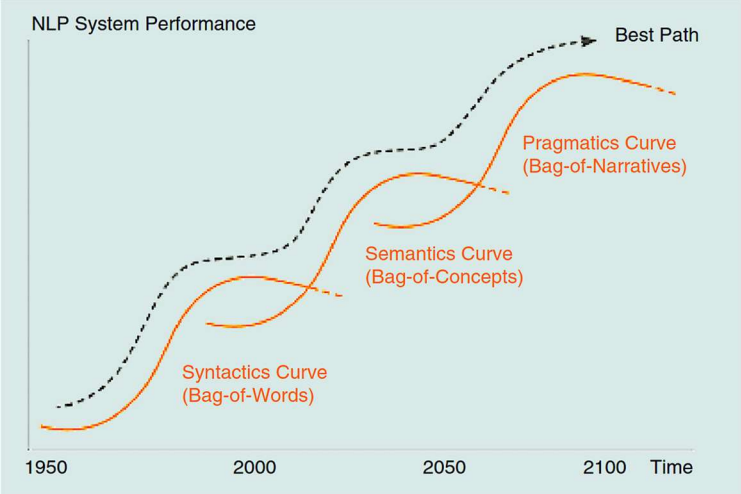


Figure 2.1 Envisioned evolution of NLP research through three different eras or curves [9].

Word embedding is the most important concept to understand semantic relations among words and thus it has been applied to a variety of NLP problems. It refers to the representation of words numerically which the machine can then comprehend by converting them into vector spaces. The words are assigned in vector spaces closely if they have a similar meaning and embeddings provide an analogy from the distance of two words. In 2013, Mikolov *et al.* [11] presented a pre-trained embedding model at Google, Word2vec improved the quality of word vectors and training speed. Skip-grams [11] and continuous bag of words [12] are two distinct objective functions for the training of models that were published by the original Word2vec authors. Each of these objective functions may be employed with a variety of different models. The Skip-gram objective takes a centre word (ω_c) as input and tries to reproduce the probability of context words as in 2.1.1,

$$\log P(\omega_{c-m}, \ldots \omega_{c-1}, \omega_{c+1}, \ldots \omega_{c+m} | \omega_c) = \log(\prod_{-m \le t \le m, t \neq 0}^{T} p(\omega_{c+t} \mid \omega_c)) \tag{2.1.1}$$

where m is the training contextual window's size. The basic Skip-gram formulation defines $P(\omega_{c+m}, \omega_c)$ using the SoftMax function in 2.1.2 [11],

$$P(\omega_O | \omega_I) = \frac{exp({v'_{\omega_O}}^{\top} v_{\omega_I})}{\sum_{\omega=1}^{W} exp({v'_{\omega}}^{\top} v_{\omega_I})} \tag{2.1.2}$$

where v_ω and v'_ω are the input and output vector representations of ω, and W is the number of words in the vocabulary. The cross-entropy metric is typically used to calculate how far apart the probabilities produced by the model, $\hat{y}$ and the probabilities in the real-world, y are from one another.

$$D(\hat{y}, y) = -\sum_{\omega=1}^{W} (y_\omega \log \hat{y}_\omega) \tag{2.1.3}$$

By reducing cross entropy, the model is trained, typically using stochastic gradient descent. Although the one layer with SoftMax and single SoftMax models are frequently introduced for educational purposes, they are rarely used in practice. Because of the sum in the denominator and the enormous vocabulary size W, computing the softmax function is computationally intensive. The researchers instead suggested two effective methods: hierarchical SoftMax and negative sampling [13]. The hierarchical SoftMax uses a binary tree to represent an output layer with W words as its leaves. Each node explicitly depicts the relative probabilities of its child nodes, then a suitable route from the tree's root to each word, ω, is determined. Let $n(\omega, m)$ be the m-th node on the path from the root to ω, and let $L(\omega)$be the length of this path, so $n(\omega, 1) = root$ and $n(\omega, L(\omega)) = \omega$. In addition, for any inner node n, let $ch(n)$ be an arbitrary fixed child of n and let $[x]$ be 1 if x is true and -1 otherwise. The hierarchical softmax then defines $P(\omega_O|\omega_I)$ as follows [11]:

$$p(\omega|\omega_I) = \prod_{m=1}^{L(\omega)-1} \sigma([n(\omega, m+1) = ch(n(\omega, m))] \cdot v_{n(\omega,m)}{}^{\top} v_{\omega_I}) \qquad (2.1.4)$$

where $\sigma(x) = 1/(1 + exp(-x))$. It can be verified that $\sum_{\omega=1}^{W} p(\omega_O|\omega_I) = 1$. This indicates that the cost of computing $\log p(\omega_O|\omega_I)$ and $\nabla \log p(\omega_O|\omega_I)$ is proportional to $L(\omega_O)$, which on average is not larger than $\log W$. Noise Contrastive Estimation (NCE), which was developed by Gutmann and Hyvarinen [14], is a substitute for the hierarchical SoftMax. The Skip-gram model only cares about learning outstanding vector representations. Therefore, even though it can be shown that NCE roughly maximises the log likelihood of the SoftMax, NCE can be optimised as long as the vector representations continue to be superior. Negative sampling (NEG) by the objective can be defined as:

$$\log \sigma({v'_{\omega_O}}^{\top} v_{\omega_I}) + \sum_{i=1}^{k} \mathbb{E}\omega_i \sim \mathrm{P}_n(\omega) \left[\log \sigma(-{v'_{\omega_i}}^{\top} v_{\omega_I})\right] \quad (2.1.5)$$

The objective is used to replace every $\log p(\omega_O | \omega_I)$ term in the Skip-gram objective. The task is thus to distinguish the target word ω_O from draws from the noise distribution $\mathrm{P}_n(\omega)$ using logistic regression, where there are specific number of negative samples for each data sample.

Figure 2.2 shows the results of using 2-dimensional PCA projection of the 1000-dimensional vectors of countries and their capital cities using the embedding model, Word2vec. As seen in this figure, the model is able to learn the relationships between the capital and the country without supervised information and arrange the coordinates of each word automatically. Other embedding models, Glove [15] and fastText[1] also developed by the scientists at Stanford University and Facebook AI, respectively.

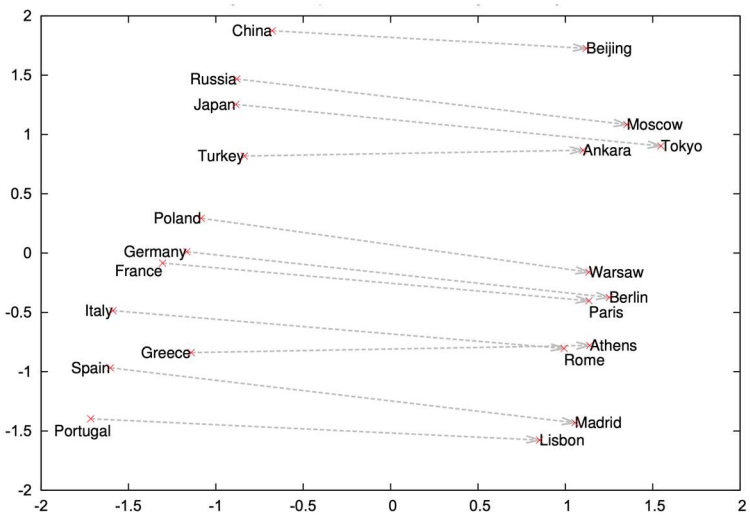


Figure 2.2 Country and capital vectors projected by PCA [11].

[1]https://github.com/facebookresearch/fastText

Although the model is trained and has captured the contextual meaning of the words, to understand a sentence, the word sequence is also crucial. To address this, Recurrent Neural Network (RNN) [16] and Sequence to Sequence (Seq2Seq) [17] models were proposed. These generate a context vector which is a fixed-length vector representation; it compresses all information of the sentence, not just the meaning of the words as well as the sequence of the words. These models achieved excellent performance particularly on the MT task compared to the traditional grammar-based language model.

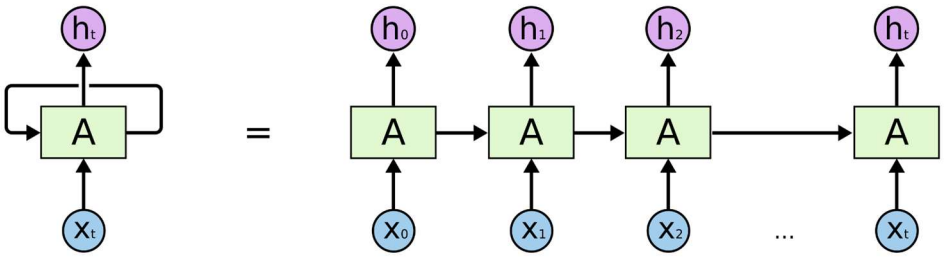


Figure 2.3 An unrolled recurrent neural network [18].

As shown in the architecture of RNN (Figure 2.3), for any input $X = (x_0, x_1, x_2, \ldots x_t)$ with a variable number of features, at each time-step, an RNN cell takes an item x_t as input and produces an output h_t while passing some information onto the next time-step. That means input data cannot be processed simultaneously and causes a huge amount of processing time as the input increases.

Another disadvantage of RNNs is the long-term dependency. Based on the RNN structure, where input data is entered sequentially, the context vector which is entering the decoder is more affected by data later in the sequence. When it has a long input sequence, the influence of the words at the earlier time steps of the sequence almost disappears. This incapability to pass on past information until the end (later time steps of the sequence) is known as a long-term dependency problem. To address this problem, LSTM (Long-term

Short Memory) [17] and GRU (Gated Recurrent Unit) [19], modified versions of Vanilla RNNs, have been proposed. Although this variant model has reduced the effect of some of the issues, the sequential structural problems have still remained. A new simple network architecture, the Transformer, based solely on the attention mechanism, was thus proposed by the team at Google [20]. The mechanism of the Transformer replaces the recurrent layers with multi-headed self-attention which contains information queries, keys and values, and it is able to perform the attention function in parallel.

Here, the attention mechanism is a part of the neural network architecture that allows the decoder to utilise the most relevant features of input data (*i.e.* sequence of words). Figure 2.4 visualises the multi-head self-attention mechanism of the encoder. The correlation between different words in the sentence is calculated and dark colours indicate more relevant words. The phrase "making... more difficult" is completed by many of the attention heads focusing on a distant dependency of the verb "making". Only the word "making" is given attention in this sentence. Different colours represent different heads.

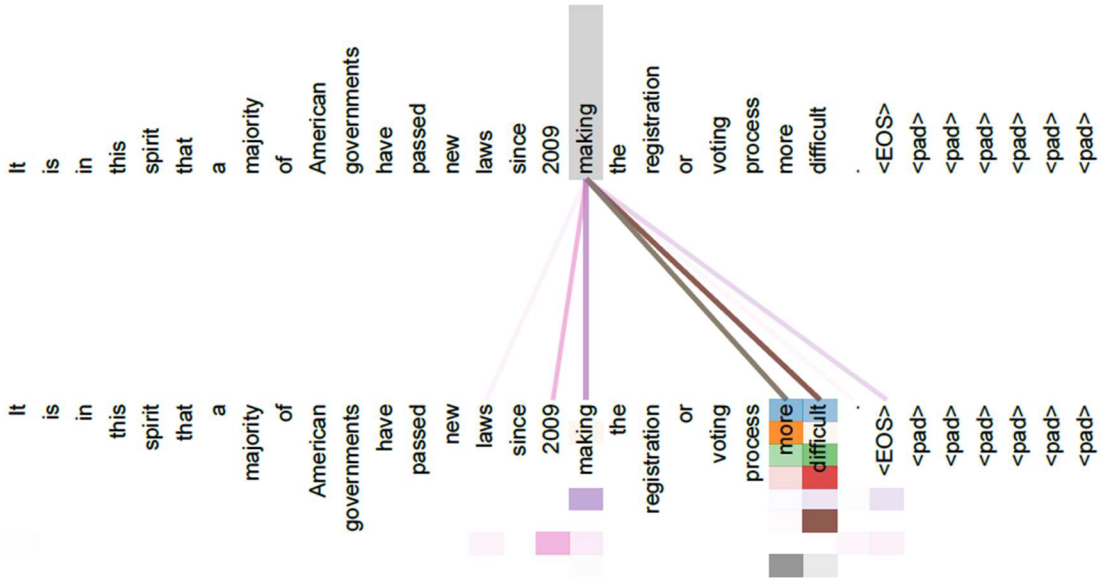


Figure 2.4 An example of attention mechanism following long-distance dependencies [20].

When human translates a sentence, they do not translate the sentence with the same weight on all words. The human translator would focus on the word which needs to be highlighted and this improves the accuracy of the translation. For example, in the sentence "I went to school.", human concatenates the meaning of each word and the positional data as "I :1, went : 2, to : 3, school : 4". Instead of putting them in order, all the words in the sentence can feed to the encoder as an input at once with the values for the order of the words. In addition, it is possible to have a context vector for each word in the sentence, so that no information loss occurs even if the sentence length increases. This improves the speed of training of the model as well as its performance while ensuring the scalability of sentence length. Both Google's Bidirectional Encoder Representations from Transformers (BERT) [21] and OpenAI's Generative Pre-Trained Transformer (GPT) [22] are high-performance and large-scale language models which utilise the Transformer technique. GPT-3 [23] which was introduced in 2020, is a language model which was pre-trained with a large scale of text data and over 175 billion hyperparameters. As the decoder in GPT-3 has already trained a huge number of sentences from the real-world data based on well-vectorised contents, it generates high quality texts which can be difficult to recognise whether it was written by a human.

Over the recent years, the emergence of large-scale language models has revolutionised NLP by achieving state-of-the-art performance on various language tasks such as question-answering and sentiment analysis. As depicted in Figure 2.5, number of parameters (*i.e.* the size of the model) has increased exponentially until the time the T-NLG is launched. Even GPT-2, the predecessor of GPT-3, which was announced by OpenAI in June 2020, used about 175 billion parameters. This uses at least 10 times more parameters than the T-NLG as in the graph, which shows how quickly the language models for NLP is scaling.

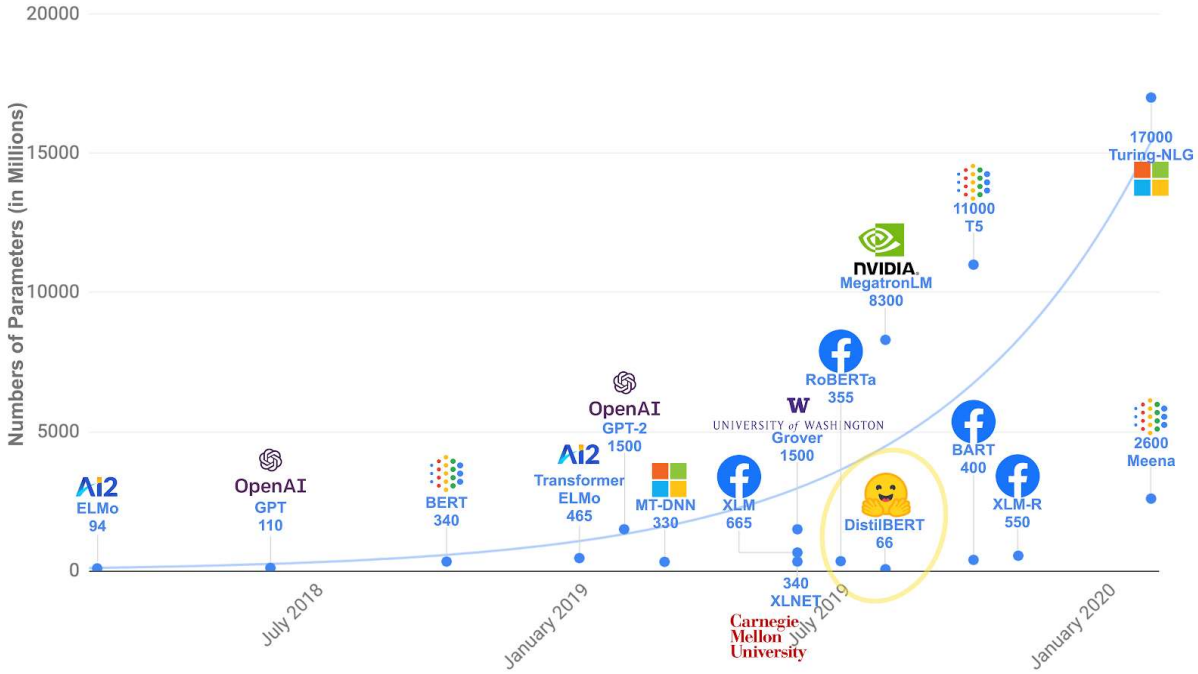


Figure 2.5 Exponential growth of number of parameters in deep learning models [24].

2.2 Machine Translation

MT was one of the first non-numeric applications of NLP as an aid tool in communication between humans [10]. In the early years of MT, building rules of translation rules such as grammar and domain knowledge databases was the most critical component in MT, however, as we entered the big data era, statistical approaches played key roles [25], [26]. This means that MT was no longer seen as a word-by-word translation problem but as a statistical and probabilistic problem. More recently, new possibilities of using a statistical model-free algorithm, neural network, in NLP have actively been discussed and this was named neural machine translation (NMT). The key benefit of NMT is that it no longer requires the pipelines (*e.g.* embeddings) used in statistical machine learning while allowing a single MT system to be trained from the source and target data directly [27]. Since then, the encoder-decoder architecture which can take a variable-length sequence as the input and transform it

into a state with a fixed shape was introduced. It was followed by the attention mechanism that enhances the encoder-decoder architecture by implementing the same action of selectively concentrating on a few relevant words in the input sequences [17], [28].

In the research of NMT, Cho *et al.* [28] developed the encoder-decoder using two recurrent neural networks (RNN) that can learn the sequences whose lengths are not known a-priori. The scores by the RNN Encoder–Decoder were found to be useful in improving the overall translation performance in terms of BLEU scores. The authors also captured the performance of the model using the word embedding of the learned phrase representation. The left graph of Figure 2.6 represents the full embedding space while the right graph shows a zoom-in view of one region. We can see that semantically similar words are clustered with each other.

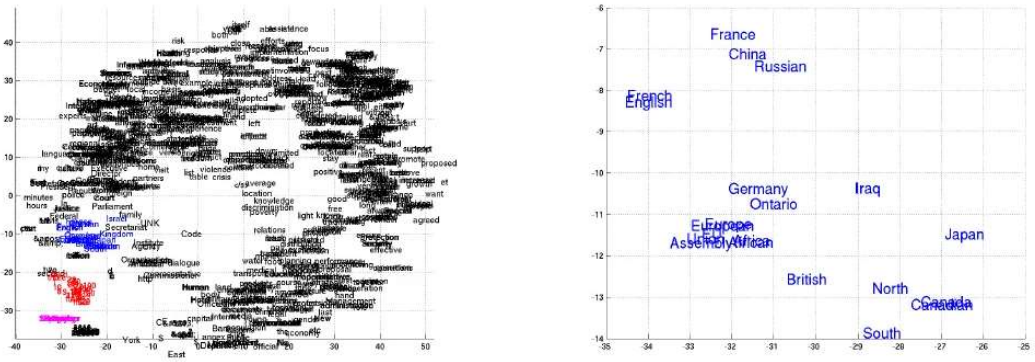


Figure 2.6 2D embedding of learned word representation [28].

As discussed, RNNs have the problem of long-term dependency, where they cannot learn effectively when the target word is far away. LSTM (Long Short-Term Memory) emerged to solve the long-term dependency problem. To address this issue, the structure of the RNN had to be modified, thus, it was LSTM that added a long-term memory with three gates.

In its basic form, RNNs had only one input path h, however, in LSTM, a new path called C (long-term memory) was added. Sutskever *et al.* [17] proposed the encoder-decoder model using a multi-layered LSTM to map the input sequence to a vector of a fixed dimensionality, plus an additional deep LSTM to decode the target sequence from the vector. The core concepts of LSTM units are the cell state and its various gates as depicted in Figure 2.7.

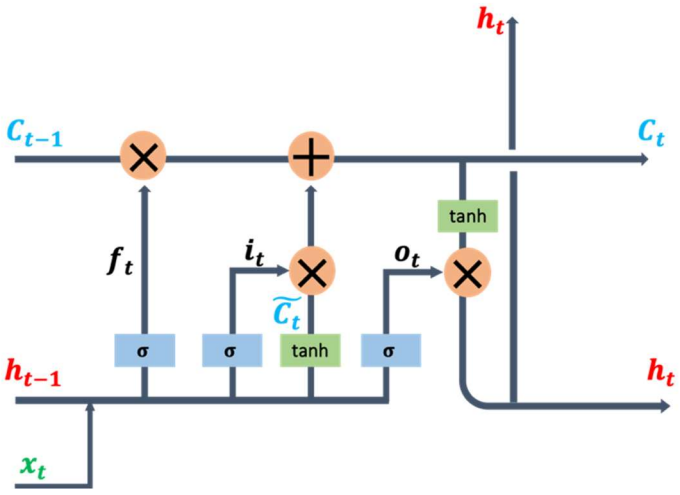


Figure 2.7 LSTM cell architecture.

The cell state regulates the amount of information the network remembers over time. It updates the old cell state, C_{t-1}, into the new cell state C_t. The compact form of the equation for an LSTM cell state is in (2.2.1).

$$C_t = f_t \cdot C_{t-1} + i_t \cdot \tilde{C}_t \tag{2.2.1}$$

The forget gates decides what is relevant to keep from prior steps using the equation (2.2.2).

$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right) \tag{2.2.2}$$

This gate takes in two inputs, h_{t-1} *and* x_t. h_{t-1} is the hidden state (*a. k. a.* short-term memory) from the previous cell or the output of the previous cell and x_t is the input at that particular time step. The given inputs are multiplied by the weight matrices (W_f) and a bias (b_f) is added. Following this, the sigmoid function is applied to this value. The input gate decides what information is relevant to add from the current step using the following equations (2.2.3), (2.2.4).

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad (2.2.3)$$

$$\tilde{C}_t = tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \quad (2.2.4)$$

The output gate determines what the next hidden state should be by deciding whether information of the Ct is visible or not using following equations (2.2.5), (2.2.6).

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad (2.2.5)$$

$$ht = o_t \cdot tanh(C_t) \quad (2.2.6)$$

As shown in Figure 2.8, this model translates a sentence "ABC" into a sentence "WXYZ" and stops reading an input after the EOS (end-of-sentence) token. Thanks to the benefit of LSTM, the model works well with a long sentence. The authors reported they surpassed the current performance on WMT' 14 English to French test set (nstst14).

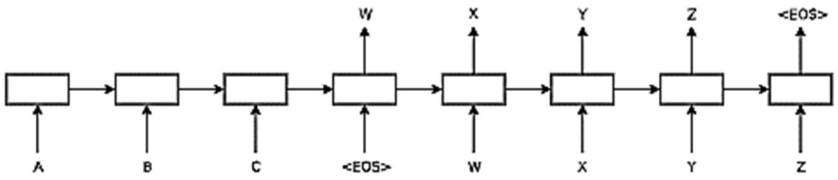


Figure 2.8 LSTM model for machine translation [17].

Although NMT systems, such as DeepL and Google Translate, have been widely used in people's daily lives, these are not error-free. Some recurring problems are frequently discussed in the literature however are still unsolved. The most well-known problem is gender bias due to unbalanced grammatical gender frequencies of words. For example, a nurse represented as a female whilst a surgeon is mentioned as a male [29], [30] Another issue is related to the context because these are not capable of translating whole texts as a single unit, but only isolated sentences. For example, the machine cannot consider the difference between technical and general terminology, it also has no knowledge of cultural conventions [29] Thus, to address these limitations, the researchers have adopted the pre-training language models like ELMo [31] and BERT [21] These can be used to extract features for the input data and are able to fine-tune for specific downstream tasks including question answering, text classification and information extraction etc. Figure 2.9 shows the overall pre-training and fine-tuning procedures of BERT. As you can see, the same architectures are used in both pre-training and fine-tuning without the output layer. Additionally, to initialise the fine-tuning model, the parameters from the pre-trained model are used. The final parameters then are learned during the training.

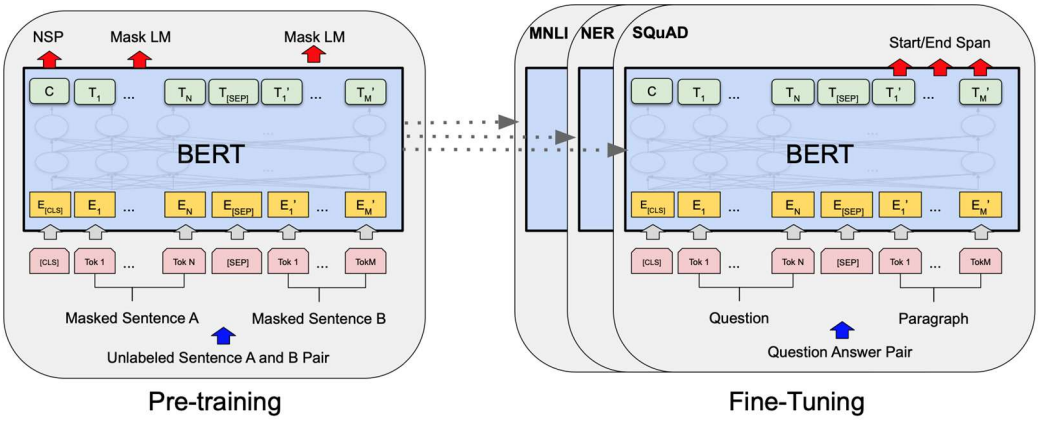


Figure 2.9 Overall pre-training and fine-tuning procedures for BERT [21].

BERT was introduced by the scientists at Google AI Language in 2018 [21] . It first applied the bidirectional learning mechanism of Transformer [20] to language modelling. For example, from the following two sentences, “I went to bank to deposit money” and “I went to the river bank”, to predict the meaning of the word “bank” by only taking either the left or the right context, then an error is unavoidable in at least one of the two given sentences. Previous approaches focused only on one-directional learning; however, bidirectional learning can have a deeper sense of language context and flow than one-directional language models. To pre-train the BERT for the bidirectional language model, the authors proposed two training strategies, Masked Language Model (MLM) and Next Sentence Prediction (NSP). MLM refers to before feeding word sequences into BERT, 15% of the words in each sequence are replaced with a [MASK] token. The model then predicts the original value of the masked words, based on the context provided by the other, non-masked, words in the sequence.

In MLM, the models are learned to understand the relationship between words in a sentence. Additionally, BERT is also trained on the task of NSP for an understanding of the relationship between sentences. BERT has already demonstrated great success in many ways. Zhu *et al.* [32] has verified the effectiveness of applying BERT to conventional NMT. Authors proposed a BERT-fused model which has two extra attention modules, the BERT encoder and decoder. An input sequence is first transformed into representations processed by BERT. Each NMT encoder layer then interacts with the representations obtained from BERT and eventually outputs fused representations leveraging both BERT and the NMT encoder. The decoder works similarly and fuses both BERT representations and NMT representations together. They conducted various experiments on supervised, semi-supervised and unsupervised machine translation. They claimed the achievement of state-of-the-art results on seven benchmark datasets.

MT is conventionally recognised for its capability to translate text between different languages. Yet, at its core, MT leverages the computational principle of sequence-to-sequence prediction. This fundamental approach has greatly influenced my methodologies in forecasting scientific research trends. Specifically, MT models, especially those anchored by LSTM or Transformer architectures, exhibit exceptional proficiency in managing long-range sequence dependencies. This ability is potentially adaptable to time series forecasting, which often grapples with intricate temporal dependencies characterised by seasonality, evolving trends, and possible non-stationarities. Moreover, the domain of machine translation leverages word embeddings, which are dense vector representations that encapsulate the semantic nuances of words. By adopting a similar approach for scientific research trend analysis, it becomes feasible to craft embeddings that pinpoint distinct categories and recurring patterns. This advancement empowers researchers with the tools to decipher, comprehend, and navigate the intricate temporal dynamics inherent to scientific trends.

2.3 Auto-summarisation

Automatic text summarisation has drawn attention as early as the 1950s. Luhn [33] used features such as word frequency and phrase frequency to extract important sentences from the text and made the automatic abstract from the high scoring sentences. While the previous researchers focused on the content words (*e.g.* keywords) only for producing the automatic abstract, Edmundson [34] in the late 1960s, added two more components; 1) the pragmatic words (*i.e.* cue words) which were used in the title and heading and 2) the structural indicators. Since then, many studies have been published to address the challenge of automatic text summarisation. Due to the exponentially increasing availability of digitised documents in recent years, text summarisation has been demanded and as a result, some of

today's NLP algorithms can produce a concise and fluent summary while preserving key information content and overall meaning. There are mainly two different types of text summarisation approaches, extractive and abstractive. In extraction-based summarisation, the contents are extracted from the original data, however, the extracted contents are not modified in any way. TextRank is the one of the most popular automated extractive summarisation methods. Although it was first introduced in 2004 [35], it is still widely used today. It is a graph-based ranking algorithm like Google's PageRank which has been successfully implemented in citation analysis [36]. Figure 2.10 represents the concept of PageRank which works by counting the number and quality of links to a webpage to determine a rough estimate of how important the website is. The reason why Page C has a higher PageRank than Page E, is that despite having fewer links, they are of much higher value.

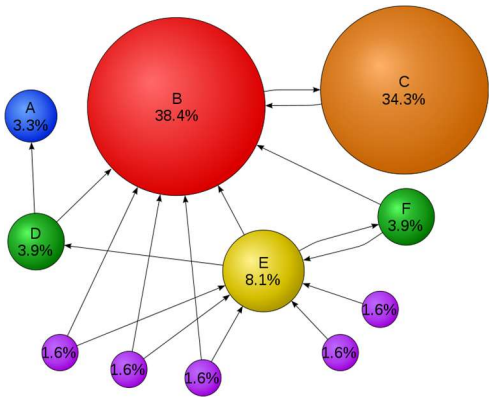


Figure 2.10 Mathematical PageRank (out of 100) for a simple network [2].

TextRank algorithm works on words and sentences instead of webpages and their links on PageRank. The algorithm first concatenates all the words or the sentences contained in the articles, and project them into a vector space. Similarities between sentence vectors are then calculated and stored in a matrix. The similarity matrix is converted into a graph, with

[2] https://commons.wikimedia.org/wiki/File:PageRanks-Example.svg#/media/File:PageRanks-Example.svg

sentences as vertices and similarity scores as edges for sentence rank calculation. Finally, a certain number of top-ranked sentences form the final summary. When the words or the sentences are converted to vectors, different methods like term frequency–inverse document frequency (TF-IDF), the word embedding (*e.g.* Word2vec), which is a deep learning neural network, or the attention-based transformers (*e.g.* BERT) could be used. Miller applied the BERT model to summarise lecture transcripts and provided a python-based RESTful service for students [37].

On the other hand, abstractive methods do not select sentences from the input texts but build an internal semantic representation of the original content and use this representation to produce a summary by interpreting the texts. Abstractive text summarisation is one of the most challenging tasks in NLP as it needs to address the issues of understanding long passages, information compression and language generation. Because of this, abstractive summarisation relies on a large-scale and well-trained language model. In 2020, Zhang *et al*. surpassed the state-of-the-art results using PEGASUS (Pre-training with Extracted Gap-sentences for Abstractive Summarisation) which is a pre-trained large Transformer-based encoder-decoder model on massive text corpora [38]. Figure 2.11 demonstrates how both Gap Sentences Generation (GSG), and Masked Language Mode (MLM) are applied to the same example. Initially, there are three input sentences. The first sentence is masked with [MASK1] and used as target generation text (GSG). The other two sentences remain in the input, but some tokens are randomly masked by [MASK2] (MLM).

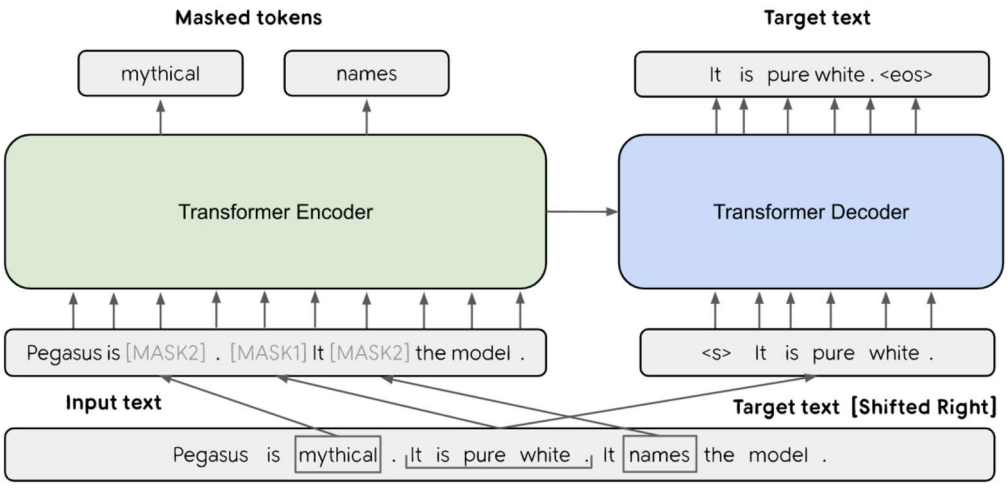


Figure 2.11 The architecture of PEGASUS [38].

The authors pre-trained the model on a sizable corpus of web-crawled texts using 5% of the parameters of the previous model. They then fine-tuned the model on 12 public down-stream abstractive summarisation datasets, producing new state-of-the-art outcomes as evaluated by automated metrics. The authors were surprised that although PEGASUS demonstrated an impressive performance with large datasets, it did not need many samples for fine-tuning to achieve the performance that was close to the previously known best results in the literature. This means that if there is a strong and reliable pre-trained language model with a massive volume of data, abstract summarisation is the one that has a promising future.

In the dynamic realm of scientific research, scholars produce an extensive array of publications daily. Given the sheer volume, it becomes an arduous task for any individual to comprehensively review each piece. Auto-summarisation, a principal task within NLP, has the capability to distil these voluminous texts into succinct summaries. This facilitates researchers in rapidly grasping the core insights of each paper, thereby highlighting avenues for deeper exploration and novel innovation. Recognising this advantage, we have endeavoured to refine our literature review methodologies in the material science domain — a field inundated with a plethora of papers. In Chapter 4, we elucidate how the auto-

summarisation tool can be a valuable asset for researchers, enabling them to extract pivotal concepts even from areas beyond their primary expertise and inclination.

2.4 Information Extraction

In the task of information extraction, structured information could be extracted from unstructured or semi-structured electronic documents. This is particularly useful for scientists as it could help reduce the amount of time and risk of error required to retrieve relevant information from the vast electronic resources. As a result, there are tremendous opportunities for large-scale automated data extraction to transform materials science into a more quantitative and data-rich field. Kononova *et al*. surveyed recent advances in automated text processing and information extraction from a large corpus of chemical, physical and materials science publications [39]. They presented that there are many advantages in using data to direct materials future research, which is motivated by things like new material discovery, property prediction, identifying synthesis methods, or figuring out device parameters. Here, data is essential to the materials informatics enterprise because they are required to use statistical methods to speed up the development process of new materials. Figure 2.12 depicts the trend of publications over the past 14 years. The top panel displays the total number of materials science publications, such as research articles, communications, letters, and conference proceedings, that are published each year. The bottom panel compares the proportion of scientific papers that are available online as image PDFs or embedded PDFs versus those that are in articles in HTML/XML format. The format of image PDFs vanished after the year 2000, and the format of embedded PDFs also decreased noticeably. The scientist thus can easily obtain the information from the more structured HTML/XML publications.

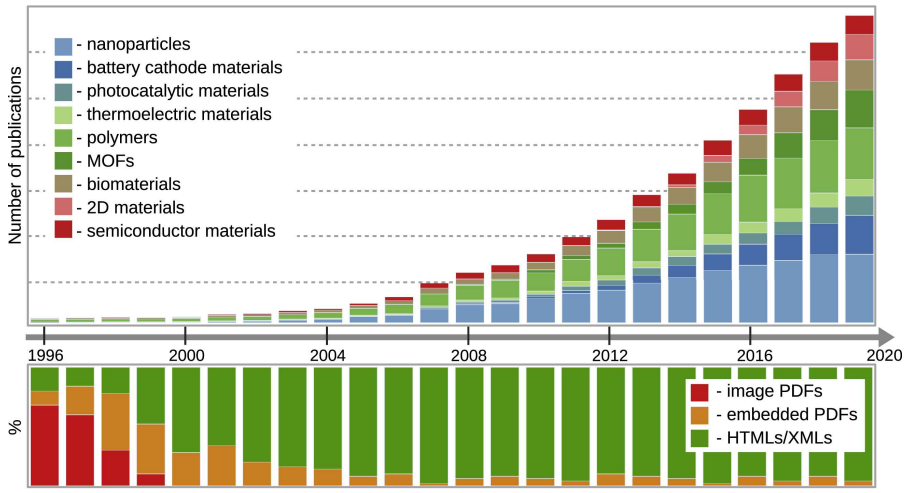


Figure 2.12 Publication trend over the past 14 years [39].

2.4.1 Automate Database Building

Recently, many researchers across a variety of fields have been exploring NLP techniques to build databases automatically from large-scale scientific publications. Tshitoyan *et al*. proposed an Elsevier and Springer APIs and abstracts related to inorganic material science [40]. They embedded 20,000 words into 200-dimensional vector spaces and these embeddings provided analogies such as Word2vec [11],[12]. Embedding is an important concept of NLP because the text cannot directly be fed into a statistical machine learning algorithm. Encoding performs this conversion from texts into a numeric form. For example, the colour of red, green, or blue could be converted into [1,0,0], [0,1,0] or [0,0,1], respectively. This encoding however is not able to represent the similarity between words and generates high dimensional vectors for all words [41]. Embedding is seen as a dense vector with similar information and Word2vec is one of the methods that can learn word embedding using neural networks. Since an embedding vector is learned with neighbour words, Word2vec can give similarity in vector representation.

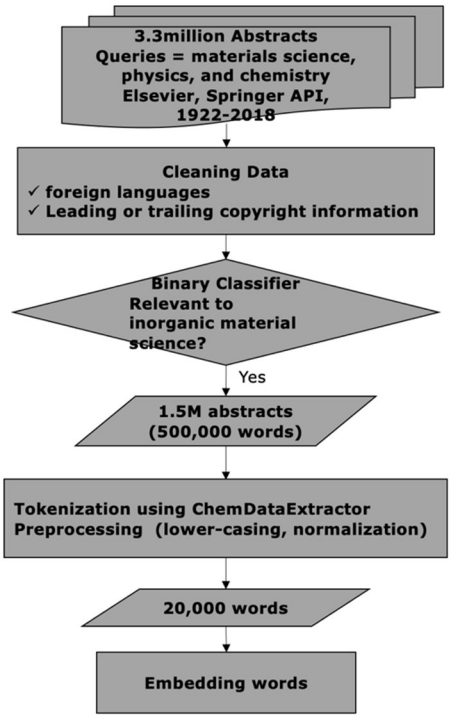


Figure 2.13 The flow chart of Tshitoyan *et al*'s work [40].

The material word embeddings discussed above enable the understanding of the relationships between words and can predict a new material using this relationship information. This embedding technique has been inspired by the idea of converting metamaterial keywords to vectors in this thesis. Figure 2.14 illustrates the outcomes of the prediction of new thermoelectric materials using these representations.

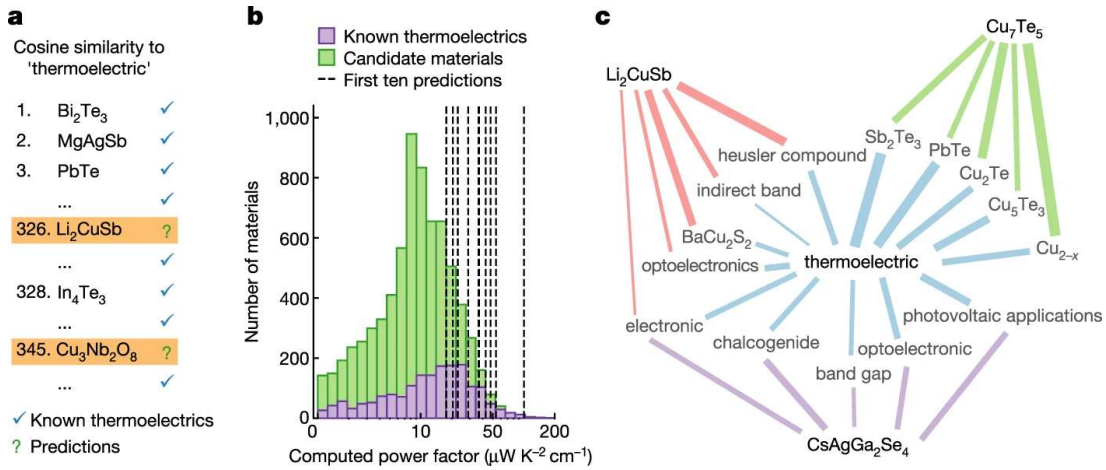


Figure 2.14 Prediction of new thermoelectric materials [40].

As shown in Figure 2.14(a), the cosine similarity of material embeddings with the embedding of the word ‘thermoelectric’ was used to create a ranking of thermoelectric materials. The authors predicted highly ranked materials that have not yet been investigated for thermoelectric applications. They also calculated distributions of the power factor for 1,820 known thermoelectrics from the literature and 7,663 candidate materials not yet studied as thermoelectrics. The first ten predictions’ power factors, which are represented by black dashed lines in Figure 2.14(b), are not studied as thermoelectric in their text corpus. The context words of materials predicted to be thermoelectric are displayed in Figure 2.14(c). They discussed the usefulness of the abstracts as the training corpus and how it led to the success of their unsupervised approach. Since abstracts are designed to give the key information in a clear and concise manner while avoiding unnecessary words, they were chosen as their training corpus. Even though their embedding model can capture delicate word relationships like negation, scientific abstracts frequently focus on positive connections. As a result, this study slightly reduces the negative effects of negation. The full texts are anticipated to have more negative relationships and, overall, more complex and variable sentences, necessitating the use of more robust techniques. This work inspired many scientists in the field as well as the study in this thesis to develop a new paradigm of machine-assisted achievements by enabling us to easily access the massive quantity of information in scientific publications.

In 2016, Swain and Cole developed a toolkit named ChemDataExtractor (CDE) that can build a chemical database from a large number of literatures automatically [42]. This work extracted useful and meaningful chemical information with their automated system using big scientific data. The pipeline that builds their database from the extracted sentences is depicted in Figure 2.15. As shown in Figure 2.15, the sentences are first separated from the texts. Each token is then separated. Each token is given a single tag that is coupled with the outputs of

the part-of-speech tagger and entity recogniser, which is processed by rule-based grammar to create a tree structure. This tree structure is then interpreted in order to extract specific chemical records for each clause. These extracted chemical records are then merged with records from other parts of the text to resolve data dependencies and provide unified records for storage in a database. The following tags are displayed: NN for noun, CD for cardinal number, VBZ for verb (present third person singular), DT for determiner, NNS for a plural noun, IN for preposition, JJ for adjective, CC for coordinating conjunction, and CM for chemical mention. Although this study has come closer to their objective of rapidly autogenerating chemical structure and property databases for materials science and other areas, accuracy-wise, individual natural language processing components still need to be improved.

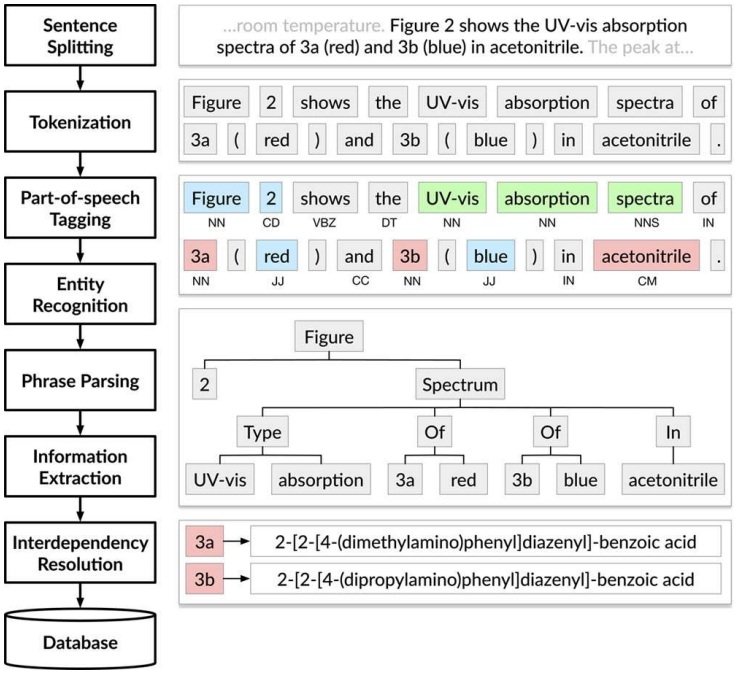


Figure 2.15 Natural language processing pipeline [42].

Using the CDE, Court and Cole automatically built a database of 39,822 records from a corpus of 68,078 journal articles containing chemical compounds and their associated Curie

and Néel magnetic phase transition temperatures [43]. The authors employed web-scraping software including the CDE 1.3 to collect all the relevant journal articles by submitting search queries for "Néel+temperature" and "Curie+temperature" to the Royal Society of Chemistry and Springer publishers' search pages. They expanded the scraper in this study to the Elsevier Science Direct websites using its Text and Data Mining API. Returning XML/HTML documents were converted into standard forms, and chemical relationships were extracted using rule-based phrase and table parsing in conjunction with probabilistic Named Entity Recognition (NER) techniques. As shown in Figure 2.16, the authors examined all database records containing $BiFeO_3$. The distribution of extracted phase transition temperatures can be developed and analysed to provide an estimate of the transition temperatures and related uncertainties. This approach achieved the average Néel (a) and Curie (b) temperatures of 644±9 K and 1,097±32 K, respectively. They claimed that these values were clearly in-line with the accepted transition temperatures for $BiFeO_3$ and they demonstrated the usefulness of the database for deriving reliable property values and associated uncertainties. However, their database still had certain limitations. For instance, the phrase and table parsing stages of the CDE are unable to discriminate between characteristics that have overlapping specifiers, and the modified Snowball algorithm stage can only extract relationships at the sentence-level.

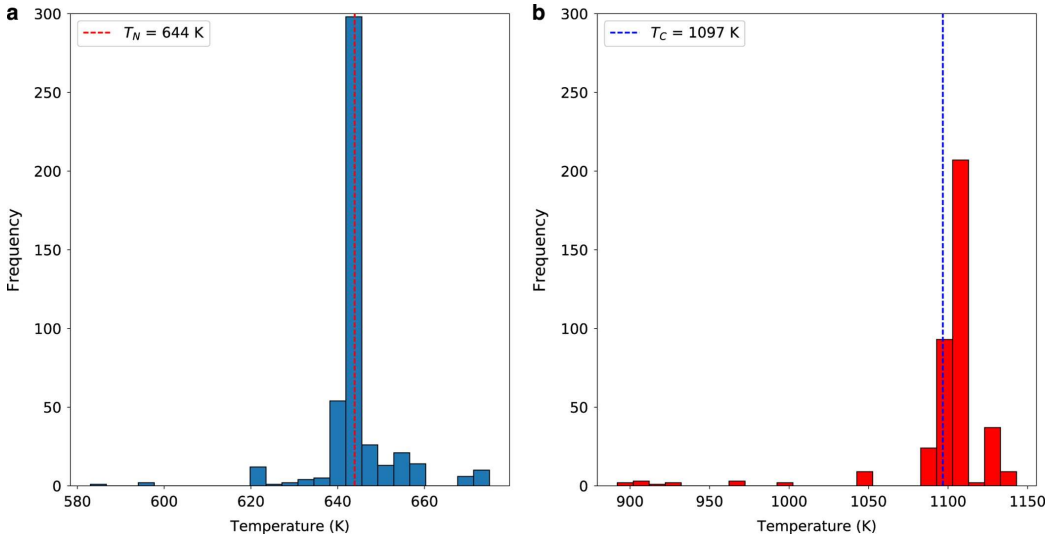


Figure 2.16 Example Néel and Curie phase transition temperature distributions for $BiFeO_3$ [43].

Although there were some limitations, CDE has been utilised and modified in ongoing approaches in the chemical and related fields. In 2020, Cole proposed a new approach that enables materials discovery based on data science with artificial intelligence which has dramatically reduced the average molecule-to-market lead time [44]. The author of this paper asserts that CDE can automatically create customised materials databases by extracting text from documents that are tailored to the chemical and property space of interest. This is a crucial factor in speeding up material discovery.

Recently, Huang and Cole automatically [45] built a database of battery materials that contains a total of 292,313 data records in an automated manner. They examined 229,061 pieces of academic literature using the modified CDE and offered Graphical User Interface (GUI) to make this database easier to utilise. The obtained data records could be employed to provide a meaningful analysis of the battery-related information contained in scientific papers. The Venn diagrams in Figure 2.17 illustrate how many chemicals share two different properties. This figure gives an indication of the degree of data correlation from massive, published works of literature, and the value of this overview is increased since it is based on scientific data rather than the opinions of human experts which may be biased.

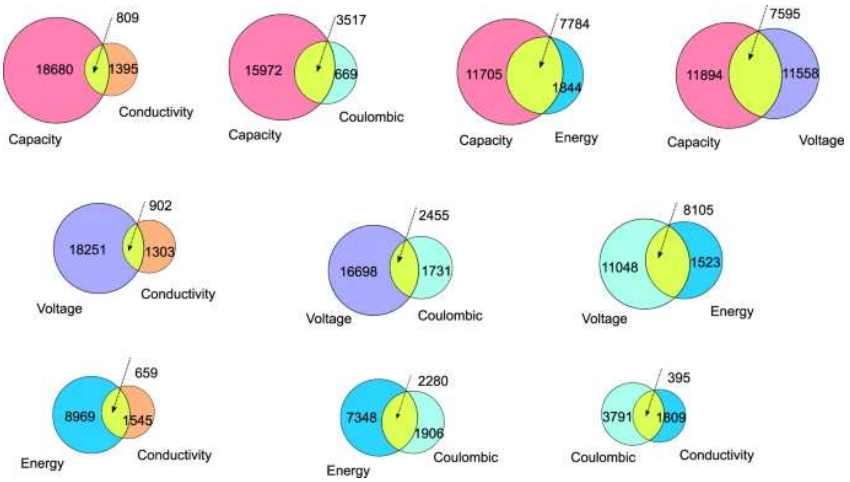


Figure 2.17 Venn diagrams of the data correlation between all possible pair-wise permutations between the five examined materials properties [45].

Another study to automate database building using the modified CDE was by Zhao and Cole, who extracted a total of 49,076 refractive indices and 60,804 dielectric constant data records on 11,054 unique chemicals from a corpus of 388,461 scientific papers automatically [46]. The employment of a new table processor from CDE version 2.0 [47] is an intriguing aspect of this study. An overview of the entire relationship extraction pipeline for the CDE used in this study is shown in Figure 2.18.

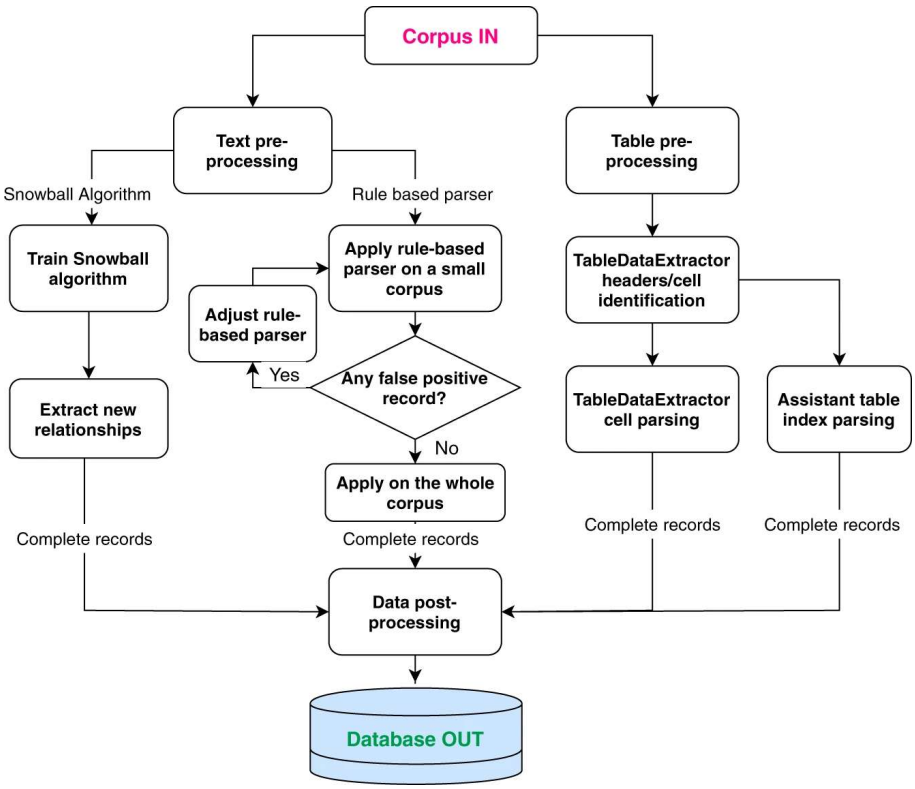


Figure 2.18 Pipeline of relationship extraction of ChemDataExtractor [46].

Table mining is one of the most challenging tasks since many tables have extremely complex structures or formats, and a materials domain's requirements for table style vary widely. In the study [46], a new table processor called TableDataExtractor was applied, and the authors stated that the database' overall precision and overall recall metrics were 77.22% and 74.48%, respectively. As seen in Figure 2.18, they used three different types of parsers, including logic rule-based data extraction from text, data extraction using the Snowball algorithm and TableDataExtractor. They pointed out that since it was the first study to use all three paring techniques, the precision is a little bit lower than the ones in earlier studies. Besides that, a trade-off between precision and recall can be seen, which is consistent with the fact that the quantity of data extracted should be proportional to the degree of loosening in the parsing logic. Tighter, more intricate parsing logic will commonly result in a smaller but more accurate database.

2.4.2 Automate Document Analysis

NLP techniques are also used in a variety of research fields to assist researchers in problem-solving and to increase the accuracy of their work. Zdravevski *et al*. analysed the new technical trends in enhanced living environments using the NLP toolkit [48] over the last 10 years [49]. Their work demonstrated that the NLP toolkit could be useful in expediting the fully automated review process while providing valuable insights from the surveying of relevant articles. In addition, their review generated by the NLP toolkit included informative tables, charts and graphs as shown in Figure 2.19 and 2.20. The majority of the review process was automated by using an NLP program to analyse papers from the IEEE Xplore, PubMed, and Springer digital libraries. In accordance with the PRISMA surveying methodology [50], they attempted to prove the NLP program's applicability by examining articles about Ambient Assisted Living and Enhanced Living Environments. The relevant articles were evaluated to find those with up to 20 attributes grouped into 4 logical categories. The analysis revealed a rise in interest from the scientific community in enhanced and assisted living settings over the previous ten years, as well as many trends in the research areas related to this scope. Figure 2.19 illustrates the relationship between several properties in terms of how frequently they appear together in a single article. This graph may be used to indicate which articles should be analysed with more depth and need to be thoroughly examined. The more articles with the associated keywords are present, the darker the edge is. It also demonstrates rare connections of some properties like 'cloud' and 'supervision' with others. The distribution of articles by keyword for each year is displayed in Figure 2.20. Notably, the number of papers for 'assistive technologies', 'smart environment' and 'enhanced living environment' is increasing over the years, while for others it is relatively small. This study is confined to present the past and current analysis only, not the future trend.

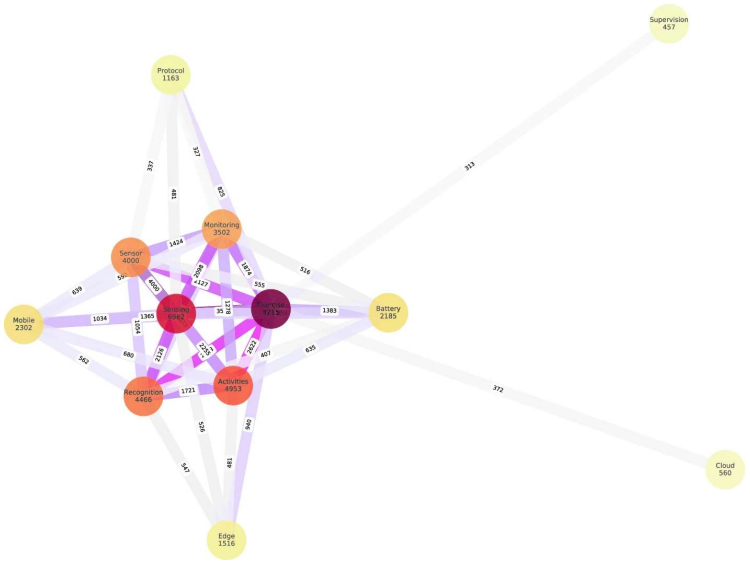


Figure 2.19 Graph visualisation with circular layout relevant articles by properties [49].

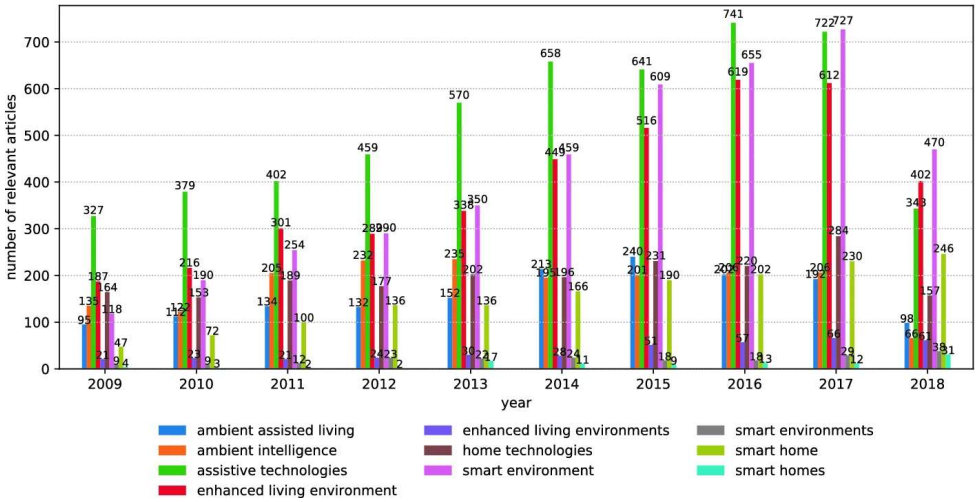


Figure 2.20 Number of relevant articles for each keyword from each source [49].

In 2020, Krenn and Zeilinger demonstrated a semantic network for predicting research trends of quantum physics from 750,000 scientific papers and knowledge from books and Wikipedia and confirmed high-quality predictions using historic data [51]. In this study, the abstracts of 100,000 arXiv articles in quantum physics categories generated lists of quantum physics concepts using one of the NLP tools named, Rapid Automatic Keyword Extraction (RAKE) [52]. RAKE is a powerful keyword extraction method which is based on keyword frequencies and word co-occurrences. The RAKE algorithm splits the text into a list of

words, removing the stopwords (*e.g.* 'the', 'is', 'in', 'where', 'at' etc.). Using this list, the algorithm creates a matrix of word frequencies and word co-occurrences, the words are then given a score. This score can be seen as the degree of a word in the matrix, as the word frequency, or as the degree of the word divided by its frequency. A word or word phrase is chosen as a keyword if its score belongs to the top T scores where T is the number of keywords to extract. This automatic keyword extraction method has been adopted for the metamaterial keyword extraction in this thesis. This study also defines emerging fields as those concepts or concept pairs that have significantly expanded over a five-year period after being introduced or first connected. As a result, Figure 2.21(A) displays the quantum physics topics that have seen the fastest increases in the number of papers containing them since their introduction between 1987 and 2017. The fastest-growing concepts from a five-year period, here are shown to not have been discussed prior to that time. Figure 2.21(B) clearly demonstrates newly connected pairs of ideas that became very influential in the scientific community over the five year period. The metamaterial keyword forecasting and antenna technology research in this thesis were also motivated by the emerging concepts and connections between the scientific discoveries in these figures.

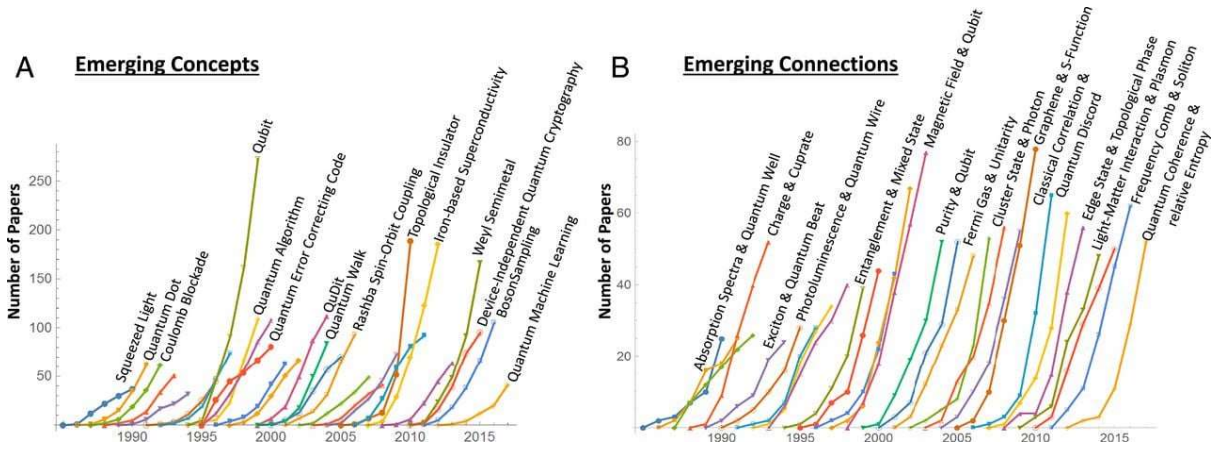
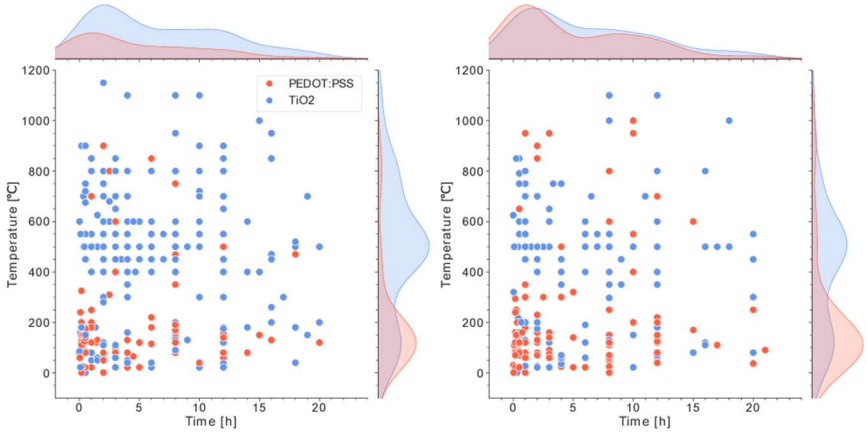


Figure 2.21 The evolution of quantum physics research observed using SEMNET reflected in the change in number of articles that contain a concept or concept pair per year from 1987 to 2017 [51].

In the field of inorganic materials, Kuniyoshi *et al*. proposed the label definitions for material names and properties and built a corpus containing 836 annotated paragraphs for training a Named Entity Recognition (NER) model [53]. NER is one of the most popular information extraction tools [54]. NER was initially developed as a text-mining approach for retrieving data from unstructured text such as names of persons, places, and organisations. The task is commonly approached with supervised machine learning, where a model learns to identify the keywords or phrases in a sentence. Entity normalisation, a process that involves mapping each entity onto a different database identifier, presents a significant challenge. The fact that a particular entity can be written in a variety of ways is the cause of problems. There are numerous efforts being made in the field of biomedical research to address this issue, but progress in the material domain is still in infancy. Kuniyoshi *et al*. achieved a micro-F1 score of 78.1% for the NER model. This model was then applied to 12,895 material research papers. With the retrieval of all the papers, this study was able to determine the trend of inorganic material research by analysing the change of keyword frequencies by year and country. As shown in Figure 2.22, the authors also visualised the temperature variation of 'PEDOT: PSS' and 'TiO_2' by year. These findings demonstrate that when synthesising 'PEDOT: PSS' and 'TiO_2', processing times for 'TiO_2' and 'PEDOT: PSS' are different in 2015. However, the processing times became similar in 2019. The processing temperatures remained constant in both years of 2015 and 2019. This indicates that there are similarities in the synthesis methods of 'PEDOT: PSS' and 'TiO_2'. They claimed that the results are useful for designing material synthesis processes.



(a) Temperature-by-time in 2015 (b) Temperature-by-time in 2018

Figure 2.22 Condition-by-year [53].

Similarly, there have been significant efforts to use NER to extract inorganic material synthesis instructions, however, no outcomes from large-scale texts of scientific papers. In 2019, Weston *et al.* applied information extraction from 3.27 million materials science abstracts and extracted more than 80 million materials science-related named entities [55]. The content of each abstract was recorded as a database entry in a structured fashion, and they were able to obtain an accuracy of 87% with their classifier. As seen in Figure 2.23, they utilised a word-level bidirectional LSTM model (a) and a character-level LSTM model (b) as the neural network architectures for NER. The goal was to train the models that encode knowledge of materials science effectively. The word embeddings and the results of the character-level LSTM ran over the same word to produce a list of entity tags which were used as the word-level features.

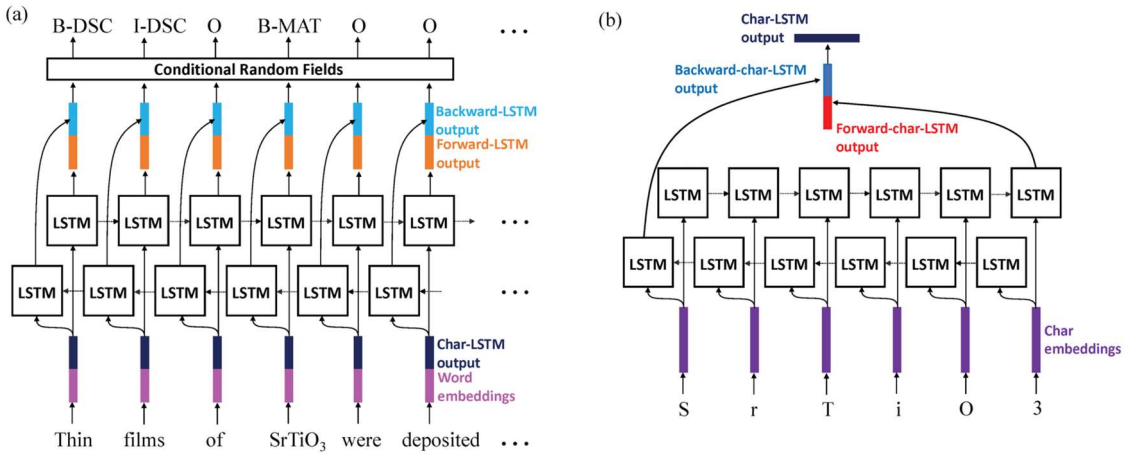


Figure 2.23 Neural network architecture for NER [55].

In this study, the word embeddings generated using Word2vec method were trained using bidirectional LSTM from a corpus of 3.27 million abstracts of materials science. Recently, Trewartha *et al.* [56] examined the performance of four different NER models on three materials science datasets, including the solid-state, doping, and gold nanoparticle synthesis datasets. All of these models consisted of a bidirectional long short-term memory (BiLSTM) and three transformer models (BERT, SciBERT, and MatBERT) that have varying degrees of pre-training from materials science. As can be seen in Figure 2.24, the MatBERT model outperformed the other models in terms of precision, recall and F1-score. They also observed that the models learned using MatBERT and SciBERT performed better within small data limit than the original BERT model. It has been demonstrated that the recently created transformer-based models offer notable performance improvements on NLP tasks. This is true since the MatBERT is a BERT-based model that was developed using the material science-specific scientific literature. The authors finally indicated that more specific pre-training would lead to improved performance.

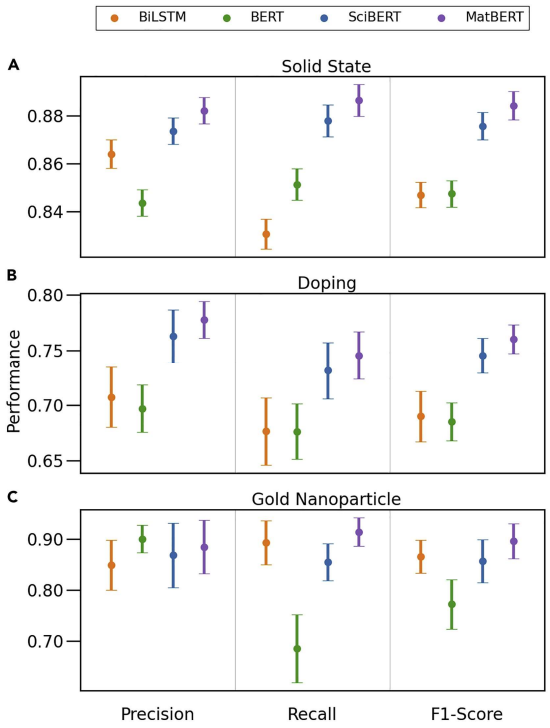


Figure 2.24 NER model precisions, recalls, and F1-scores [56].

In this way, researchers in biological science have made a lot of effort in developing BERT's domain adaptation abilities. A notable achievement is BioBERT [57], Other domain specific BERTs are SciBERT trained on scientific and biomedical corpus [58], clinicalBERT trained on 2 million clinical notes [59], mBERT [60] for multilingual machine translations tasks, and FinBERT [61] for financial tasks. Gupta *et al.* trained materials science domain-specific BERT named MatSciBERT and achieved state-of-the-art results on three downstream tasks, NER, relation classification, and abstract classification in comparison to SciBERT [62]. Figure 2.25 shows some comparisons between the selected image captions and the corresponding manual annotation by Venugopal *et al*. [63]. The task of assigning tags to each caption was carried out by human experts. While only one word was assigned per image caption in the previous studies, using the MatSciBERT NER model, multiple entities were extracted for the selected five captions. This illustrates that the large amount of information could be captured using the MatSciBERT NER model proposed in this study.

Captions with entities extracted using MatSciBERT NER model	Manual labels [10]
The comparison of XRD patterns of glass ceramic heat treated at 725 °C for 5h and rhombohedral Ba4Gd3F17. The superstructure reflections are marked with o. Inset: enlarged sections of XRD patterns.	Reflection
HRTEM image and the corresponding FFT pattern taken from as-deposited sample B(80/1) (a) and annealed sample D(30/1) (b); identifying rutile TiO2 crystal grains.	FFT
The illustrative schemes: a) The bonding of hexagonal ZnO nanocrystals to the glass surface. b) The structure of multi-layers coatings.	Crystal
(a) XRD patterns of the glass-ceramics sintered different holding times; (b) intensity of μ- and α-cordierite peaks count at (101) and (110) plane respectively as a function of sintering holding time.	XRD
Photoluminescence spectra of PbBr-based layered perovskites with an organic layer of naphthalene-linked ammonium molecules. Profiles: (a) 1; (b) 2; (c) 3; (d) 4; (e) 5.	Luminescence
Labels: Application (APL), Characterization method (CMT), Descriptor (DSC), Material (MAT), Property (PRO), Synthesis method (SMT), Symmetry/phase label (SPL)	

Figure 2.25 Comparison of MatSciBERT based NER tagging with manually assigned labels [62].

In 2021, Venugopal *et al.* successfully extracted useful knowledge from inorganic glasses' literature using NLP. They also demonstrated their framework on automating abstracts and image categorisation using an unsupervised NLP algorithm named Latent Dirichlet Allocation (LDA) to classify and search for semantic connections among publications. In this study, LDA was enabled to automatically divide the corpus based on the topic (*i.e.* 15 topics) where each topic was defined by the set of words that have the highest probability of occurrence within the topic and this enabled a text corpus to be clustered quickly and effectively with little assistance from humans. Using Term Frequency-Inverse Document Frequency (TF-IDF), which converts each article in the corpus into a unique vector in a higher dimensional space, each abstract in the corpus was vectorised. These vectors were then projected onto a 2D space using *t*-distributed Stochastic Neighbour Embedding (t-SNE), which groups the vectors according to their cosine similarity. Figure

2.26(A) shows their LDA plot that identifies the bioactive glass clusters with 94,207 abstract vectors which have been marked with the presence of fluorine and chlorine. Figure 2.26(B) illustrates the overlap between the red pixels corresponding to the ‘F’ and ‘Cl’ containing abstracts and the green pixels corresponding to abstracts on bioactive glasses. Upon examination of the LDA plot, it was observed that there are three areas of significant overlap. The authors note that the ability to identify scientific papers that meet the combined criteria of thematic category and chemistry is a noteworthy skill, as there is currently no alternative method available to achieve this. In Figure 2.26(C), randomly chosen images from this dimensional space are shown for reference (i–viii). The captions of the images verify that the figures do match the chosen image type, and their abstracts cover a wide range of subjects. The authors assert that each of these has a broad relationship to the subgroup of bioactive glasses being studied and this approach enables quick access to very specialised and subtle information sets in scientific data.

They also claim that a researcher who needs to access the microstructure of fluoride or chloride glasses without undertaking a thorough comprehensive literature review will find such a method to be of tremendous assistance. Although the LDA and CCPs offer a very specific, comprehensive graphical overview of the corpus of glass literature that is currently accessible, this analysis still requires assistance from domain experts to establish a topic list. Furthermore, there are now only a limited number of topics that can be detected using LDA. Since LDA is an unsupervised technique, the human experts must manually classify the images after understanding the topic from the keywords of literature. It is notable however that experts’ advice built from extensive literature reviews may introduce bias into the results causing by it being their personal area of expertise.

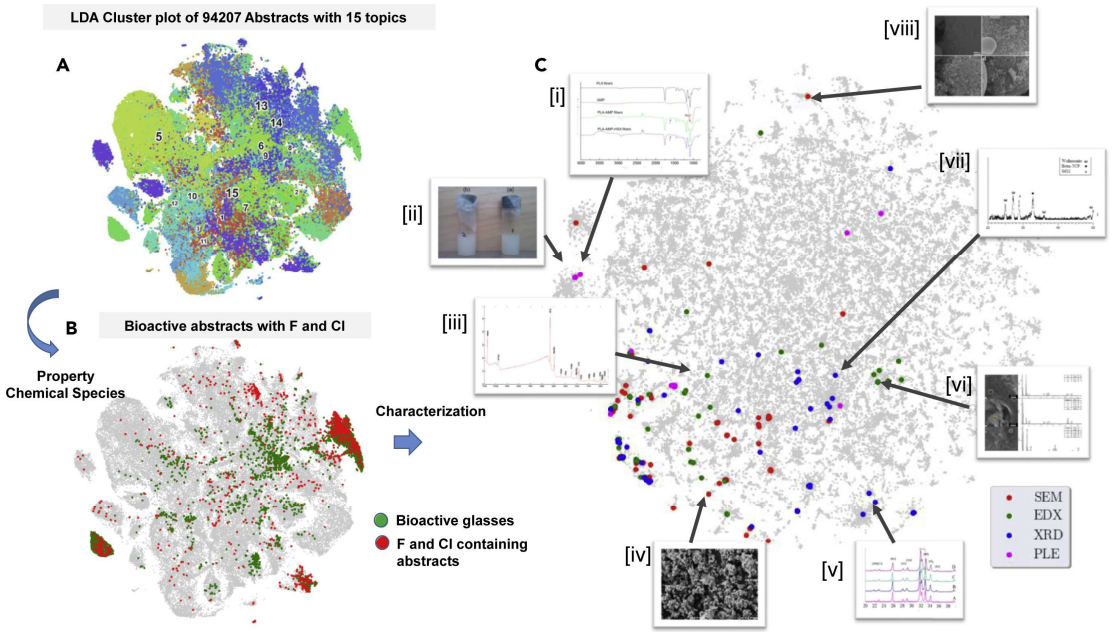


Figure 2.26 Knowledge extraction combining CCP, LDA and elemental maps [63].

In a related but more recent study, Wang *et al.* developed 35,675 solution-based synthesis recipes of organic materials by extracting information from the scientific literature using ML and NLP approaches [64]. Each recipe includes crucial synthesis information, such as the precursors and target materials, their quantities, the synthesis activities, and the relevant properties. The reaction formula is an addition to every recipe. Although there are many datasets for the synthesis of organic materials, there is currently no large-scale database of inorganic synthesis for AI-assisted design and optimisation. Human-written descriptions of syntheses need several degrees of interpretation to be converted into a codified, machine-operable format. This is known to be one of the main challenges. The first effort to establish a text mining pipeline to create a large-scale database collection of solid-state ceramics synthesis procedures was made by the same authors' group [65]. This collection is comprised of balanced chemical-reaction equations as well as the synthesis operations, their properties, and the end products in addition to the initial ingredients and finished goods. This study has developed a more advanced extraction pipeline which uses BERT and Materials entity

recognition (MER) model to extract recipe data for solution-based inorganic materials synthesis procedures from the scientific literature.

The researchers pointed out that solution recipes are more intricate than those for solid state synthesis and require the exact extraction of the component compounds and their relative concentrations. More complex organic and mixed organic-inorganic chemicals are also used to solubilise ions or control solution conditions. From more than 4 million articles, their extraction process ultimately established 35,675 solution-based inorganic materials synthesis formulas. The target material, precursors, and information about the synthesis procedures and their attributes were all extracted. The reaction formula for each synthesis process was then constructed using knowledge of the objectives and precursors. This dataset, which is the first large-scale collection of solution-based synthesis recipes, could open the door for future data-driven approaches to the synthesis and synthesisability of inorganic materials as well as the development of improved synthesis protocols for automated experimentation.15,638 precipitation synthesis reactions and 20,037 hydrothermal synthesis reactions were included in the dataset. A single JSON object was used to represent each record, which corresponds to a synthesis recipe taken from a paragraph. When many materials were synthesised in a paragraph, the corresponding reactions were broken out into separate data records. The metadata for each reaction also included the chemical formula as well as the data structure from their earlier work, which includes the DOI of the paper, the synthesis paragraph, chemical details, and operations with their corresponding attributes. The materials and appropriate quantities were also specified in the metadata. Table 2.1 provides information about the data format in detail. In addition to being recorded as a dictionary with lists of the precursors (`left_side`) and target materials (`right_side`) in the reaction, the chemical formula for the reaction is also stored as a string (`reaction_string`).

Table 2.1 Format of each data record: description, key label, data type [64].

Data Description	**Data Key Label**	**Data Type**
DOI of the original paper	`doi`	*string*
Snippet of the raw text	`paragraph_string`	*string*
Chemical formula	`reaction`	Object (*dict*): `-left_side`: *list of strings* `-right_side`: *list of strings*
Chemical formula in string format	`reaction_string`	*string*
Target material data	`target`	Object (*dict*): `-material_string`: *string* `-material_formula`: *string* `-composition`: *list of Objects* `-additives`: *list of strings* `-elements_vars`: {`var`: *list of strings*} `-amounts_vars`: {`var`: *list of Objects*} `-oxygen_deficiency`: *boolean* `-mp_id`: *string*
List of target formulas obtained after variables substitution	`targets_string`	*list of strings*
Precursor materials data	`precursors`	*list of Objects* (See **`target`**)
List of solvent formulas	`solvents_string`	*list of strings*
Sequence of synthesis steps and corresponding conditions	`operations`	*list of Objects* (*dict*): `-token`: string, `-type`: *string* `-conditions`: *Object* `--temperature`: *list of Objects* `--time`: *list of Objects* `--atmosphere`: *list of strings* `--mixing_device`: *list of strings* `--mixing_media`: *list of strings*
Materials with corresponding quantities	`quantities`	*list of Objects (dict):* `-material`: *string*, `-quantity`: *list of Objects5*
Synthesis type	`type`	string

The information on this table guides on what information should be extracted from scientific journals and how to specify the data type for organic material synthesis recipes. Elton *et al.* did not just extract from scholarly journals, but also gathered textual data from a variety of sources (*e.g.* journal articles, conference proceedings, the US Patent & Trademark Office, and the Defence Technical Information Centre archive on archive.org) and successfully extracted meaningful chemical-chemical and application-chemical relationships by performing computation with word vectors and without hand-labelling [66]. They developed a customised NLP pipeline to collect and identify the names of chemical

compounds, related function words (*e.g.* underwater, rocket and pyrotechnic) and property words (*e.g.* elastomer and non-toxic). All the obtained words were then vectorised using the two embedding methods (*i.e.* Word2vec) and Glove. Relationships between chemicals and their applications were found by calculating cosine similarity with word vectors. The cosine similarity $S_C \in [-1, 1]$ between two word-embedding vectors w_1 and w_2 is defined below:

$$S_C(w_1, w_2) = \frac{w_1 \cdot w_2}{\|w_1\| \|w_2\|} \quad (2.3.1)$$

Latent information about energetic materials was derived from the cosine similarity, allowing related materials to cluster together in the word embedding space. The most common chemical names found in the corpus are listed in Table 2, and the authors also compiled a list of 80 application words related to chemical names like ‘airbag’, ‘underwater’, ‘nontoxic’, ‘nanoparticle’, ‘binder’, and ‘desensitizer’. For each chemical name, the application words that are most similar are displayed. The overall frequency of each name is displayed (*N*), along with the number of times CDE recognised it as a chemical name (N_{CDE}). As shown in the table, CDE fails to recognise a significant portion of instances for most chemical names. This study explains that CDE could only recognise chemical names that are contained within sentences and not able to recognise chemical names in a list, which helps to explain its low detection rate. Although there was no hand-labelling for extracting meaningful relationships of chemical-chemical and application-chemical in this study, the authors should create a list of applications which are related to the chemical names.

Table 2.2 Most common chemical names in the corpus [66].

	chemical name	*N*	N_{CDE}	*Word2vec application words*	*GloVe application words*
1	HMX	5592	456	plastic, explosive	binder, plastic
2	RDX	5098	9	insensitive, explosive	binder, plastic
3	AND	4369	5	blasting, detonator	primary, oxidizer
4	TNT	3626	304	explosive, underwater	plastic, explosive
5	nitrocellulose	1960	1749	plasticizer, propellant	binder, plasticizer
6	TATB	1890	1779	insensitive, explosive	plastic, insensitive
7	PETN	1877	1736	explosive, detonator	plastic, insensitive
8	3-nitro-1, 2, 4-triazol-5-one	1484	13	primary, secondary	elastomer, thermoplastic
9	5-nitro-2H-terazole	1484	4	primary, secondary	elastomer, thermoplastic

Another study using data from human experts to pre-train a model and improve annotated samples manually was published by Vaucher *et al.* [67]. They created a set of synthesis actions with predefined properties for all the operations required to carry out the corresponding chemical reactions in order to transform unstructured experimental procedures written in English into structured synthetic steps (*i.e.* action sequences). They also developed their own rule-based NLP algorithm for the extraction of actions from experimental processes with associated chemical substances, amounts, and reaction conditions. The combined dataset used in this study added '`Yield`' actions from the Pistachio-based extraction to the ones produced by their NLP techniques. A deep learning sequence-to-sequence model which is regarded as cutting-edge for neural machine translation built on the transformer architecture was also used to translate experimental protocols into action sequences. The translation model was used to produce new translations after being trained on the action sequences created by combining the NLP and Pistachio approaches [59]. A random subset of 1.0 million experimental procedures was used to generate 4.66 million pairs of sentences and action sequences using their algorithm. Table 2.3 gives an example of an action sequence from actual experimental procedures such as `Add`, `Wait`, `Quench` and `Wash`.

Table 2.3 Action sequence extracted from an experimental procedure [67].

1	**MakeSolution** with methyl 3-7-amino-2-[(2,4-dichlorophenyl) (hydroxy)methyl]-1H-benzimidazol-1-ylpropanoate (6.00 g,14.7 mmol) and acetic acid (7.4 mL) and methanol (147 mL);
2	**Add** SLN;
3	**Add** acetaldehyde (4.95 mL, 88.2 mmol) at 0 °C;
4	**Wait** 30 min;
5	**Add** sodium acetoxyborohydride (18.7 g, 88.2 mmol);
6	**Wait** 2 h;
7	**Quench** with water;
8	**Concentrate;**
9	**Add** ethyl acetate;
10	**Wash** with aqueous sodium hydroxide (1 M);
11	**Wash** with brine;
12	**DrySolution** over sodium sulfate;
13	**Filter** keep filtrate;
14	**Concentrate**;
15	**Purify**;
16	**Yield** title compound (6.30 g, 13.6 mmol, 92%).

Figure 2.27 illustrates, for the actions observed in the ground truth, the action types predicted by the transformer model. Action types predicted by the transformer model are shown by labels on the x-axis, whereas ground truth action types are represented by labels on the y-axis. They noted that majority of the wrongly predicted actions are related to `NoAction` and `InvalidAction.` Interesting mistakes include confusing `MakeSolution` and `Add` (three times), forecasting `DrySolution` instead of `DrySolid` (two times) and `wait` instead of `Stir` (two times) or mistaking a `PH` action for an `Add` action. This plot was produced by counting all the action types that were correctly predicted and correspond to values on the diagonal. The remaining incorrectly predicted actions were then used to identify the off-diagonal elements. Actions exclusive to the projected set or the ground truth, respectively, were compiled in the final row and column. Although their predictions on the test set produced a perfect match of the action sequence for 60.8% of sentences, a 90% match for 71.3% of sentences, and a 75% match for 82.4% of sentences, there is a limitation resulting from their specific selection of action types and corresponding properties, which are not able cover all the operations in organic chemistry.

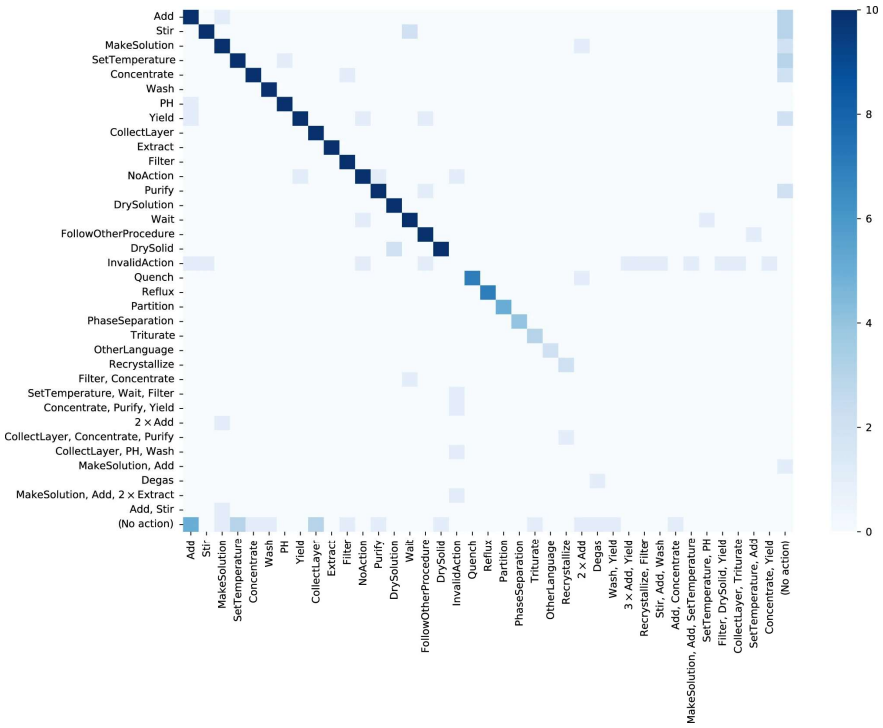


Figure 2.27 Visualisation of the correctness of predicted action types [67].

Data-driven approaches for fully automated extraction are limited by the completeness and substance of the data resources used. Using a wealth of gold nanoparticle synthesis and characterisation data available for data-driven approaches in an unstructured form in the scientific literature, Cruse *et al.* utilised NLP and text-mining techniques on 5 million materials science publications to extract the recipes and their outcomes [68]. The collected dataset is made up of 5,154 data records, each of which represents a single article on the synthesis of gold nanoparticles. Each record contains codified synthesis protocols and extracts morphological data from a total of 12,519 characterisation and 7,608 experimental paragraphs. As shown in Figure 2.28, they illustrated relationships between the use of specific precursors and the resulting AuNP morphologies. Each cell represents the proportion of morphologically targeted articles (*e.g.* the proportion of sphere-related articles) that use one or more of the precursors ($AuCl_4^-$, Citrate, BH_4^-, CTAB, Ascorbic Acid, and Ag^+) that specific cell uses in the recipe. The top left cell, for instance, demonstrates that the precursor

$AuCl_4^-$ is used in more than 90% of studies that are only focused on the synthesis of AuNP. Despite its limitations, the finding of this study give researchers studying AuNP a profound understanding of the subject.

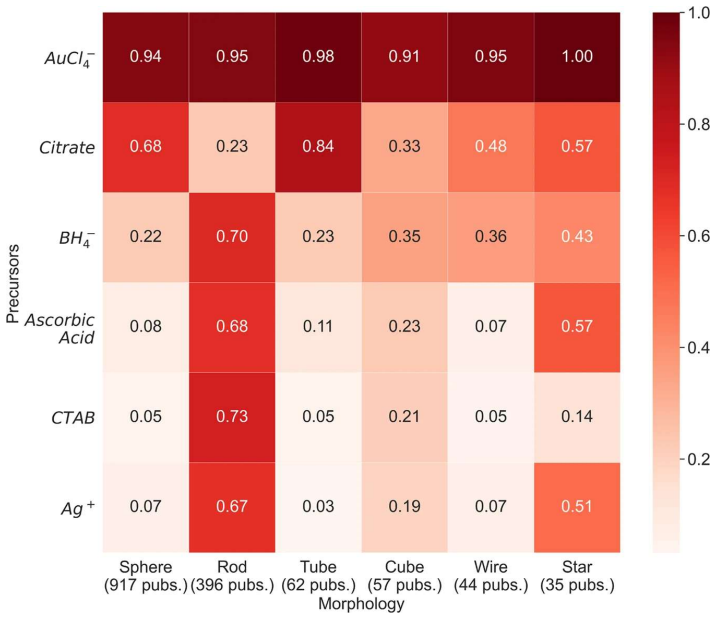


Figure 2.28 Heatmap depicting correlation between precursors and resultant AuNP morphologies [68].

2.5 Summary

This chapter presents an overview of the historical evolution of NLP and provides a summary of significant works that have employed NLP techniques in material science research. NLP techniques have been employed in material science to extract essential information, predict material properties, integrate data, and design new materials with desired properties. However, there are still significant challenges that need to be addressed in order to fully harness the potential benefits of NLP in material science. These challenges include the scarcity of annotated data, particularly in specialised domains such as material science, which

hinders the development of accurate and robust NLP models. Moreover, the scientific language used in material science is often complex and technical, with domain-specific concepts and structures that pose a challenge to NLP models. Addressing these challenges requires significant research and innovation to develop models capable of handling the linguistic complexity and scarcity of annotated data in material science. Overall, NLP has made significant strides in recent years, but there is still much room for improvement and innovation to fully leverage its potential in material science research.

In this thesis, advanced information extraction methodologies are employed to identify novel phase change materials (PCMs) that are both facile in fabrication and environmentally benign. Drawing inspiration from previous research, automated literature retrieval tools, notably CDE was utilised. To better cater to the specifics of our research, we have enhanced the capabilities of CDE. This was achieved by defining specialised functions that meticulously extract data pertinent to PCMs, followed by refining the model based on this curated data. Recognising the challenges posed by ambiguities in chemical nomenclature from earlier works, we have also integrated domain-specific ontologies and knowledge bases. This inclusion serves to enrich the context and provide a comprehensive understanding of the terms in use.

Chapter 3

Prediction of Metamaterials Research via Hyperdimensional Keyword Pool and Memory Learning

Conventional methodologies for trend forecasting in metamaterial research have largely depended on human experts. Such methods, while valuable, can be subject to individual biases, varied levels of expertise, and personal interests. These strategies often struggle to identify nascent keywords in their incipient stages. In this chapter, an innovative NLP-based technique that facilitates automated forecasting of scientific research trends by analysing unstructured texts and publicly accessible literature in the realm of metamaterial research is introduced. To realise this, a tailored recurrent neural network was developed, specifically the encoder-decoder LSTM, tailored to project forthcoming trajectories in metamaterial science. This initiative marks a significant stride towards harnessing automated, data-centric, and impartial mechanisms for anticipating future metamaterial research directions.

Metamaterials research in the modern era has been ongoing for more than 20 years and it has gained much public and scientific interests. Recently, there has been compelling evidence for its commercial success ranging from sensing, imaging to communications. Anticipating the possibilities for future research has become very popular recently based on big data and machine learning techniques. There have been many expert opinions forecasting on the road ahead for metamaterials, notably, in lieu of "knowledge tree" in 2010. Ten years on, this chapter proposes to re-examine these claims by using automated computer tools, such as NLP, to extract research information for processing and analysing from unstructured texts in publicly available scientific publications. As a results, a fully auto-generated database of 43,678 abstracts related to metamaterials published between 2000 and 2021 was built using Scopus Search API (Application Programming Interface). For assessing the popularity and trends of research themes, hyperdimensional vector spaces of keywords, clusters, and relationships between them also can be visualised. Finally, a trained neural network model was developed based on the encoder-decoder Long Short-Term Memory (LSTM) architecture to predict future directions and theme evolutions in the next four years for selected topics. This study not only provides vital information in terms of impact of metamaterials research but also lays down a solid foundation for the development of future metamaterial research roadmap in the form of Gartner's hype cycle.

3.1 Introduction

In scientific research, there is no doubt that setting a correct path is of great significance and importance to all key stakeholders in the scientific community, as it will ensure the proposed research to be effective and productive while reducing its cost by effective forward planning.

This could be accomplished by the analysis of the plethora of data of relevant and past studies using the predictive analytics.

Current human-expert-based approaches however could only be performed within one's area of expertise. Unlike the systematic review, where a formulated research question shall be provided with a collection of secondary data and meta-analysis, objectivity is often considered for scientific inquiry, as a good reason for valuing scientific knowledge, and as the basis of the authority of science in society [69]. The prediction made by a human expert is likely to be influenced by his or her particular scientific perspectives or personal interests leading to biased objectivity. In addition, the performance of such predictions is difficult to be measured quantitatively. Quantification in scientific prediction has a number of advantages such that it can lead to direct comparisons, time saving and large-scale analysis. Many researchers have therefore believed that every aspect of science can, and in fact should be quantified [70]. Thanks to electronic and open access publications, scientific data can now be made available in the task of automated information extraction from unstructured or semi-structured electronic documents, including articles, tables, and figures etc. This is particularly useful for researchers not only to retrieve an increasing amount of digital data from literature, but also to search information from patents, papers, and theses [67].

In 2010, Zheludev predicted the road ahead for metamaterials with a "Tree of Knowledge" describing the negative-index media as "forbidden fruit" [71]. The author articulated in his paper that the study of chiral, negative index and artificial magnetic metamaterials has matured, while the theory and technologies behind transformation optics, materials with high/low epsilon and designer dispersion have been "ripe". The author also reported that switchable metamaterials based on arrays of micro- and nano-electromechanical devices were also being developed and this research was highly interesting. This is also true with active metamaterials, such as those based on non-foster's theorem [72], which loss and

bandwidth issues of metamaterials may be overcome. Most of metamaterials exhibit strong local field enhancement near their resonances and it makes metamaterials attractive for nonlinear optical applications. Sensor applications were another growth area in metamaterials research where 2D materials such as graphene and TMD (transition metal dichalcogenides), which a single molecular layer of carbon can induce a multifold change in the transmission of metamaterials [73]. Finally, the author predicted that superconducting metamaterials would find their applications in exploiting quantum coherence with a multilevel quantum structure replacing the classical plasmonic resonators.

Meanwhile, ten years on, there have been many bodies of review articles summarising recent research and commercial developments of metamaterials [74], [75]. These papers and reports are often focused on specific subject areas and fail to present a whole picture of recent metamaterial research. None of these studies has validated previous predictions and research roadmap development with a systematic study of a large volume of historical and recent research outcomes. More importantly, many of reviews and technical reports have been written in line with author's area of expertise.

This chapter introduces a novel framework that combines automated data extraction from scientific databases with NLP for intelligent forecasting. As the scientific findings in this domain are published in a variety of academic venues, this study uses Elsevier's Scopus API which allows access to 50 million abstracts of over 20,500 peer reviewed papers from more than 5,000 publishers, capturing articles published in virtually all scholarly journals of any significance in the world, including the American Association for the Advancement of Science, Springer-Nature, the Institute of Electrical and Electronics Engineers (IEEE) and the American Institute of Physics (AIP) [76].

This study differs in several aspects from the previous ones. First, all keywords of metamaterials research are extracted automatically using the RAKE algorithm, these

keywords then categorised into objects and properties depending on their attributes. The list of 3,187 keywords is vectorised as unstructured natural texts not only for visual analytics, but also future trend forecasting based on their frequencies appeared in the search as well as a sequential machine learning algorithm. A clustering analysis of all keywords via unsupervised mapping and labelling is then conducted. As a result, 10 mostly appearing keywords from 8 clusters in the field of metamaterials have been obtained. This study also demonstrates the feasibility of using a modified recurrent neural network (*e.g.* encoder-decoder LSTM) for predicting research trends for the next four years from a sequence of published data collected between 2000–2021. Finally, a hype cycle is automatically generated to trace the evolution of metamaterials research as they pass through successive stages pronounced by the peak, disappointment, and recovery of expectations.

An overall architecture of the proposed approach can broadly be described including four key steps, namely, data extraction, hyperdimensional keyword pool building, keywords relationship visualisation, and future trend forecasting as shown in Figure 3.1. This framework is not designed simply to assist a human expert but rather substitute the conventional literature review and roadmap development process via its four aforementioned novel concepts used within the framework, and it alone can perform any prediction tasks accurately on any subject of scientific research while reducing associated project costs.

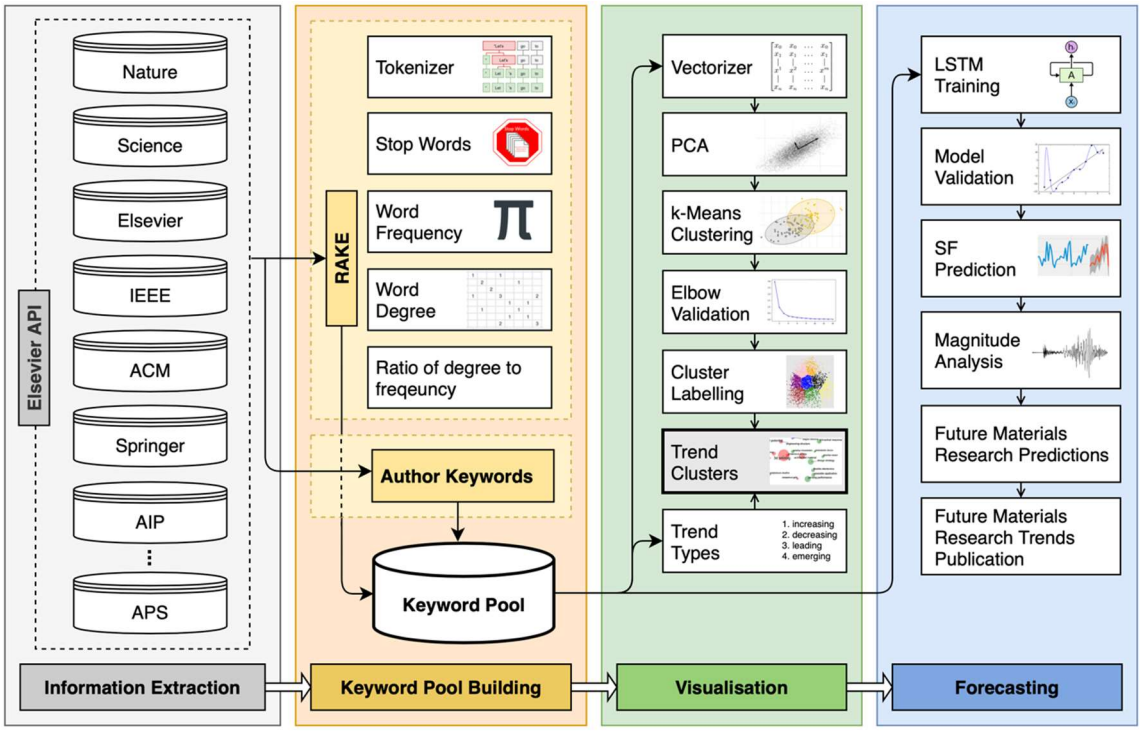


Figure 3.1 The architecture of the proposed keyword prediction system.

The keyword prediction system is composed of four main components, namely, Information Extraction (IE), Keyword Pool Building (KPB), visualisation and forecasting. The Keyword Pool (KP) is built as a result of the RAKE algorithm and author's keywords in the KPB phase that processes 43k abstracts. The visual analysis of KP is performed to seek the relationships among the selected keywords in KP. The Normalised Frequencies (NFs) of keywords in KP are fed into the encode-decoder LSTM and the learned models are validated for further processing such as magnitude analysis. The future materials research trends are finally predicted and published.

3.2 Data Collection

The keyword of metamaterial(s) on Elsevier's Scopus returns the various data on 43,678 papers (published until the second quarter of 2021) that include title, abstract, author's

keywords, published date etc. Figure 3.2 shows the number of metamaterial publication per year. In the year of 2000, the first metamaterial-related paper recorded in Elsevier Scopus. Since then, the number of metamaterial publications has increased steadily until 2019. In 2004 and 2005, the number has doubled. To learn an accurate model that achieves a desired level of performance using such time-series data, as many as examples are needed and the 20 data points exist from 2000 to 2021 would never be enough. Instead of artificially increasing the sample size (*e.g.* data augmentation), as in Table 3.1, the number of publications per term has been adopted and normalised.

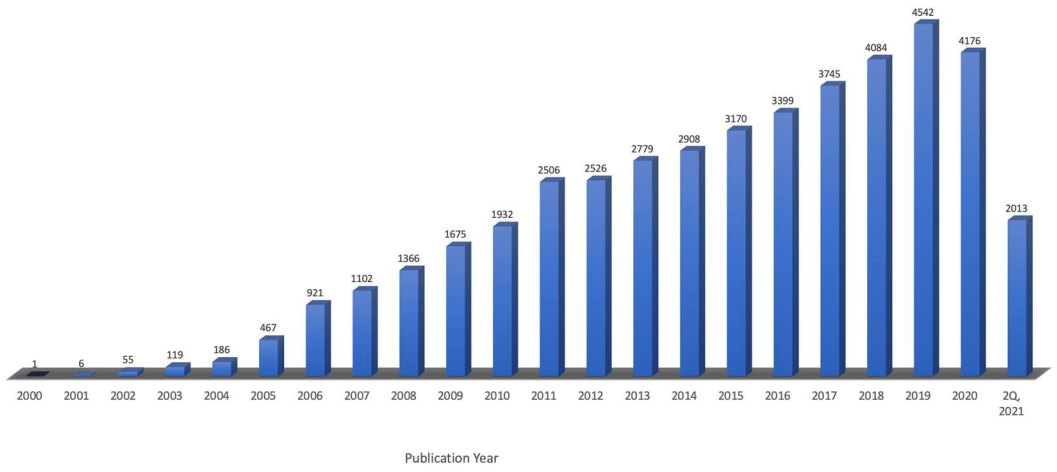


Figure 3.2 Number of metamaterial publications by year since 2000.

Table 3.1 Number of metamaterial publications by quarter.

Year	*1st quarter*	*2nd quarter*	*3rd quarter*	*4th quarter*	***Sum***
2000	0	1	0	0	**1**
2001	5	1	0	0	**6**
2002	25	3	18	9	**55**
2003	42	5	29	43	**119**
2004	77	17	35	57	**186**
2005	125	56	48	238	**467**
2006	380	70	79	392	**921**
2007	360	143	139	460	**1102**
2008	452	207	239	468	**1366**
2009	422	213	239	801	**1675**

2010	486	340	291	815	**1932**
2011	579	566	445	916	**2506**
2012	601	422	433	1070	**2526**
2013	1017	406	491	865	**2779**
2014	2057	229	306	316	**2908**
2015	1439	381	465	885	**3170**
2016	1141	611	708	939	**3399**
2017	1395	647	633	1070	**3745**
2018	1521	676	794	1093	**4084**
2019	1757	861	1002	922	**4542**
2020	1546	715	1004	911	**4176**
2021	1309	704			**2013**

An abstract of published research articles is a self-contained, short and powerful statement that describes or summarises a whole paper [77]. Learning from abstracts for keyword prediction offers a number of benefits: First, it helps text pre-processing and learning by reducing the complexity that comes from irrelevant information potentially contained in other parts of a paper. Second, the computational cost of text pre-processing and learning could significantly be reduced due to a smaller amount of texts in an abstract than those in a whole paper. This study thus not only improves the efficiency of the required NLP and learning processes but also leads to accurate learning for forecasting.

Scopus client is one of the Elsevier's APIs which allows an access to its largest database of abstract and citation of the literature and relevant web sources. Scopus API particularly indexes 'metadata' from abstracts and references of thousands of publishers including Elsevier. For the data extraction task, connecting the Scopus repository[3] with the newly coded X-ELS-APIKey, multiple JavaScript Object Notation (JSON) files were downloaded. These files contain the information of each paper related to metamaterial studies such as title, publication date and abstract from year of 2000 to 2021. Figure 3.3 displays an example of the data we downloaded from the year 2010.

[3] https://dev.elsevier.com/

	Title	Author Keyword	Coverdate	Term	Citation	Publication Name	Abstract
0	Radiation pressure force enhancement in metamaterials		2010-12-31	4	0	European Conferenc	A new generation of ultra-narrow band gap materi
1	Dynamic self-assembly and co	Defocusing | Hydrodynamic int	2010-12-28	4	147	Proceedings of the N	Engineered two-phase microfluidic systems have r
2	Performance enhancement of terahertz metamaterials on ult		2010-12-27	4	129	Applied Physics Lett	We design, fabricate, and characterize split-ring re
3	Cavity-involved plasmonic metamaterial for optical polarizati		2010-12-27	4	93	Applied Physics Lett	We experimentally demonstrate a plasmonic assis
4	Three-dimensional microcoils as terahertz metamaterial with		2010-12-27	4	12	Applied Physics Lett	A metamaterial consisting of three-dimensional su
5	Sum rules and physical bounds in electromagnetic theory		2010-12-27	4	1	Symposium Digest -	Sum rules are useful in many branches of physics a
6	A geometrically simple benchmark problem for negative inde		2010-12-27	4	1	Symposium Digest -	The proposed model problem is geometrically very
7	Nanolithography in the evane	Evanescent near-field | Gain-a	2010-12-27	4	0	Proceedings of SPIE	Surface Plasmon polaritons are electromagnetic wa
8	Experimental verification of t	Inverse Doppler effect | Negat	2010-12-27	4	1	Proceedings of SPIE	Research of negative-index material (NIM) is a very
9	Thin wideband absorber with optimal thickness		2010-12-27	4	7	Symposium Digest -	The known methods for designing nonmagnetic ab
10	Multipole model for metamaterial homogenization		2010-12-27	4	0	Conference Proceed	Homogenization of optical metamaterials is one of
11	Transmission line model with X-circuit for a metamaterial lay		2010-12-27	4	1	Symposium Digest -	In this paper we analyze the propagation through a
12	Electromagnetic investigation about composite right/left han		2010-12-27	4	0	IEEE International Sy	This paper investigates about the main properties
13	Sum rules and constraints on passive systems with applicatio		2010-12-27	4	0	Symposium Digest -	A passive system is one that cannot produce energ
14	Conference Proceedings - 5th International Conference on Ac		2010-12-27	4	0	Conference Proceed	The proceedings contain 96 papers. The topics disc
15	Broadband terahertz modulators based on reconfigurable me		2010-12-27	4	10	Symposium Digest -	A new scheme for broadband terahertz modulatior
16	Metamaterial-based microstrip antenna with ground slots for		2010-12-27	4	2	Symposium Digest -	In this paper, a novel design of the metamaterialba
17	Micro-/nano-photonic device structures applied to communic		2010-12-27	4	1	Conference Proceed	Device structures in high refractive index materials
18	A transformation-optics-inspired route to sensor invisibility b		2010-12-27	4	1	Symposium Digest -	In this paper, we introduce and explore a transforn
19	Radiation from an electric dipole axially mounted above a sph		2010-12-27	4	2	Symposium Digest -	An oblate semi-spheroidal cavity is flush mounted
20	Tunable nonlinear metamater	Metamaterials | Nonlinear plas	2010-12-27	4	0	Conference Proceed	We discuss tunability and nonlinear properties of r
21	Zero reflection from anisotropic metamaterial stratified struc		2010-12-24	4	4	Progress In Electrom	A method of solving the scattering problem for ger
22	Synthesis design of metamate	Absorber | Genetic algorithm |	2010-12-24	4	2	Conference Proceed	In this paper, a simple and efficient method, genet
23	Variable magnetic and electri	copper thickness | electric and	2010-12-24	4	0	International Journa	Electric and magnetic resonances of split-ring reso
24	2010 International Symposium on Signals, Systems and Electr		2010-12-24	4	0	Conference Proceed	The proceedings contain 181 papers. The topics dis
25	PV metamaterial based on nanostructured Si		2010-12-24	4	0	Materials Research S	There are several ways to nanostructure Si. Some c

Figure 3.3 The example of data collecting from Scopus API.

3.3 Hyper-dimensional Keyword Pool

To build a domain-specific keyword pool, this study used various techniques of locating and defining keywords, namely, unsupervised clustering, NLP-based keyword extraction, hierarchical/agglomerative clustering, and word-to-vector embedding.

3.3.1 Keyword Pool Building using Frequencies

Firstly, the keywords based on only word frequencies were extracted as the words that appear most frequently being the most general descriptors of the scientific content. An NLP algorithm that reads 43,678 abstracts from the .csv files created during the data extraction phase were then developed. The NLP algorithm could help reduce the complexity of the vocabularies in the downloaded abstracts. Each character was converted to a lower case and unwanted texts such as punctuation, numbers, non-alphabets, and other characters which might not be a part of language are removed. The expansion of contraction was in turn performed (*e.g.* thz → terahertz) and the abbreviation of a meaningful word sequences was

created (*e.g.* split ring resonates → srr and electromagnetic band gap → ebg). The pre-processed texts were now tokenised to cleansed texts in which stopwords (*e.g.* of, are, the, and it) were removed. The error rate of recognising a same meaning word as a different meaning word thus must be reduced. Stemming has been the most widely adopted morphological technique for information retrieval [78]. Stemmer reduces a word to its root term by defined rules (*e.g.* dielectric → dielectr, tunable → tunabl and absorption → absorpt). This study adopted this technique in order to reduce the number of the cluster of a same meaning word so that the total number of keywords could significantly be reduced. It should be noted that some keywords were manually removed by domain expert as they were identified as non-keywords (*e.g.* frequency, band and wave). 172 potential keywords from both green and yellow zones of Figure 3.4 based on their frequencies only were then selected.

To extend this one-dimensional nature (*e.g.* frequency) of the keyword pool to hyperdimensional, the selected 172 keywords were plotted onto the hyperdimensional spaces using Mat2Vec [40], a pretrained word embedding model which extracts multidimensional vector representations of each word. Mat2Vec is built from an unsupervised word embedding using 3.3 million scientific abstracts. Each of our selected keywords was vectorised in this pre-trained model. In the following experiments, two groups of selected keywords, namely, KP-T1K and KP-T2K were used. KP-T1K is the keyword pool which has 84 hyperdimensional keywords from the top 1,000 most frequently appeared words (T1K) as indicated in green zone while KP-T2K is the keyword pool which has 88 keywords from yellow zone. Each and every word in both KPs were vectorised into 200 dimensions. All selected keywords are provided in Appendix A Section A.1.

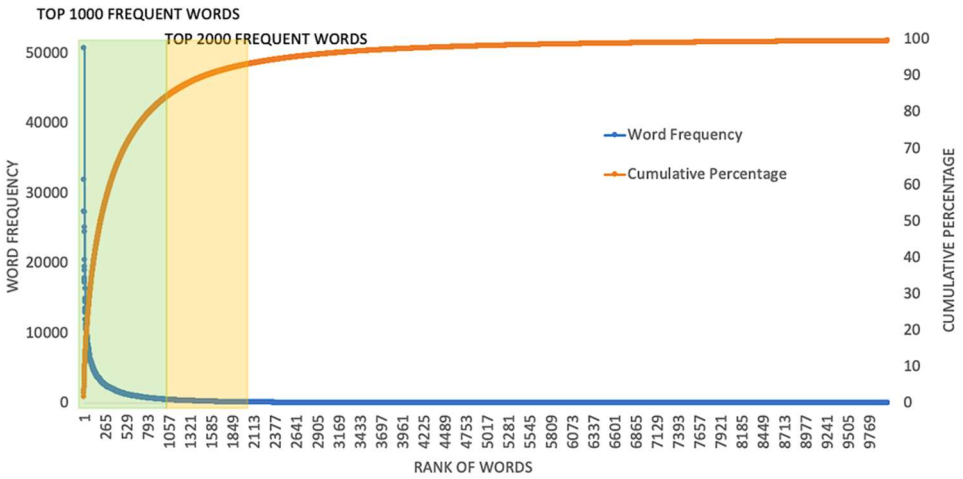


Figure 3.4 The coverage of T1K and T2K words.

The numeric frequencies of the words are then re-arranged in descending order. In the total of 130,000 words, the 2,000 most frequently appeared words (T1K and T2K) cover over 90% of accumulates percentage. 84 words in green zone (T1K) were the most frequently appeared (*e.g.* 500–50,000 occurrences since 2000). 88 words in yellow zone (T2K) were the next most frequently appeared (*e.g.* 100–500 occurrences).

The words in the current keyword pools are in different hierarchical dimensions. Each word could be categorised as object, property, or both. For an example, an object 'lens' has its properties of 'transmissive', 'optical' and 'refractive'. The automatic labelling of each keyword with such hierarchical dimensional information saves the time and resources spent on. Clustering analysis performs an unsupervised mapping and labels each word based on identified clusters in the vector space. For clustering analysis, the high dimensionality of the words in the current KPs (*e.g.* 200 dimensions) could be undesirable. The curse of dimensionality theory [79] well demonstrates that as the data is moving into higher dimensions, the sparsity of data and statistical error grow exponentially. The current number of dimensions were reduced by computing the Principal Components (PCs) that simplifies the complexity in high-dimensional data while increasing interpretability. The resulting plots are depicted in Figure 3.5.

Clustering could be categorised into two broad sub-categories, hierarchical and partitional. The hierarchical clustering considers a view of the data at different levels of

granularity and organises the data points into a disjoint cluster. A concept of hierarchy defines a sequence of mappings from a set of low-level concepts to higher-level or general concepts. In the case of this study, the words cannot be split into a distinct disjoint group because some words could belong to more than one clusters (*e.g.* object and property). In this hierarchy concept, each word must belong at least one of the higher-level or general concept as depicted in Figure 3.6.

Agglomerative clustering builds a cluster hierarchy that is commonly displayed as a tree diagram called a dendrogram. The algorithm begins with each object word in a separate cluster. At each step, the clusters with high similarities are merged to form a new single cluster. When the clusters are merged, the Ward's method was adopted as a linking method. The Ward's method, unlike other linking methods which measure the inter-cluster distances, uses the cluster similarities based on the increase in squared error when clusters are merged. According to the Ward's method, the distance between two clusters, A and B, is calculated as follows (3.3.1).

$$C(A,B) = \sum_{i \in A \cup B} (x - m_{A \cup B})^2 - \sum_{i \in A} (x - m_A)^2 - \sum_{i \in B} (x - m_B)^2$$

$$= \frac{n_A n_B}{n_A + n_B} (m_A - m_B)^2 \qquad (3.3.1)$$

where m_A is the centre of cluster A, n_A is the number of points in cluster A. C is the merging cost of combining the clusters A and B. In agglomerative clustering, the sum of squares starts out at zero and then grows as it merges other clusters. The Ward's method keeps this growth as small as possible.

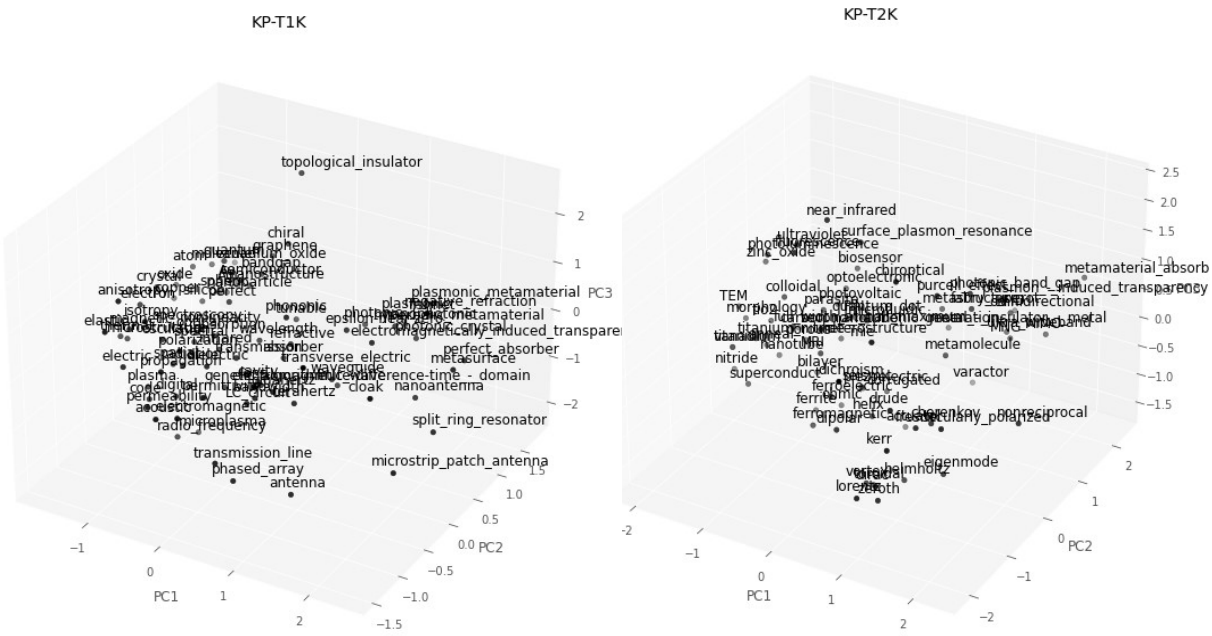


Figure 3.5 The visualisation of KP-T1K and KP-T2K using 3 PCs.

As depicted in Figure 3.6, a tree-like dendrogram could be visualised as the results of agglomerative clustering. By adjusting the threshold point (*r. h. s* of Figure 3.6) the number of clusters in this dataset could be chosen. Using agglomerative clustering on KP-T1K, the optimal number of clusters was found.

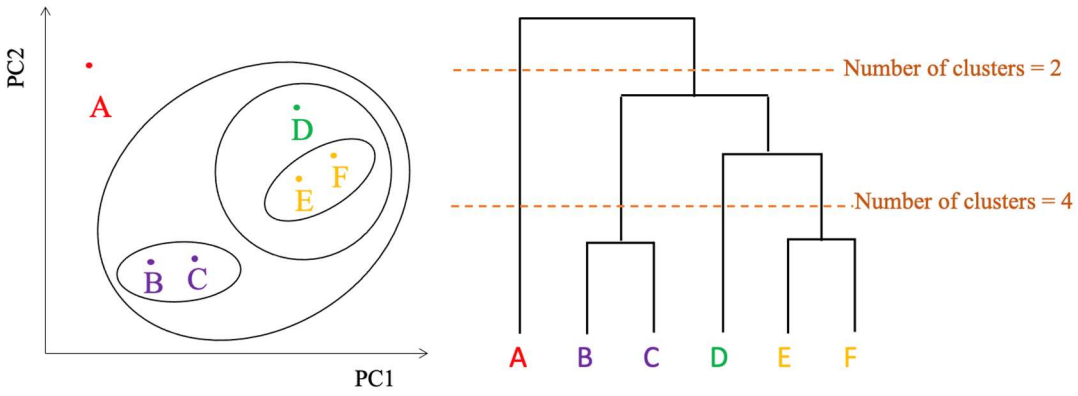


Figure 3.6 The dendrogram with different threshold points and its representation.

To demonstrate the performance quantifiable measures of the clustering techniques, seven clustering methods were validated using SC, CHI and DBI as shown in Table 3.2. As observed, Agglomerative showed the highest validation scores in two of the three validation categories (*e.g.* SC and DBI) demonstrating its suitability. SC and CHI, the ratio of the sum of inter-clusters dispersion for all clusters, give higher scores when clusters are dense and well separated. DBI relates to a model with better separation between the clusters, the lower the score, the better the performance.

Table 3.2 The validation results of the clustering performance.
The validation results of the seven clustering algorithms are compared using Silhouette Coefficient (SC), Calinski-Harabasz Index (CHI) and Davies-Bouldin Index (DBI). As for partitional algorithms, two types of k-means, standard k-means and Mini-batch k-means, as well as Spectral and Gaussian Mixture Model (GMM) are used. The best validation results are underlined.

	Partitional				Hierarchical		Density-based
	k-means	k-means (MB)	Spectral	GMM	Agglomerative	BIRCH	DBSCAN
SC	0.35	0.35	0.34	0.27	<u>0.36</u>	0.32	0.07
CHI	<u>88.42</u>	85.27	67.74	57.35	86.19	79.94	10.53
DBI	0.88	<u>0.83</u>	0.85	1.14	<u>0.83</u>	0.88	1.90

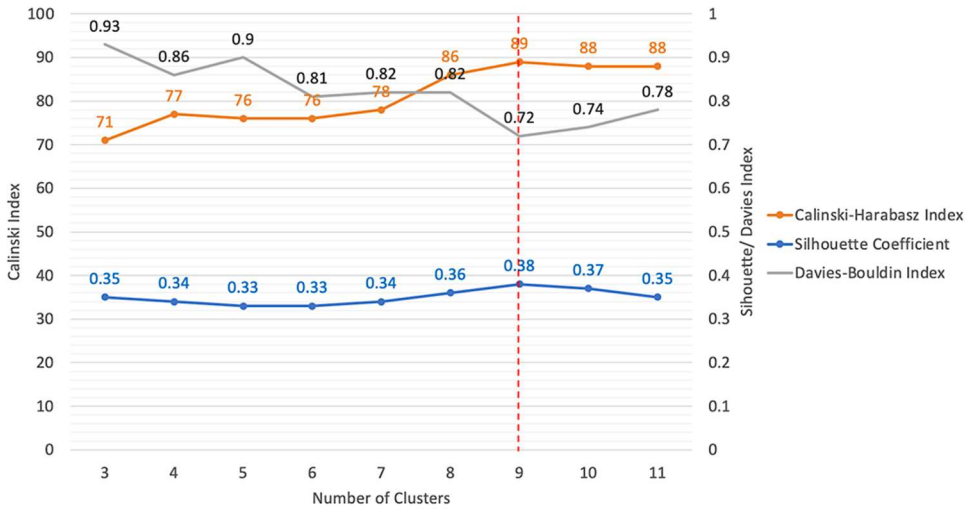


Figure 3.7 The validation results from KP-T1K.

Figure 3.7 shows that the agglomerative clustering algorithm give best scores when there are nine clusters in KP-T1K. The clustering results using 2PCs are plotted in Figure 3.8 and the corresponding dendrogram is visualised in Figure 3.9. The clustering results and its dendrogram of KP-T2K is provided in Appendix A Section A.2.

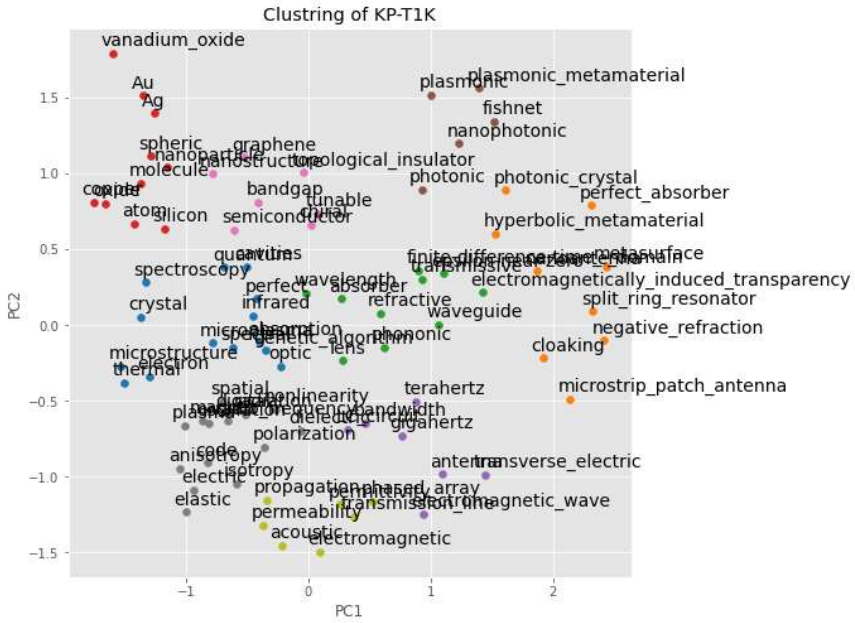


Figure 3.8 The result of agglomerative hierarchical clustering with 9 clusters from KP-T1K.

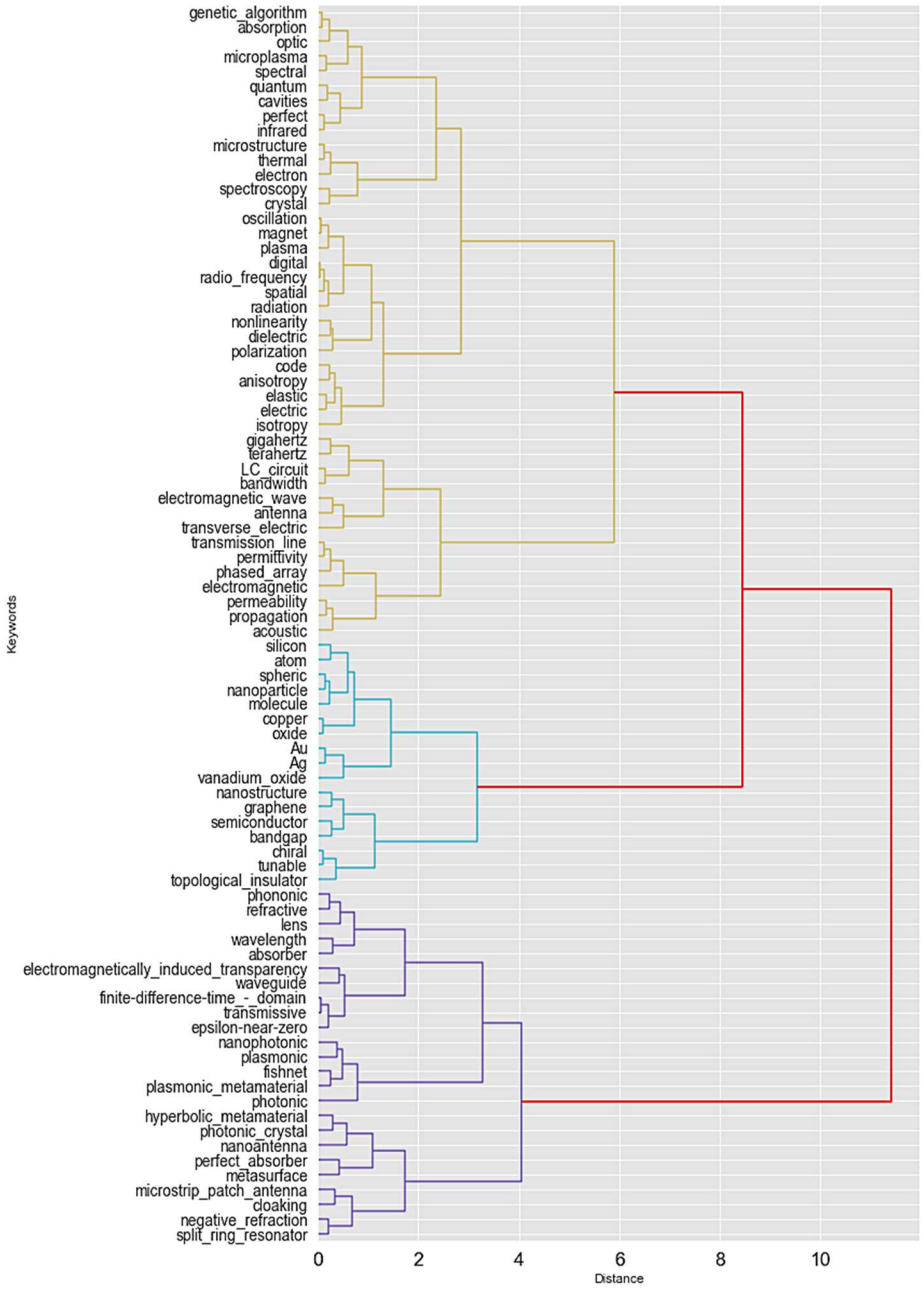


Figure 3.9 The dendrograms from KP-T1K.

Using word frequency to build a keyword pool has some significant limitations. When selecting a keyword, you can only use one word rather than a key phrase. However, single word could categorise as either an object, a property, or both. As a result, using a keyword with just one word won't yield meaningful results. Consequently, the agglomerative clustering algorithm was used to identify various hierarchical dimensions (*e.g.* objects and

properties). It was noticed that categorising without the assistance of a human expert is difficult.

3.3.2 Keyword Pool Building using RAKE

RAKE finds the most relevant words or phrases in a piece of text. Using such key phrases is of particular importance as object-property pairs could be identified. One potential issue however is that too many candidate keywords could be selected due to the way to find keywords-pairs, which is mainly based on word occurrence or frequency. For example, some of the less meaningful keywords such as 'huge potential', 'grown rapidly', 'current issues' could be selected as they have been frequently used in the literature.

To address this issue, the less-meaningful keywords were filtered out by using the words in the title and the list of author's keywords as the reference. This indicates that RAKE's chosen keywords have been confirmed by what human experts consider substantial. Some of published papers accessible via Elsevier's API do not have author's keywords. In this case, the words appear in both the titles of the literature and the RAKE generated word list were selected. If the paper has author's keywords, the words appear in both the list of author's keywords and the RAKE generated word list were selected. Here, RAKE was set to choose up to 2-words phrase. Setting RAKE to choose 3- or more-word phrases may bring additional complexity in searching candidate keywords. The example of keyword selection from a typical paper published in 2020 is shown in Figure 3.10(a) [80]. This is a fully automated process that creates a KP which contains 3,187 keywords. A simple statistical analysis of the KP shows that the most frequently occurring keywords over the most recent 20 years are 'transformation optics', 'photonic crystal', 'negative refraction', 'mutual coupling' and 'negative permittivity'.

For further analysis, all the selected words in the KP were vectorised onto the hyper-dimensional space using Mat2Vec [40]. Each word in KP is however either 1-word or 2-word. While the 1-word keyword could easily be vectorised, in the case of 2-word keyword, each word in the 2-word keyword was vectorised then added together by using the word embedding analogy which allows vector-oriented reasoning based on the offsets between the pair of words. For example, the male/female relationship is automatically learned, and with the induced vector representations, "King – Man + Woman" results in a vector very close to "Queen." [81], [82]. Likewise, a vector using the analogy for each 2-word keyword was computed. For example, 'acoustic cloaking' was embedded onto the sum of the vectors of 'acoustic' and 'cloaking' as shown in Figure 3.10(b). Each keyword in the KP was respectively vectorised into 200 dimensional embeddings as shown in Figure 3.11, which represents the process how the KP is built and visualised. These high-dimensional embeddings also can graphically be represented via TensorBoard Embedding Projector. Rendering of the high-dimensional embeddings into two or three dimensions helps the analysis, examination and understanding of the embedding layers. All the embedded keywords data are provided in Appendix A Section A.3.

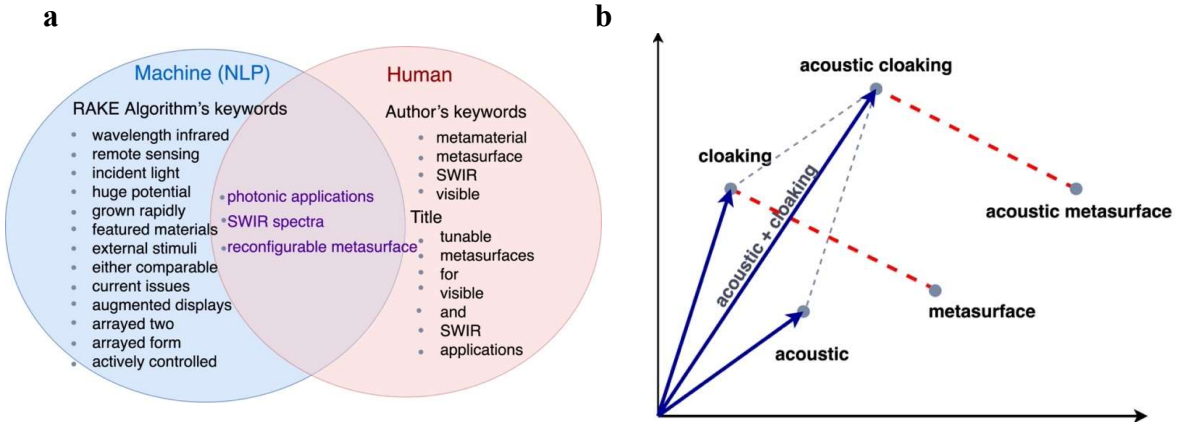


Figure 3.10 The methods of selecting keywords.

a The words in blue circle are the ones selected by RAKE using the abstract while the words in red circle are from paper's title and author's keywords. The words in the intersection are the selected keywords. **b** In word embedding, each word is represented in vector space thus the words could be added or subtracted for analogy. For example, 'acoustic cloaking' can represent the sum of 'acoustic' and 'cloaking', and 'acoustic cloaking' is also embedded with amount of a distance between 'metasurface and 'cloaking' from 'acoustic metasurface'.

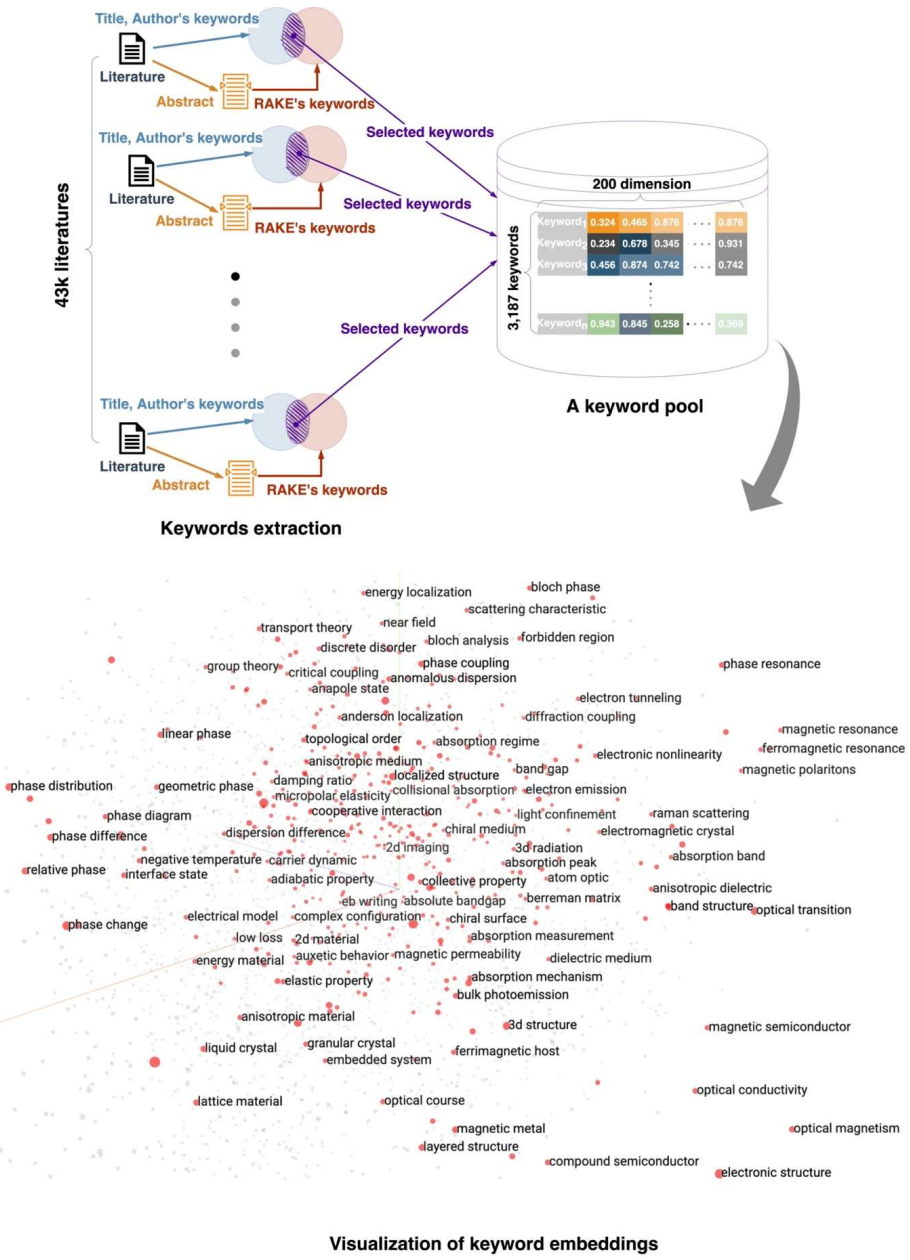


Figure 3.11 The process of selecting keywords and embedding.
From 42k literatures, the overlapping words between RAKE and author's keyword and title are selected as the keywords in KP. The KP has 3,187 keywords which are then vectorised into 200 dimensional spaces for the visualisation on 3 dimensional spaces (r.h.s).

To gain a high-level view of all relevant topics in metamaterials research, keyword clusters were created by merging words with similar meaning. This stage had 3,187 embedded keywords in 200-dimensional spaces with no label. A clustering analysis was thus used in order to perform an unsupervised mapping and label each word based on pre-identified clusters in the vector space. For the clustering analysis, the high dimensionality of the words in the KP was undesirable. The curse of dimensionality theory [79] demonstrates that as the data is moving into higher dimensions, the sparsity of data and statistical error will grow exponentially. The current number of dimensions was reduced, as illustrated in Figure 3.12, by computing the principal components that simplify the complexity in high-dimensional data while increasing interpretability. It draws on a distance-based clustering algorithm, k-means [83], [84], which aims to minimise the sum of squared Euclidean distances to each cluster mean (centroid) defined as follows.

$$\min \sum_{c=1}^{k} \sum_{x \in S_c} \| x - \mu_c \|^2 \tag{3.3.2}$$

where $\{\mu_c\}_{c=1}^{k}$ are the cluster means and S_c are all vectors assigned to cluster c. The algorithm alternates between reassigning vectors to the closest cluster means, and then updating the means [85]. The number of clusters, k, is found using the elbow method [86], [87]. In the elbow method, the x-axis changes are used to select the point that exhibits the smallest rate of change while the y-axis displays the sum of squared distances. The elbow method was used to determine the value of 8 ($k = 8$).

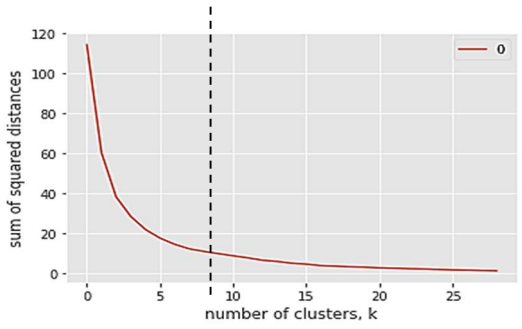


Figure 3.12 The elbow method for choosing the k.

The resulting plots are depicted in Figure 3.13 demonstrating labelled clusters and 10 most frequently appeared keywords for each cluster. The data space is split into 8 clusters and each cluster is labelled. In this way, the keywords could be analysed as a group instead of individuals. Green cluster seems to be focused on applications category, however, 'pin diode' shall be in red cluster for fabrication and enabling materials category. 'liquid crystal' in pink cluster which seems to focus on material constitutive parameters however shall be in red cluster. Blue cluster may represent modelling and optimisation category, yet 'energy harvesting' from blue cluster shall be in green cluster for applications. Purple cluster is focusing on wave phenomenon; yellow cluster focuses on 'antennas' while orange cluster is for optical properties; black cluster seems to focus on guided wave devices category.

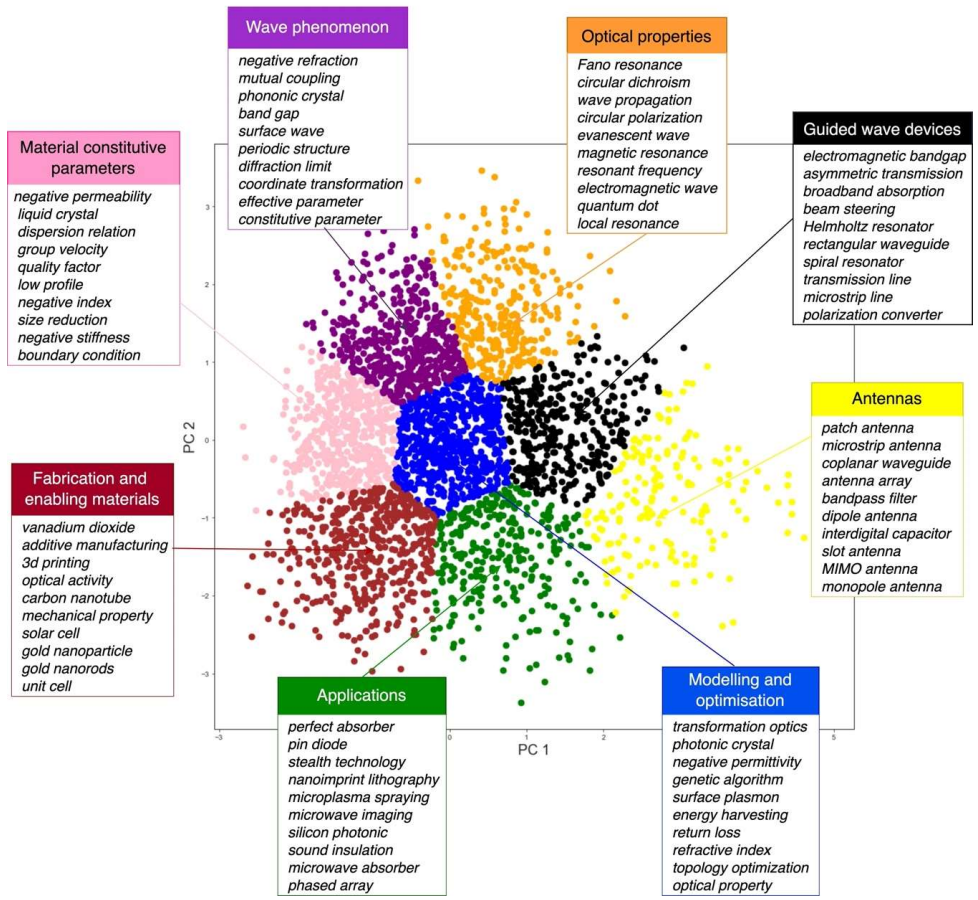


Figure 3.13 The labelled clusters using k-means and 10 most frequently appearing keywords from each cluster.

3.4 Time-series Data Building

To analyse the trend of word occurrences over the years, time series data related to appearance frequencies of each keyword was built. The frequencies of each keyword per year over the duration of the first quarter of 2000 to the second quarter of 2021 was first recorded. To reduce the bias in time period selections, the value of frequencies for each keyword was normalised using (3.4.1), which corrects all data in the time series to a common scale. As in (3.4.1), NF is a normalised value that is computed as the frequency of the given word frequency (WF) divided by the total number of literatures from certain period i. Figure 3.14 shows how the frequencies change with respect to the number of literatures per year.

$$NF_i = \frac{WF_i}{total\ no.of\ literatures} \times 100 \tag{3.4.1}$$

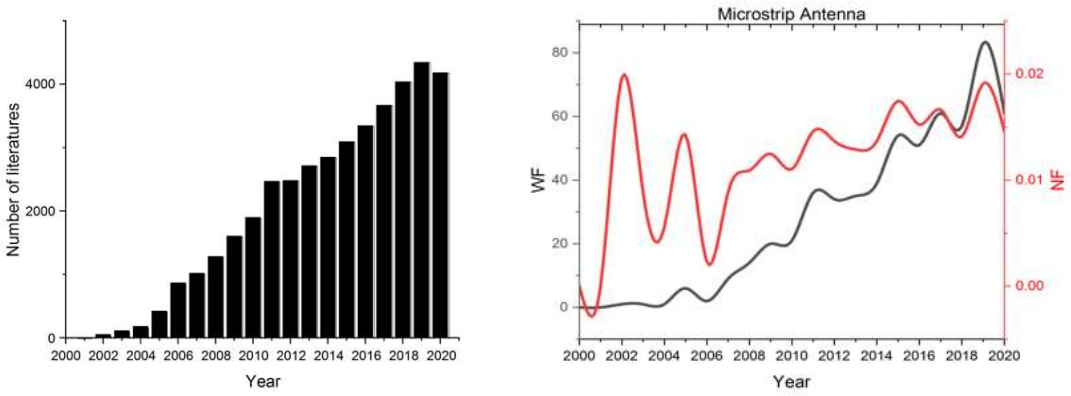


Figure 3.14 The bias reduction of frequencies per published year.
The number of papers published varies annually. Therefore, we adjusted the word frequency (shown by the black line) based on the number of publications each year, resulting in the normalized values (represented by the red line).

Figure 3.15(a) demonstrates that the NF range (*i.e.* y-axis) of every keyword is different. For direct comparisons, this difference was normalised using Min-max scaler that makes all the features to be transformed into a given range, $[-1, 1]$ as shown in Figure 3.15(b). The scaled frequency (SF) is defined by (3.4.2). All the WFs in KP were normalised into SFs in

order to have the same y-axis scale. Using this normalised data, which keyword has an increasing trend, a decreasing trend, or an emerging trend could be identified. After dividing the whole 20-years period into 4 phases evenly, the increasing trend keywords could be selected if the keyword of interest gives a positive value using the equation (3.4.3).

$$SF = \frac{NF - M \quad (NF)}{Max(NF) - Min(NF)} \tag{3.4.2}$$

$$\frac{\sum_{i=2016}^{n} SF_i}{n} - \frac{\sum_{i=20}^{n} \quad SF_i}{n}, where\ n = 5 \tag{3.4.3}$$

The average SF value in phase 4 (years of 2016–2020) should be greater than the average SF value in phase 1 (years of 2001–2005) if the occurrence of the keyword of interest in the literature grows. Contrarily, it gives a negative number if the occurrence of the keyword interested in phase 1 is greater than in phase 4. In this case, the keyword is considered as a decreasing keyword. Similarly, an emerging trend keyword was defined by looking at this resulting polarity of phases 3 and 4 of the words first appeared since 2010. As seen in Figure 3.15(c), emerging keywords generally show sudden increase during a short period of time (*e.g.* 2015–2020). The high frequency keywords generally give the highest or near highest NFs throughout the observed period.

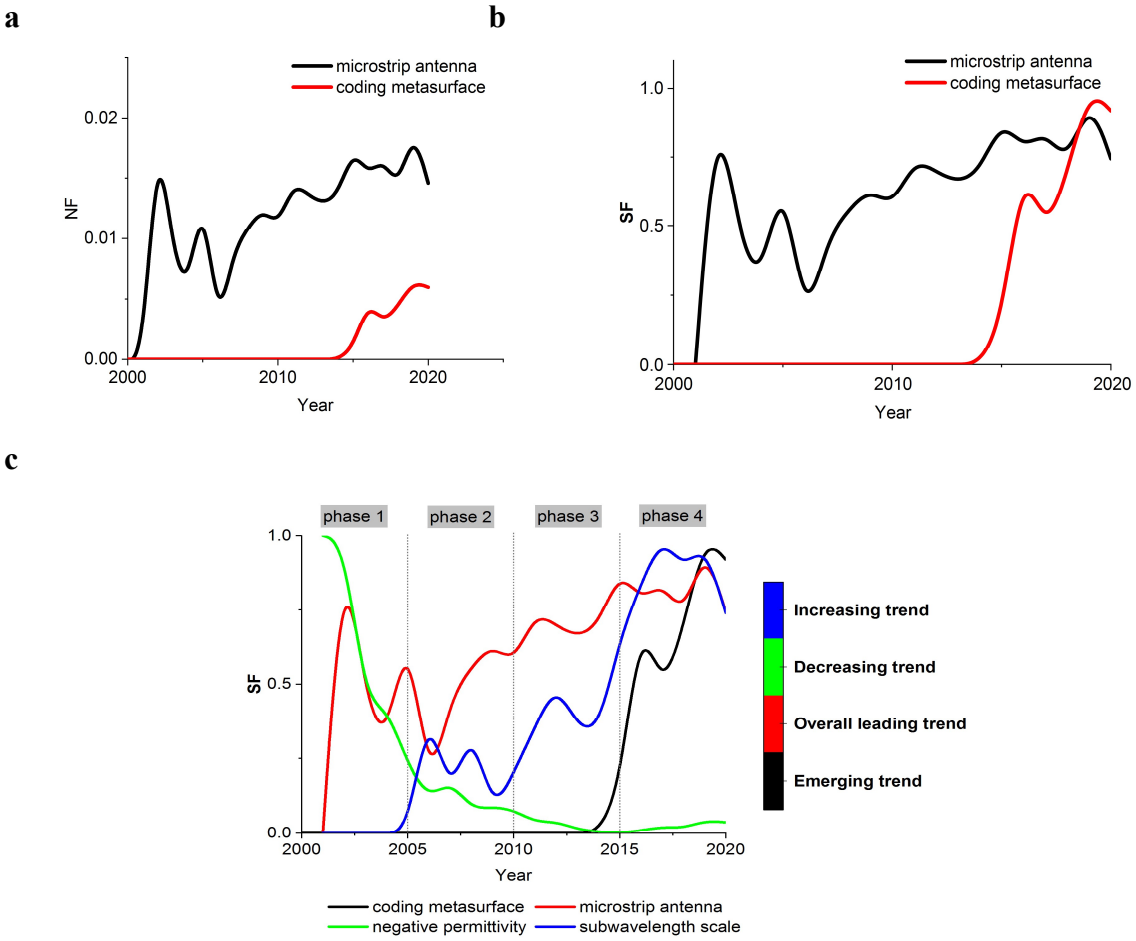


Figure 3.15 The bias reduction of keywords frequencies and the four different types of trends. **a** When a high frequency word (*e.g.* microstrip antenna) is compared to a low frequency word (*e.g.* coding metasurface), the changes in the low frequency word seems minor due to large changes in the high frequency word. **b** To address the issue in b, min-max scaler is utilised (*e.g. SF*). As both words are now compared on the same y-axis, the trend changes are easily compared. **c** One keyword for each type of trend is plotted using SF.

3.5 Keywords Analysis

To understand the past and current trends of keywords in metamaterial research, each of the increasing, decreasing, leading, or emerging trend, together with our own reasoning was presented. Table 3.3 shows that the 20 most notable keywords in four different types of trends since the year of 2001. It was noticed that keyword frequencies of 'plasmonic nanoantenna', 'graphene' and 'terahertz wave' have increased steadily over the past 20 years. On the other hand, the keywords such as 'negative refraction' and 'effective permeability'

fall under the typical decreasing trend. The keywords such as ‘coding metasurface’, ‘2D materials’ and ‘additive manufacturing’ have emerged over the recent 5 years truly reflecting the current research trend. Figure 3.16 visualise a 100% stacked area chart for overall leading keywords and heatmaps for the other three trends.

Table 3.3 The 20 most notable keywords in four different types of trends.

	Increasing trend	***Decreasing trend***	***Overall leading trend***	***Emerging trend***
keywords	*plasmonic nanoantenna*	*negative permittivity*	*resonator*	*engineering structure*
	modulation depth	*negative refraction*	*bandwidth*	*nanoporous gold*
	surface current	*rectangular waveguide*	*ring resonator*	*coding metasurface*
	chiroptical response	*composite medium*	*refractive index*	*nonlinear metasurface*
	subwavelength scale	*ring resonator*	*unit cell*	*doppler radar*
	stealth technology	*magnetic permeability*	*transmission line*	*kirigami structure*
	polarization conversion	*effective permeability*	*split ring*	*interfacial layer*
	remote sensing	*resonance cone*	*magnetic field*	*absorber sensor*
	slow light	*dispersion equation*	*metasurface*	*2d material*
	graphene	*grounded slab*	*negative index*	*dielectric metasurface*
	polarization sensitive	*dispersion characteristic*	*electromagnetic wave*	*lagrangian model*
	high gain	*layered structure*	*frequency range*	*mechanical displacement*
	perfect absorption	*thin wire*	*wave propagation*	*sensing characteristic*
	engineering structure	*FDTD method*	*graphene*	*acoustic metasurface*
	quantum emitter	*subwavelength focusing*	*numerical simulation*	*auxetic behaviour*
	circular ring	*transmission line*	*photonic crystal*	*angle insensitivity*
	local resonance	*negative permeability*	*negative permittivity*	*3d printing*
	radiation force	*superlens*	*negative refraction*	*spinodal decomposition*
	optical response	*enhanced transmission*	*laser*	*chiroptical response*
	terahertz wave	*negative index*	*surface plasmon*	*shock isolation*

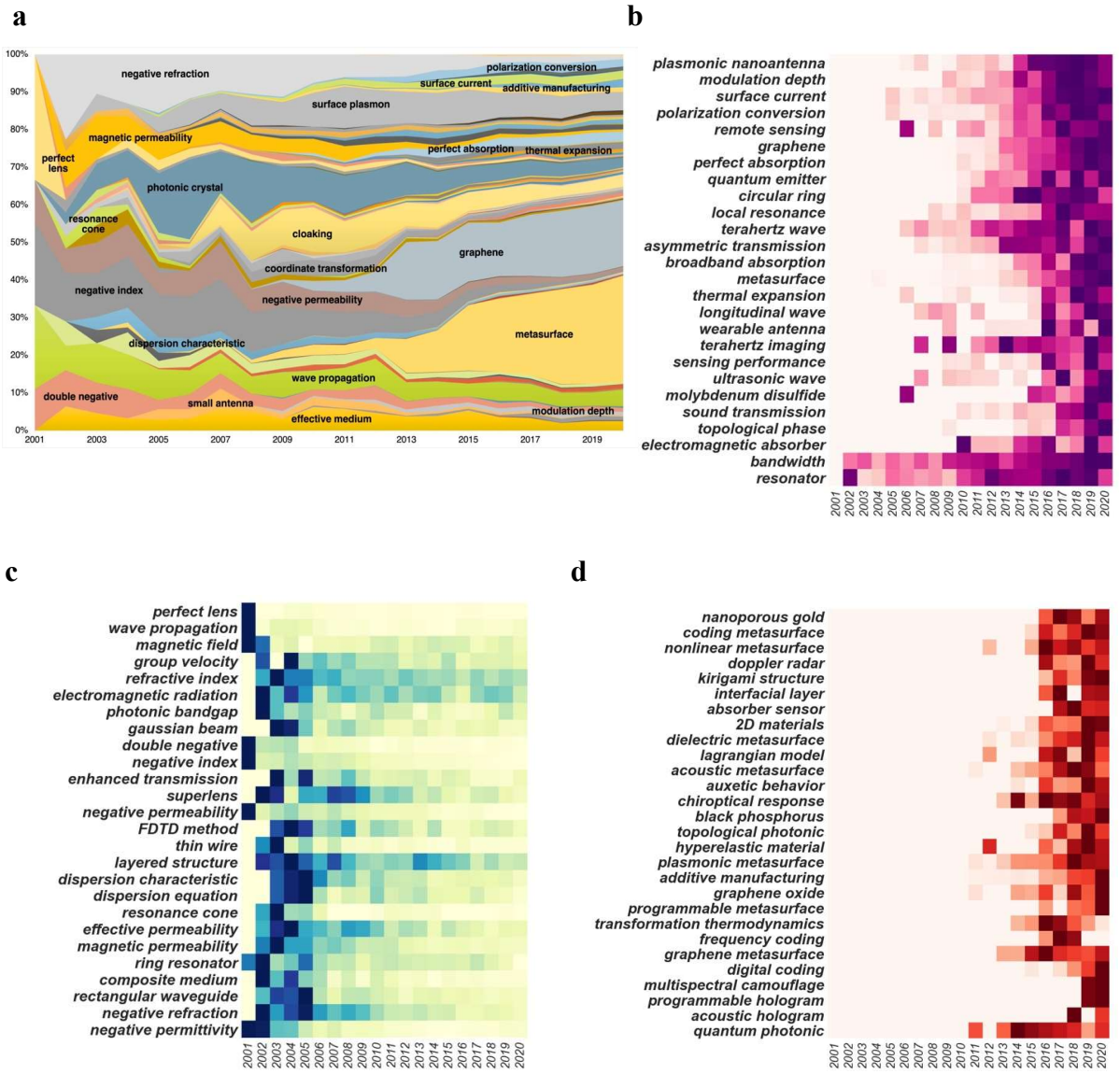


Figure 3.16 The 100% stacked area chart and heatmaps for each trend.
a This 100% stacked area chart shows how high frequency keywords (the overall leading tend) of a whole KP have changed over 20 years. The y axis scale is 100%. Each area of colour represents one keyword of the whole KP. The parts are stacked up, vertically. The height of each coloured stack represents the percentage proportion of that keyword at a given point in time. **b Increasing trend, c Decreasing trend, d Emerging trend:** Each heatmap shows a 2D graphical representation of the scaled frequency (SF) of each keyword in a data matrix. Each cell reports a numeric frequency; however, the numeric frequency is accompanied by a colour, with larger frequencies associated with darker colourings.

To analyse the relationships between the labelled clusters and each type of trends, one hundred keywords were selected showing the greatest SF difference from each type of trend by counting the number of keywords in its corresponding clustering category. Figure 3.17

shows that the words in the "modelling and optimisation (blue)" cluster have the majority portion in the overall leading and decreasing trends. Figure 3.17 illustrates the keywords which are related to "applications (grey)" and "fabrication and enabling materials (green)" have large proportions in the trends of emerging and increasing. This indicates that the number of studies which gained popularity in the early era of metamaterial research (*e.g.* keywords in "modelling and optimisation" as well as "optical properties" categories) has decreased while the interests in translating metamaterials into practical applications (*e.g.* keywords in "fabrication and enabling materials" and "applications" categories) have increased. Besides, the cluster of "antennas" shows less than 10% in all trends. Figure 3.18 visualises 30 keywords with the emerging trend in 8 clusters, each of which shows the most popular research area over the recent years in each labelled category.

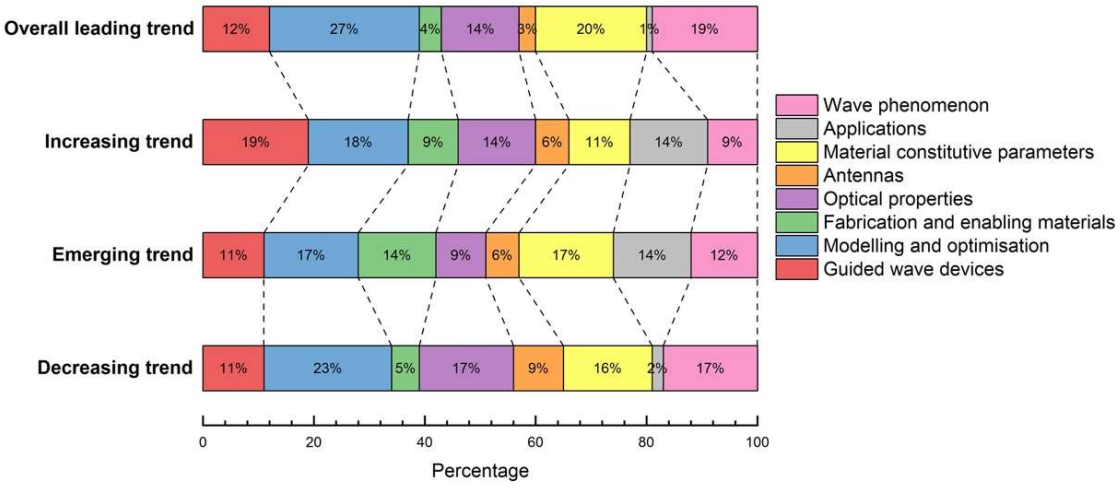


Figure 3.17 The number of keywords from each cluster in 4-different type of trends.

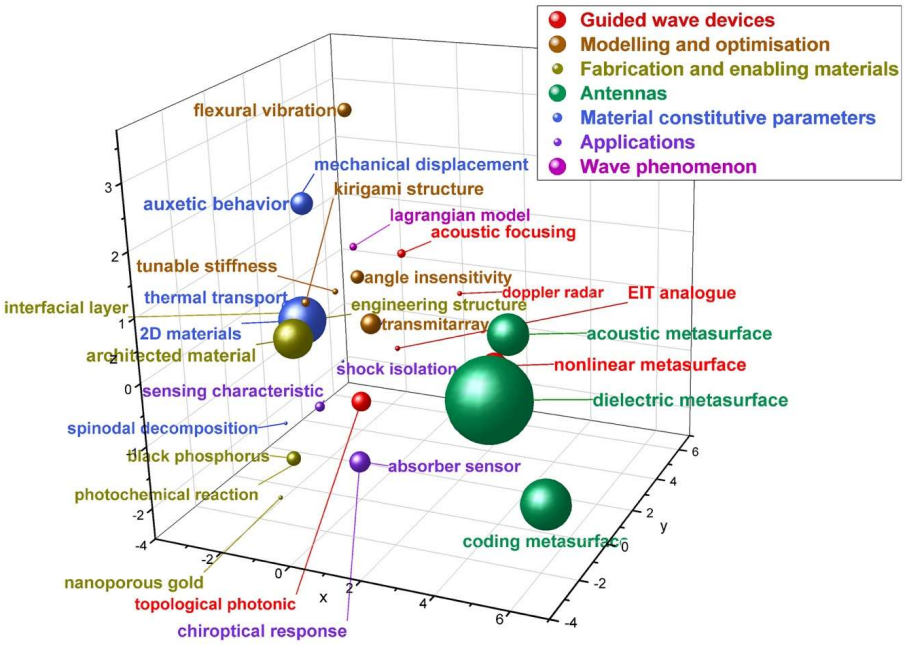


Figure 3.18 The 30 emerging trend keywords from each cluster and their extent.

This 3D plot shows 30 words with the largest increase in frequency over the past 5 years. The size of sphere represents the frequency growth rate. Based on the size of sphere, the emerging keywords could be identified. Emerging trend keywords are ‘dielectric metasurface’, ‘coding metasurface’, ‘2D materials’ and ‘acoustic metasurface’. As its colour represents, each of these keywords belongs the cluster of antenna and material constitutive parameters.

3.6 Keyword Forecasting

The NFs were now fed into a sequential memory-based ML algorithm for the prediction of the frequency of each word. To date, many different types of time-series analysis algorithms have been proposed. Methods for time series analysis may be categorised into two classes, namely, frequency-domain and time-domain methods. The former includes wavelet analysis [88] while the latter comprises autocorrelation [89]. The approach for time series analysis

could also be categorised into parametric and non-parametric methods. Parametric methods rely on assumptions about the shape of the distribution in the underlying population while non-parametric methods presume that the data distribution cannot be defined in terms of such a finite set of parameters thus take flexible and model-free approaches. This study uses the LSTM, a time-domain and non-parametric method given the complex and temporal nature of input data. We believe that the future research trend could be predicted via the analysis of predicted keyword frequency and its rate of changes in a sequential format. The encoder-decoder architecture of LSTM was thus adopted for this task.

3.6.1 The Architecture of Encoder-decoder LSTM

The model consists of three main components, namely, encoder, intermediate (encoder) vector and decoder. Encoder and decoder use a multi-layered LSTM unit to map the input sequence to a vector of a fixed dimensionality, and then another LSTM is used to decode the target sequence from the vector [28], [90], [91]. These are trained using the input data while maximising the conditional probability of the target sequence for a given sequence. As the data used in this study is in a time sequential format, the current keyword frequency has been predominantly calculated from the legacy data in the past years. A many-to-one architecture model was designed so that it was able to look back several sequences and predict a new sequence ahead. If annual data was utilised removing outliers from the years 2000 and 2001, it was left with only 19 data points from the years 2002 through the second quarter of 2021. This results in a shortage of data for predictions four or five data points out.

To avoid the data scarcity for training the model, especially for new keywords from new emerging research in metamaterials, the data was computed quarterly thus there are in total of 78 data points until the second quarter of 2021, instead of using the frequency of keywords calculated on the annual basis (*i.e.* 19 data points). Figure 3.19 illustrates how the

data points for the keyword ‘cloaking’ change when the frequencies are gathered by year, half-year, and quarter of year, as well as the outcomes of the model validation in accordance with data point changes. As noticed from Figure 3.19, there is a trade-off between the quantity of data points and the prediction precision with the scaled frequencies. Despite the noise in the quarterly data increasing, a high number of data sequences is undoubtedly beneficial for an accurate LSTM model learning process. Thus, frequency data by quarter is gathered; for instance, if articles from the year 2018 were published, the abstracts of articles are divided into four groups by three months. The frequency of each group is then calculated in turn.

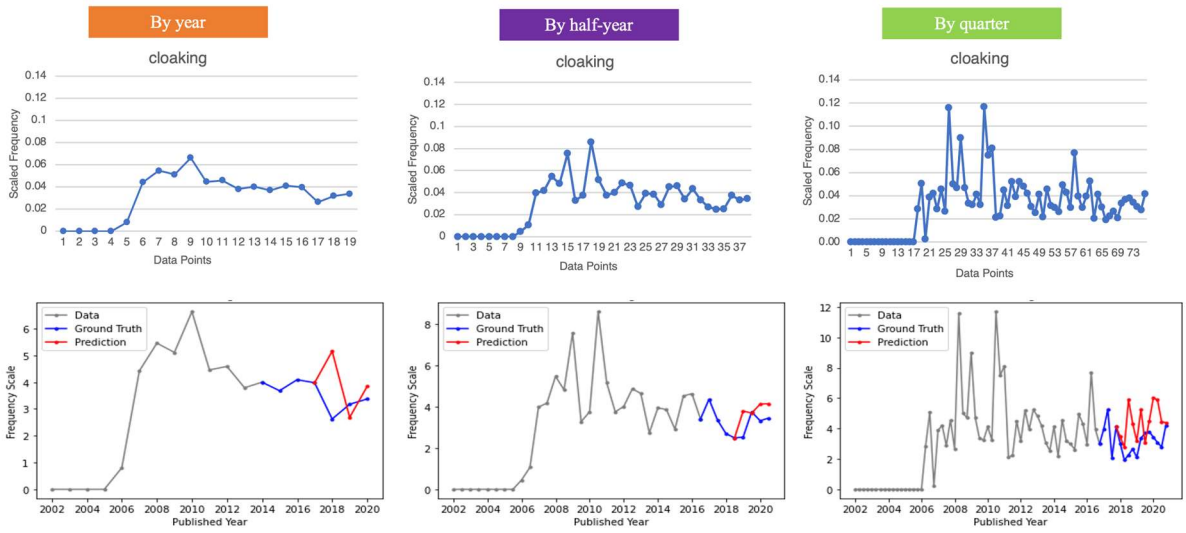


Figure 3.19 Data points augmentation and the results of model validation of ‘cloaking’.

The calculated 78 data points were then used as an input to the many-to-one encoder-decoder model as shown in Figure 3.20. The learned model was fed with an input data of a sequence of every 6 frequencies in the form of sliding window for predicting the next data point. To predict the research trend in the next 4 years, the above process was repeated 16 times (*i.e.* quarterly data points of 4 years). The modified architecture of the encoder-decoder

model has two LSTM units (*e.g.* encoder and decoder) and each unit is composed of 300 layers with rectified linear unit (ReLU).

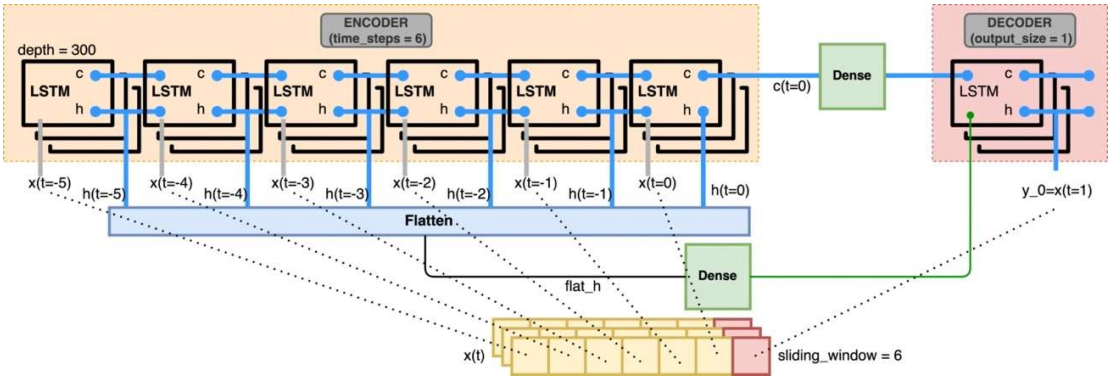


Figure 3.20 The architecture of encoder-decoder LSTM.
The architecture has 6 sliding window input LSTM cells and 1 output LSTM cell. Every cell is connected via cell state (c) and hidden state (h). Each LSTM cell in Encoder block is designed to have the depth of 300. The encoder block is connected to a decoder block via dense layers.

3.6.2 Model Validation

Given the small number of sequences, the LSTM models were trained with the objective of minimising Mean Square Error (MSE) which is defined by (3.6.1).

$$MSE(\hat{x}_{T,h}) = E\left[\left(x_{T+h} - \hat{x}_{T,h}\right)^2\right] \tag{3.6.1}$$

The optimal h-step-ahead forecast of x_{T+h} at time T is the conditional expectation, $E[x_{T+h}|\Omega_T]$. Therefore, if $\hat{x}_{T,h}$ is any h-step predictor at time T, it follows as (3.6.2).

$$MSE(\hat{x}_{T,h}) \geq MSE(E[x_{T+h}|\Omega_T]) \tag{3.6.2}$$

To build an optimal model, the MSE between the predicted sequence and the target sequence has been minimised [92], [93]. Figure 3.21 compares the predicted values with the ground truth. The model trained with the sequence data of keyword 'metasurface' has 0.274 of MSE while the keyword 'small antenna' has 0.108 of MSE. The average MSE that is computed

over 91 selected keywords is 5.9e-0.2 only, which have validated the model with good accuracy.

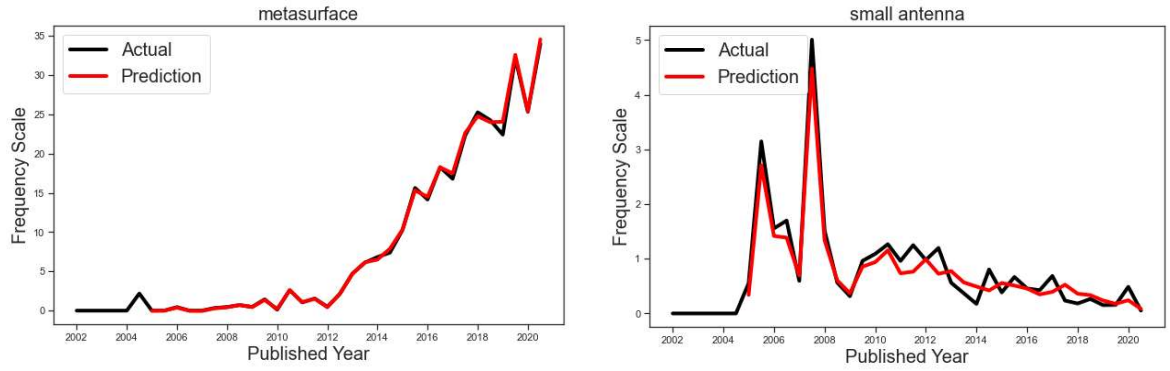
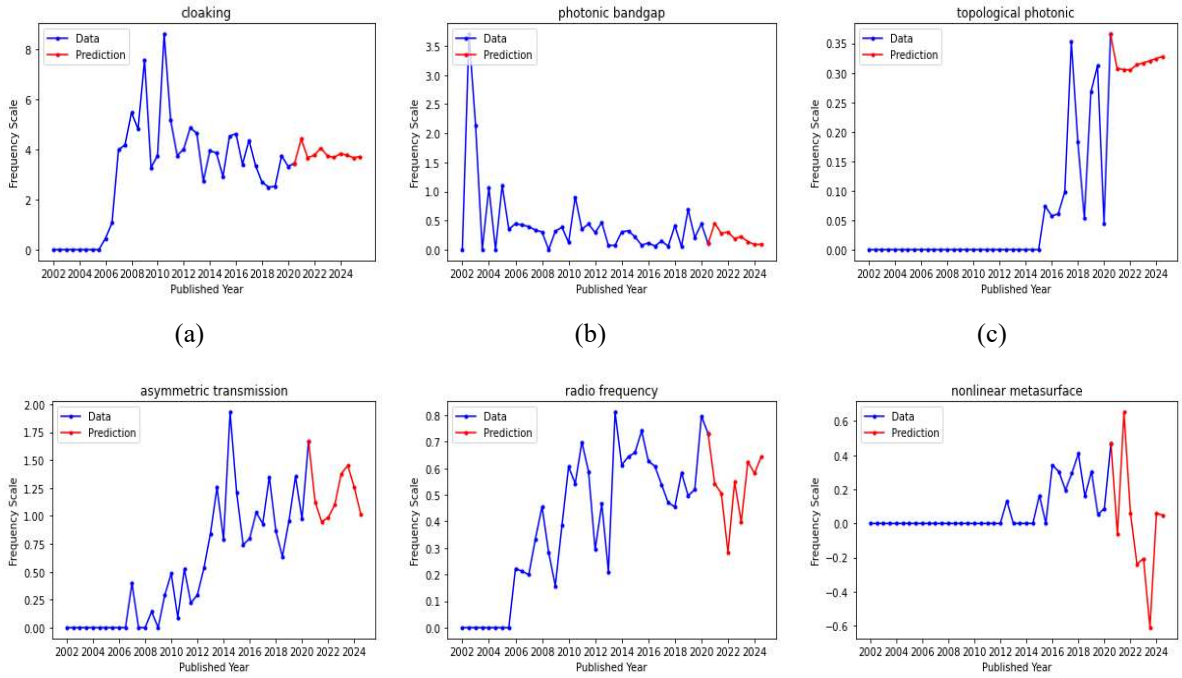


Figure 3.21 The validation results of encoder-decoder LSTM models.

3.6.3 Prediction Results

Using 78 data points (from the first quarter 2002 to the second quarter of 2021), the learned model forecasts 16 data points ahead (from the third quarter of 2021 to the second quarter of 2025). The prediction outcomes of several keywords are provided in Figure 3.22. As you can see, the model worked well for the keywords in (a), (b), (d) and (e) however, it seemed not working well for the keywords in (c) and (f). It was noticed that if frequencies contain a lot of '0' data, the model cannot be trained properly and seems to have just half the training data, which increases the probability that it would predict the wrong things. On newer or emerging keywords, this kind of pattern is frequently observed. In Appendix A Section A.4, there are further examples of emerging keywords that cannot be effectively trained. The training parameters and the number of sliding windows were thus fine-tuned since the frequencies contain a lot of data that is "0", preventing the model for these keywords from correctly learning from historical data.



(a) (b) (c)

(d) (e) (f)

Figure 3.22 The prediction results of encoder-decoder LSTM models.

To reduce the impact of fluctuating noise from the quarterly data, a moving average was applied to the current prediction process. The moving average method, which is popularly used with time series data to smooth out short-term fluctuations and highlight longer-term trends or cycles, is a straightforward but effective statistical technique for analysing data points by averaging a number of different subsets of the full data set. Let time series data-points be $p_1, p_2, \ldots, p_n$. The mean over the last k data-points is denoted as MA and calculated as:

$$MA = \frac{p_{n-k+1} + p_{n-k+2} + \cdots p_n}{k}$$

$$= \frac{1}{k} \sum_{i=n-k+1}^{n} p_i$$

Figure 3.23 shows the outcomes of the keyword prediction using the moving average method.

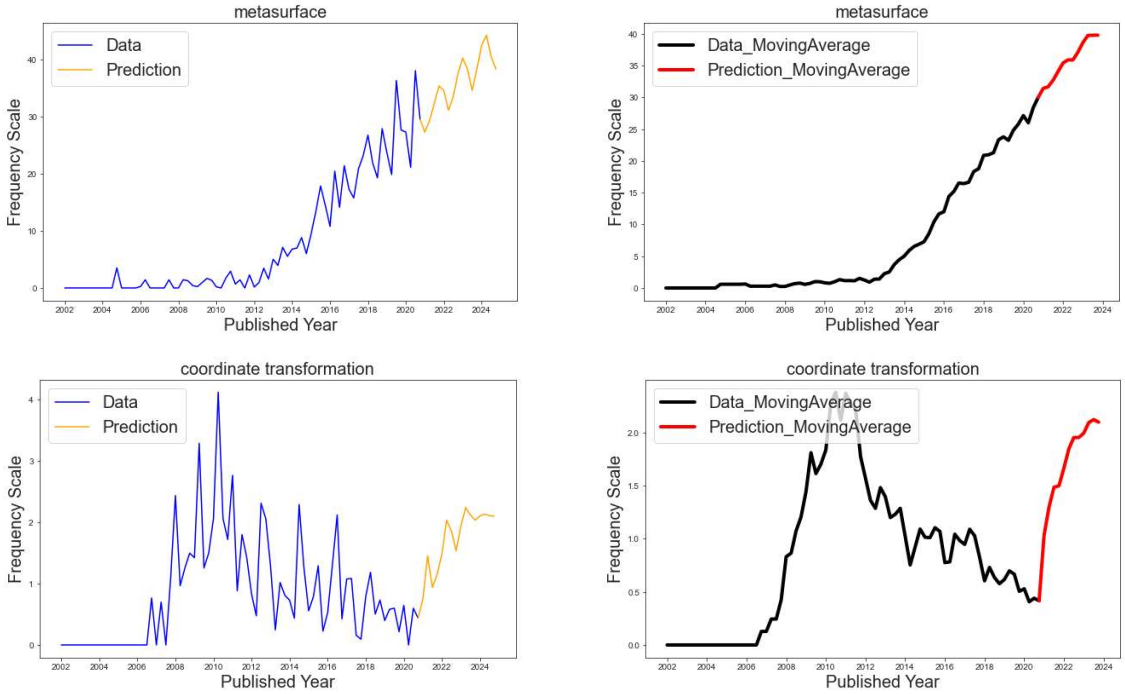
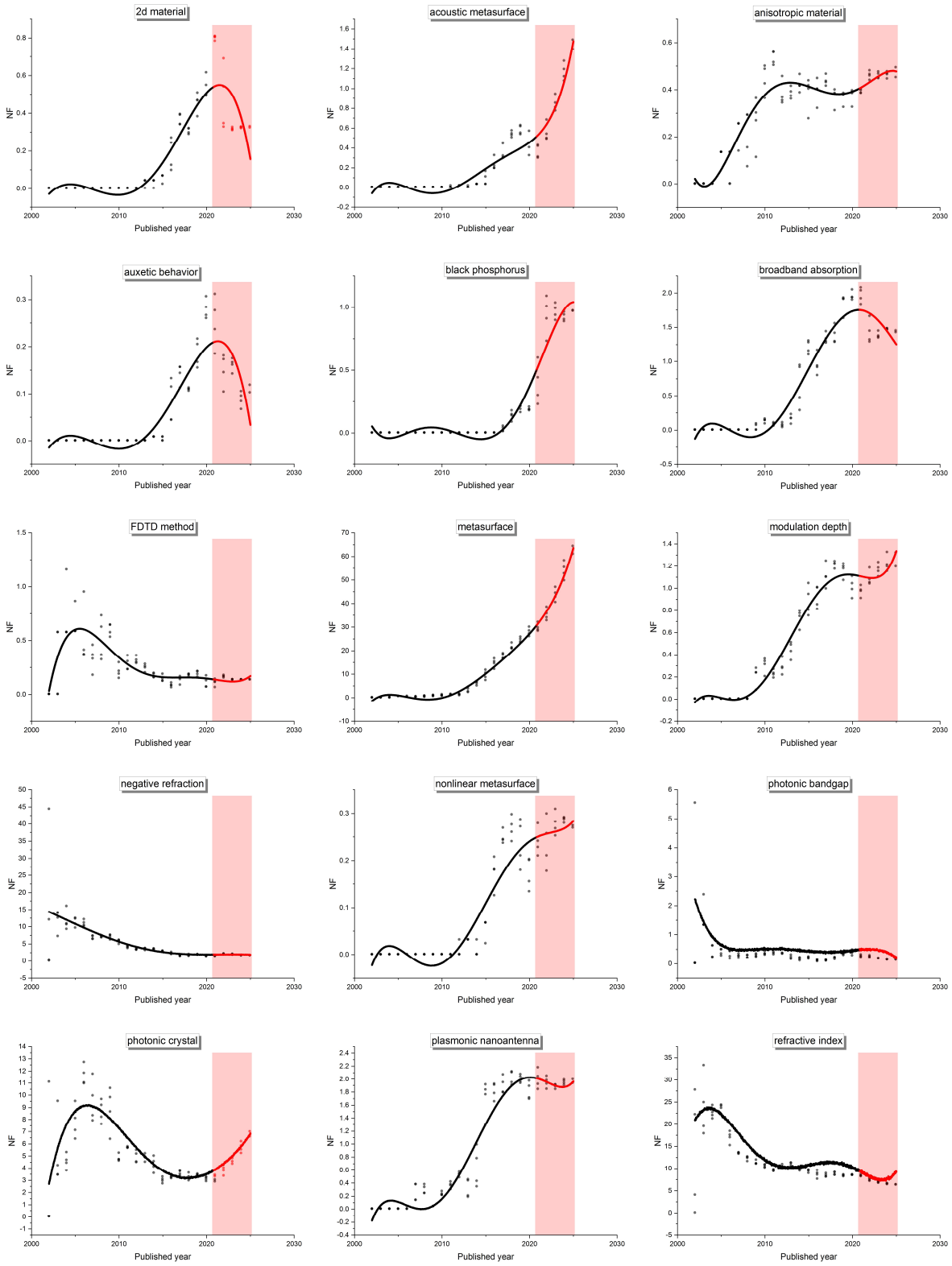


Figure 3.23 The results of the comparison using the moving average method.

Another attempt for reducing the fluctuated noise, a polynomial fit was performed with 6 orders which gives the improved visualisation of increasing and decreasing trends. From the preliminary experiments of this study, it was found that polynomial regression gives the most accurate approximation of the relationship between the dependent and independent variables. All in all, as in the equation below, the expected value of y as an n^{th} degree polynomial is modelled yielding the general polynomial regression.

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \cdots + \beta_n x^n + \varepsilon$$

With the 6th degree polynomial (n = 6), the 20 keyword predictions in Figure 3.24 are obtained.


2d material
acoustic metasurface
anisotropic material
auxetic behavior
black phosphorus
broadband absorption
FDTD method
metasurface
modulation depth
negative refraction
nonlinear metasurface
photonic bandgap
photonic crystal
plasmonic nanoantenna
refractive index
NF
Published year

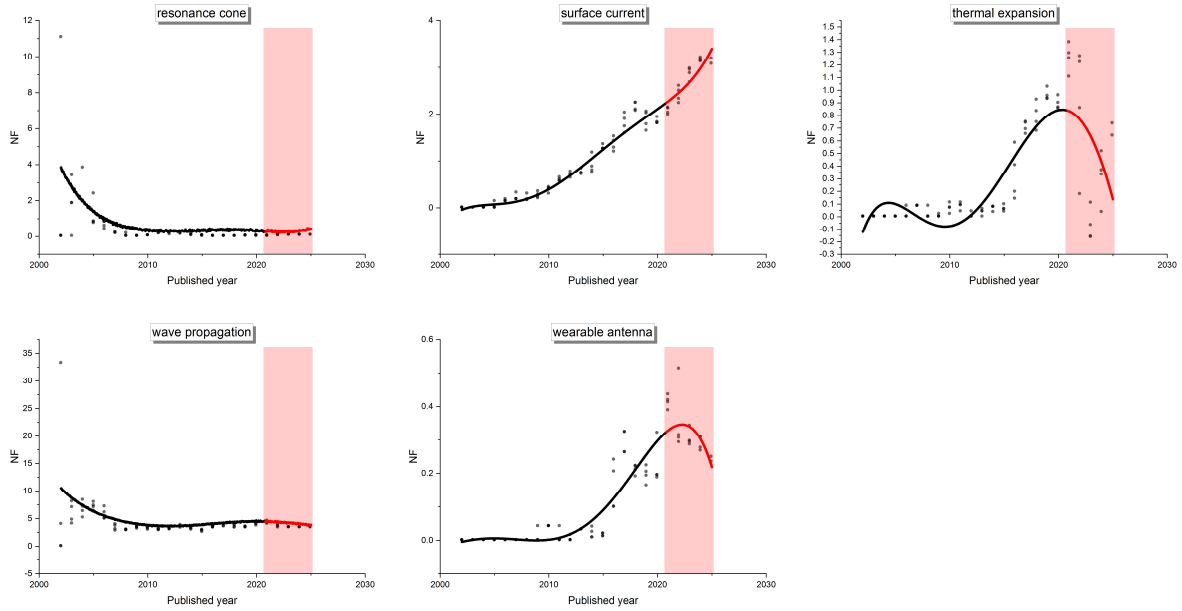


Figure 3.24 The results of prediction for next 4 years using encoder-decoder LSTM models.

The validated models were used for 16 steps ahead predictions for the period of the third quarter of 2021 to the second quarter of 2025 based on the data collected for the duration between the first quarter of 2002 and the second quarter of 2021. As a result of prediction, the frequencies of appearance for every keyword in the KP were estimated. The prediction results are also shown in Figure 3.25 along with a polynomial fit that has a 95% prediction and 95% confidence band. In statistical analysis, a confidence band is used to depict the level of uncertainty in a curve or function estimate that is based on sparse or noisy data. The uncertainty about the value of a new data point on the curve, which is vulnerable to noise, is similarly represented by a prediction band. Bands of confidence and predictions are frequently utilised in the graphical display of regression analysis results.

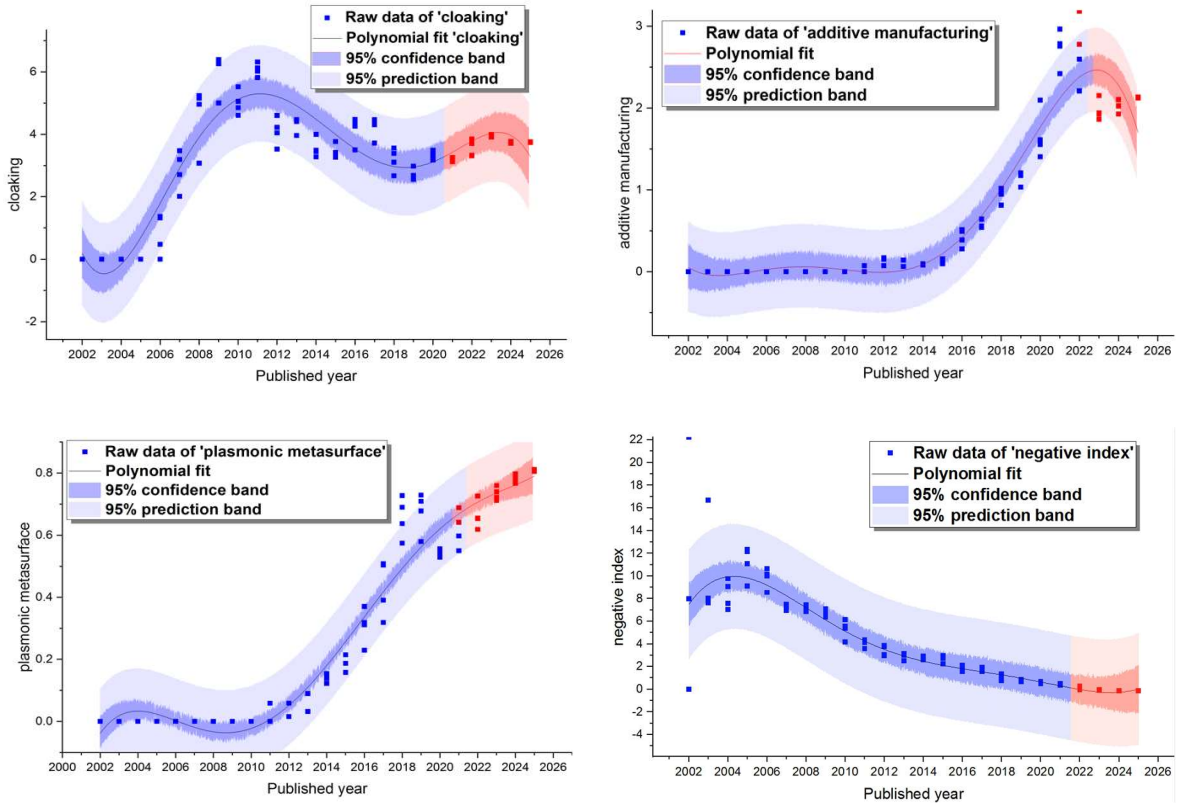


Figure 3.25 The results of prediction with 95% prediction and 95% confidence band.

The trend of scientific research and technology development often follows a S curve [94], which measures the effort (*e.g.* time and money) on its x-axis against technical performance on its y-axis. It means that at the beginning of technology lifecycle, there is a great deal of investment with relatively little performance improvement. However, when it reaches some tipping point, its performance increases significantly, while towards the end, the performance becomes less distinguishably improved and reaches a plateau.

Gartner's Hype Cycle, a graphical depiction of a common pattern that arises with each new research and technology, was proposed in 1995. It graphically visualises the technology lifecycle, however, in different dimensions with time on its x-axis and expectations on its y-axis. Its emphasis is particularly on the expectations of technical performance in the marketplace, and it is generally accepted that although it focuses on specific technologies, the same pattern of hype and disillusionment applies to more high-level

concepts [95]. The Hype Cycle usually starts when an event generates public interest in a technology innovation. The expectation increases dramatically after the stages of technology trigger (TT) and peak of inflated expectation (PI). In the phase of trough of disillusionment (TD), as many experiments and implementations fail to deliver, expert groups react negatively. The technology is then more widely understood so that more industries fund pilots cautiously in the phase of slope of enlightenment (SE) followed by the phase of plateau of productivity (PP) when is the mainstream adoption starts to take off. It is undoubtable that the hype cycle also applies to the research of metamaterials, it would be therefore of great help in setting a correct path for future research ensuring it to be efficient and cost effective. It was noticed, however, that in metamaterial research, some research topics do not slide into SE or PP phases, however, they rather disappear. When the inflated expectations begin to die down via the phase of TD, they start to decrease, and this trend continues over the phases of SE and PP as indicated with a red line in Figure 3.26. Among the four trend types that this study defined, it is obvious that emerging and increasing keywords fall into TT and PI phases. However, decreasing keywords should be divided into the words sliding into SE and PP phases (continue following blue line) or the words disappearing (following the red line) based on its prediction results. The order of each keyword in the Hype Cycle is decided within each phase, based on the difference of corresponding SF values.

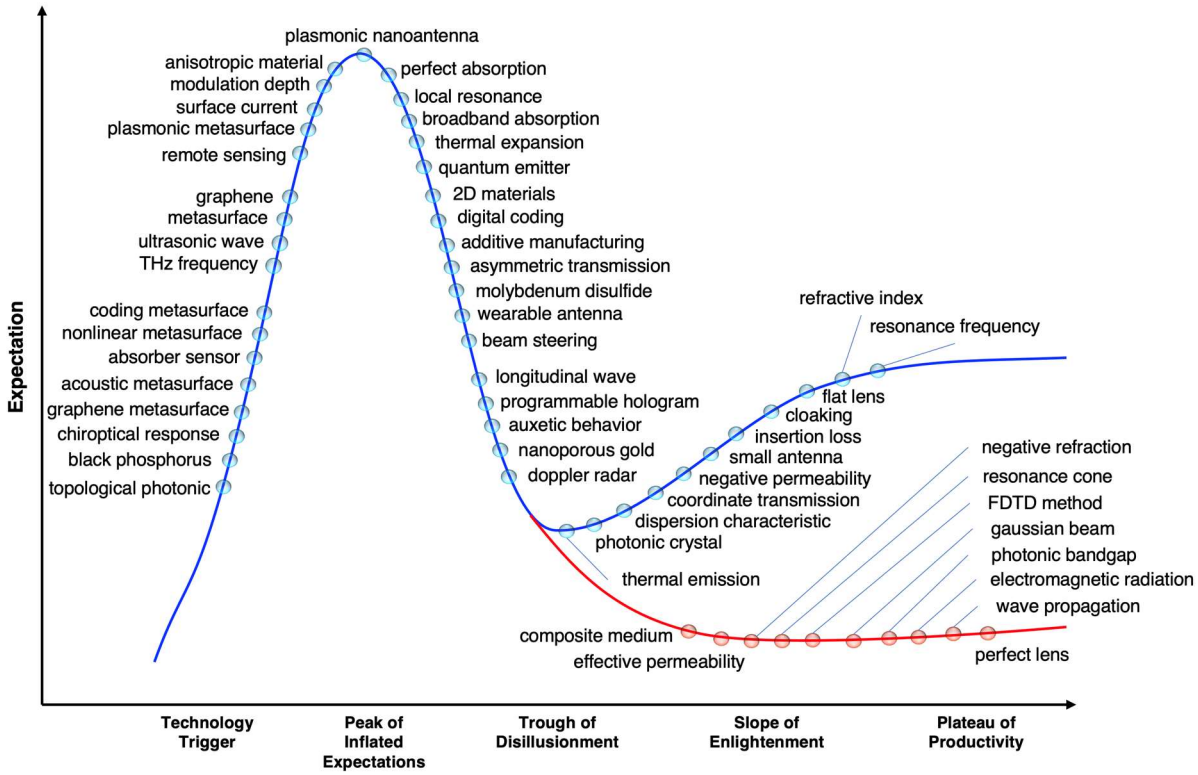


Figure 3.26 The results of prediction using the hype cycle in year of 2025.

On the basis of the preceding fully automated analysis of the past metamaterial research trend and its forecasting, it can be said that, for the past 20 years, we have witnessed substantial effort in translating metamaterial research from fundamental studies to engineering practices including industrial applications, large scale manufacturing and multi-material integration etc. Aligning with the argument of Zheludev [71], the research on 'graphene', 'plasmonic metasurface' and 'nonlinear metasurface' has also gained the significant attention among metamaterial researchers. Although he stated in his article [71] that the switchable metamaterials will bring major benefits via material's properties (*e.g.* vanadium oxide (VO_2)), its role and significance in coding metasurface or programmable metasurface to harness the power of both computer science and metamaterials were not anticipated. Most of relevant research is still based on the use of III-V semiconductor-based diodes and varactors. On the "forbidden fruit" of negative-index media, although the concept

was widely accepted in the scientific community, research activities on ‘negative refraction’, ‘double negative media’ and ‘negative index media’ have significantly decreased [96]. It is partially due to the difficulty in fabricating low loss materials and demonstrating ground-breaking applications such as ‘perfect lens’, and, meanwhile, ‘metalens’ consisting of millions of meta-atoms with nanoimprint lithography process has been successfully demonstrated and commercialised [97].

Another cornerstone of metamaterial research is ‘transformation optics’, *a.k.a.* ‘coordinate transformation’, where the concept of ‘cloaking’ [98] has drawn significant attention from the public and academic community. Despite of being an important mathematical tool in the design of electromagnetic devices, and, more broadly any device with operating principles governed by partially differential equations, the volume of its research in metamaterials has never been significant, and also seen the decreasing trend. Coupled tightly with the concept of ‘transformation optics’ is ‘cloaking’, which still remains to be an active subject of research [99], far from being “ripe” as predicted in article of Zheludev [71], due to the complexity of material designs and inherent loss/bandwidth limitations, leading to plateau of productivity (Figure 3.24). Technological trigger or user pushing may be needed to regain some momentum of basic research on ‘negative refractive index’ materials and ‘perfect lens’. ‘nonlinear metamaterials’, ‘amplifying metamaterials’ and ‘switchable metamaterials’ can be broadly classified as ‘active metamaterials’ have seen continuous growth in subject areas ranging from microwave to optics in topical areas of ‘asymmetric transmission’, ‘broadband absorption’, and etc [100]. They are largely driven by applications in all optical modulation and switching enabled by silicon and graphene based nanophotonics. Metamaterials have their uniqueness to manipulate polarisation states, modulation depths, and absorption of propagating waves as well as radiation from plasmonic nanoantennas.

The development of ‘quantum metamaterials’ is accelerated by a vast amount of recent new funding invested by major agencies worldwide, in line with the growth of strong research activities in ‘quantum emitter’, ‘topological phase’ and ‘superconducting’ etc. ‘sensor metamaterials’ maintain a strong growth in areas such as ‘remote sensing’ and ‘THz imaging’. Two distinctive subjects missing from the “Knowledge of Trees” [71] are ‘acoustic metamaterials’ and ‘thermal metamaterials’, both of them witness a strong growth in topics around ‘thermal expansion’, ‘sound transmission’, ‘transformation thermodynamics’ and ‘ultrasonic waves’, especially, for the latter, the focus of study shifts from local resonances to broadband sensing performance, and the transmission of longitudinal waves. ‘mechanical metamaterials’ such as those based on ‘Kirigami structures’ have a capability of being hyper-elastic and possessing on-demand auxetic behaviour. Three key topics, namely, ‘metasurfaces’ [101] , ‘graphene’ [102] , and ‘surface plasmon’ dominate the current metamaterial research, it can be evident by the fact that “new” emerging subjects emerge in ‘nonlinear metasurfaces’, ‘plasmonic metasurface’, ‘quantum photonics’ and ‘topological photonics’, all of which are enabled by “wonder” materials such as graphene, graphene oxide, black phosphorus and nanoporous gold etc. ‘programmable metasurface’ and ‘coding metasurface’ have drawn attention from microwave and communication engineers, with analogical terms such as ‘reflecting intelligent surface (RIS)’ etc., which are largely made of conventional PIN diodes controlled via FPGA (field programmable gate array), a semiconductor-integrated circuit where a large majority of the electrical functionality inside the device can be changed.

Discoveries of novel tunable and phase changing materials are urgently needed and accelerated by machine learning and artificial intelligence [103] . Hologram metasurfaces seem to be common for all spectra for applications including camouflage, even for acoustic waves. Finally, ‘additive manufacturing’ ranging from 3D/4D printing to roll-to-roll (R2R)

processing have become a key technological enabler, which sustains the growing trend of metamaterial research. Looking into future, for the next four years, we will continue to see the growth in studies of 'metasurfaces' including 'plasmonic metasurfaces' (Figure 3.25), 'acoustic metasurfaces' and 'nonlinear metasurfaces' (Figure 3.24). The popularity of 'cloaking' and 'additive manufacturing' studies will reach its peak of recent cycle and start to decline (Figure 3.25). It is interesting that, despite of a declining trend overall in '2D materials' research, studies on 'graphene' and 'black phosphorus' will remain strong for the next four years (Figure 3.26). Metamaterials applied to antenna applications including wearable antennas will decline but 'photonic crystals' research will regain its momentum, by focusing manipulating light in deep 3D structures consisting of a large number of nanopores etc.

3.7 Summary

This chapter proposed a state-of-the-art method that enables an automatic extraction of the information from unstructured texts, analyses of publicly available publications and prediction of future research trends in materials science. This is of paramount importance for those areas where setting a correct future research path ensures the research to be effective while reducing its cost. As it is well believed today that the most respected sources of scientific data are coming from scientific publications such as journals and patents, this chapter automatically built a KP from over 43,000 abstracts of various scientific publications and all the words in KP were vectorised into hyperdimensional spaces for a mathematical algorithm to understand the meaning of word via correct representations. The proposed keyword prediction framework not only effectively visualises the trends and relationships of each word in KP but also forecasts the future research trend in the form of hype cycle until 2025. The experimental results of this chapter demonstrated that the proposed approach is

valid and versatile, and arguably it can be applied to any research field of interest. To the extent of our knowledge, this is the first study that proposes a fully automated future keyword prediction framework in materials science and provides useful benchmark to future metamaterial research based on newly built KPs.

Chapter 4

NLP-assisted Study on Body-Centric Wireless Communications

A literature survey traditionally relies on data gathered from literature database searches. Manually sifting through such vast information can result in overlooking pivotal studies or essential findings. Utilising extractive summaries can greatly enhance the creation of scientific review papers by ensuring critical details are not missed. With this context, this chapter suggests a framework for how text summarisation, one of the main NLP tasks, can boost the thoroughness of review papers using an advanced text summarisation technique, and the validation results of the suggested approach were made available for comparison with well-referenced datasets for text summarisation. Additionally provided are a forecast of the number of publications in this field of study and the keywords trend in earlier publications.

Over the past decade, the Internet of Things (IoT) has evolved and gained considerable attention in various fields. This significant development in the area of mechanically flexible, wearable and skin electronics can accelerate the broadening of the utility of IoT technologies. Thanks to the recent development of big data and sensing technologies, the IoT devices are able to collect and share huge amount of data directly with other devices through the cloud environment, allowing vast information to be gathered and analysed for data-analytics processes [104]. This platform is used not only to make life easier for people, but also to address health and security issues that are critical for our well-being. As COVID-19 has become more pressing than ever, any type of data from human as a part of the Things is considered a great importance [105]. This enables the researchers to focus on studying the electronic skin sensors or wearable sensors. Analysis of prior publications on the methods used to conduct these studies is required to comprehend the cutting-edge technologies of the research fields. Review papers are written with the intention of identifying and synthesising pertinent literature in order to assess a particular research question, substantive domain, theoretical perspective, or methodology and finally the review papers provide readers a current understanding of the research topic [106].

4.1 Introduction

The recent development of NLP allowed for the collection and extraction of useful information from published articles (*e.g.* keywords extraction and text summarisation). For the colleting of the publications in journal websites, Elsevier API key that requires users to sign up to make web scraping a legal and valid process was used in this study. In addition, the request involves query search keywords. Most publishers provide various APIs for scientists to download and use their abstract, author's keywords, published data, author details and full

texts. Accordingly, the abstracts, published year and author's keyword for each query search keywords are downloaded from the Elsevier Developer Portal (https://dev.elsevier.com). To understand the research trend of the electronic skin sensors or wearable sensors, 'flexible electronic skin sensors', 'flexible wearable sensors', 'body-centric wireless communication', and 'wireless implant sensor' were searched using Elsevier API from 1993 to 2019. Figure 4.1 depicts the number of papers found in relation to search terms and publication years. The Elsevier API is used to download the entire abstract as well as additional information, such as publication dates, authors, affiliations, and journal names, to an Excel file.

Search Query	Published Year	Number of Papers
Body-centric wireless communications	2005-2019	376
Flexible electronic skin sensors	1998-2019	810
Wireless implantable sensor	1993-2019	1123
Flexible wearable sensors	1997-2019	3576

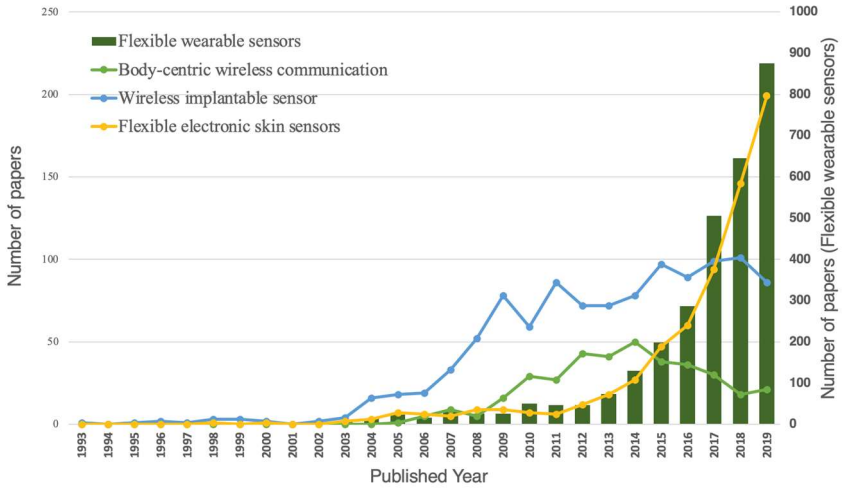


Figure 4.1 The number of papers for each search query by the year of 2019.
Since there are more papers on 'flexible wearable sensors' than on other topics, this sequence data was plotted using a different y-axis to provide a clear contrast.

Using past publication data by year, this study makes a forecast for the number of papers over the next five years. Time series data forecasting is a crucial topic in economics, business, and finance. In the past, a number of methods, including univariate Autoregressive (AR) [107], univariate Moving Average (MA) [108], Simple Exponential Smoothing (SES)

[109], and most notably Autoregressive Integrated Moving Average (ARIMA) [110] with its many variants, have been used to accurately predict upcoming time series data. Figure 4.2 depicts the results of the ARIMA model's prediction of the number of papers related to body sensing technologies over the next five years.

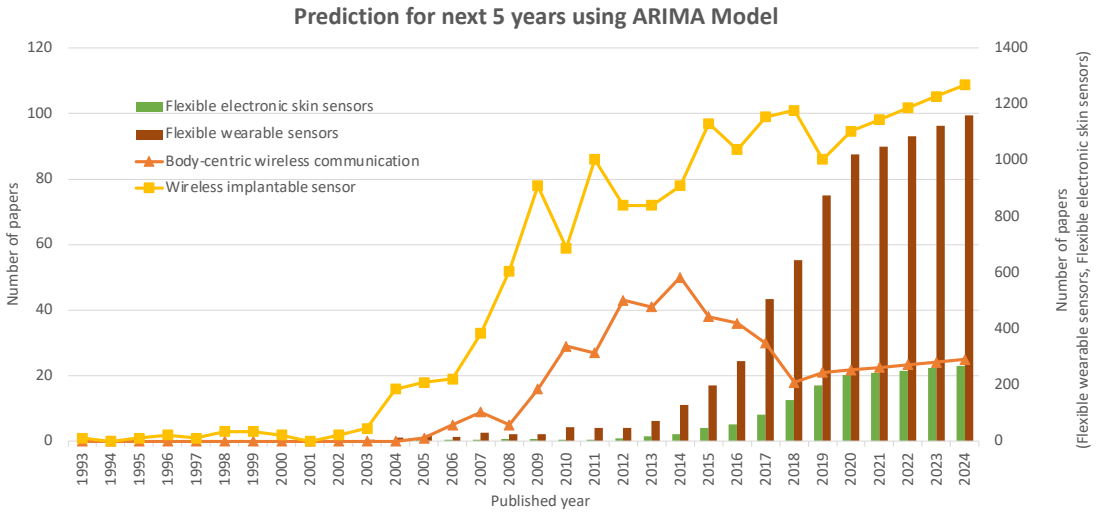


Figure 4.2 Prediction for number of publications in next 5 years using ARIMA. For comparison, the y-axis of the line graphs for 'body-centric wireless communication' and 'wireless implantable sensor' is on the left of the figure, while the y-axis of the bar graphs for 'flexible electronic skin sensors' and 'flexible wearable sensors' is on the right of the figure.

Many researchers have recently discovered deep-learning-based time series forecasting algorithms, including LSTM, outperform conventionally based algorithms, such as the ARIMA model [111], [112]. While LSTM does not call for the setting of such parameters, ARIMA requires a set of parameters (p, q, d) that must be calculated using data. To learn models from deep-learning-based algorithms, however, some hyperparameters must be tuned. The time series data was fed into the sequence-to-sequence LSTM model, which was built in Chapter 3, as input, to compare the statistical approach and the machine learning approach. The four search keywords were used to validate the LSTM model, and Figure 4.3 demonstrates the results. As seen in Figure 4.3, the y-axis range for the keywords 'body-centric wireless communication' and 'wireless implantable sensor' is narrow, which makes

the error rate seem so high. However, if each y-axis has the same scale, the difference between the ground truth and the predicted value is not as great.

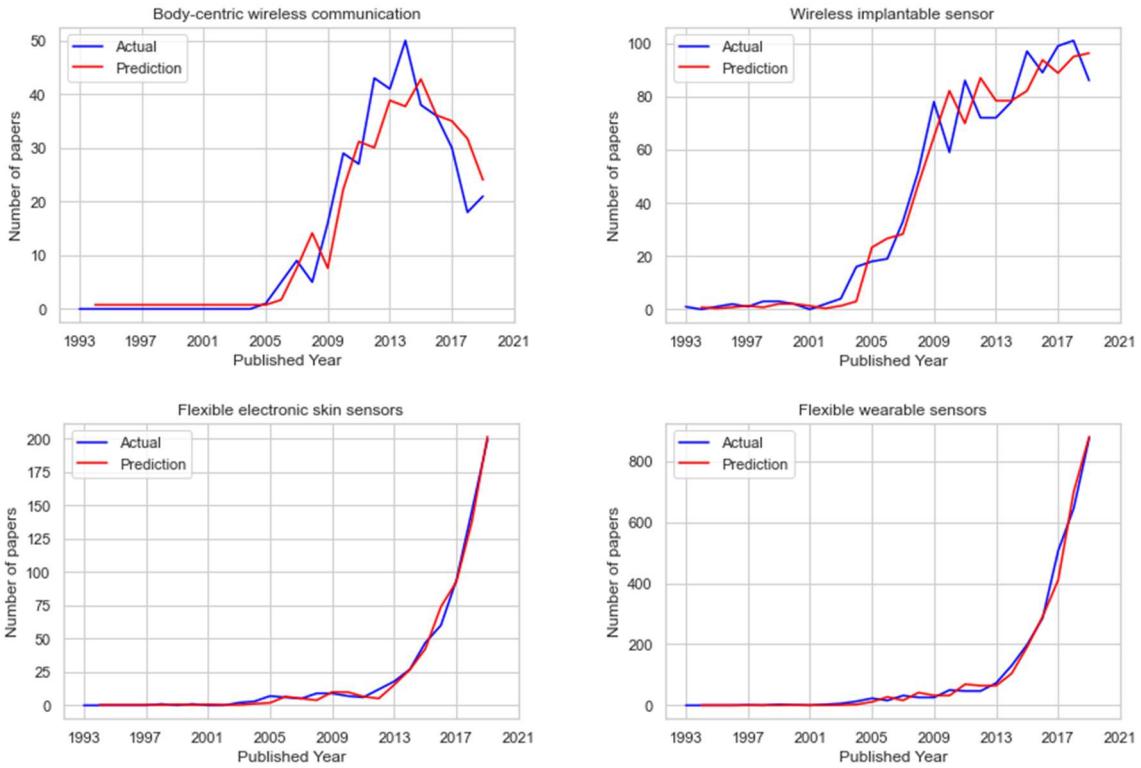


Figure 4.3 Validation results of LSTM model.

Using the validated sequence-to-sequence LSTM model, the number of papers is predicted for each query search keyword in the future period from 2020 to 2024. As seen in Figure 4.4, the number of published papers which have the keywords ‘wireless implant sensor’ and ‘body-centric wireless communication’ observed since 1996 and 2005 respectively. The literatures on ‘body-centric wireless communication’ and ‘wireless implantable sensors’ have declined, and the number of publications looks like saturated in the next 5 years. The number of publications of ‘flexible electronic skin sensors’ and ‘flexible wearable sensors’ are increasing exponentially, and the prediction shows that this will continue to rise until the year of 2024. Therefore, it can be said that the interest in the e-skin sensors or wearable sensors reflects the need of society as well as the researchers.

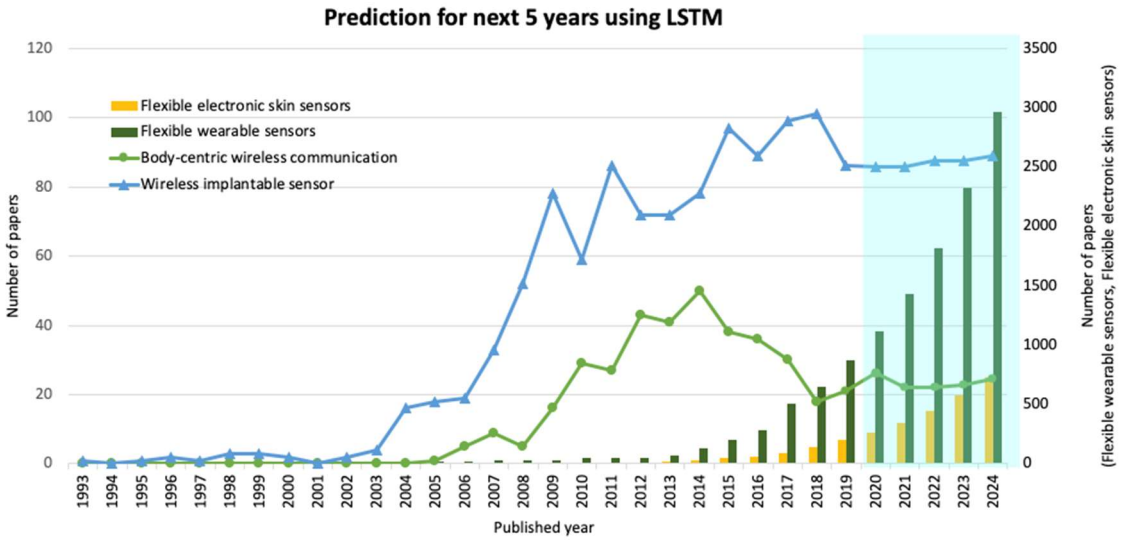
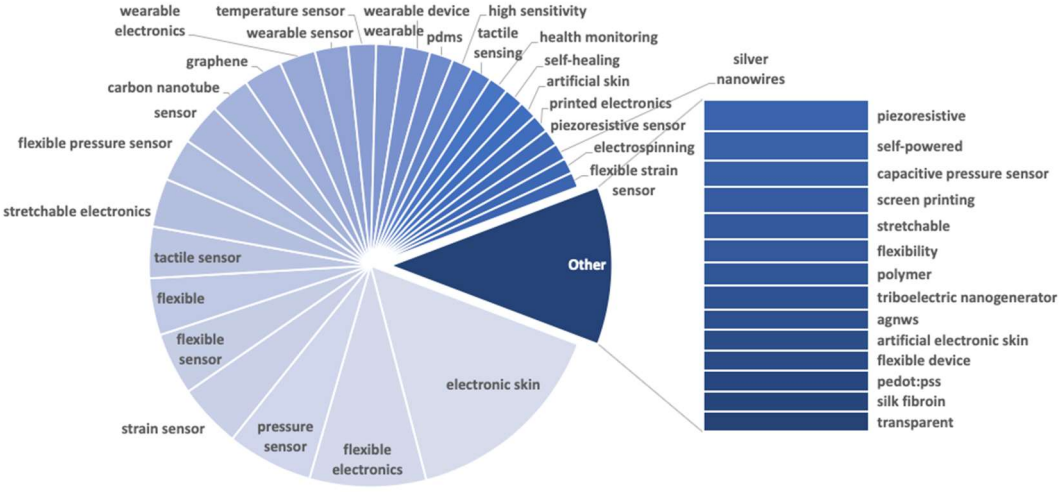


Figure 4.4 Prediction for number of publications in next 5 years using LSTM.
Like Figure 4.3, the y-axis of the line graphs for 'body-centric wireless communication' and 'wireless implantable sensor' is on the left of the figure, while the y-axis of the bar graphs for 'flexible electronic skin sensors' and 'flexible wearable sensors' is on the right of the figure.

Figure 4.5(a) shows the frequent keywords that authors frequently use to describe their works, while Figure 4.5(b) shows the frequent keywords that the Gensim NLP toolkit [113] automatically extracts from the literature on flexible electronic skin sensors. Clearly, author's keywords provide more relevant information because they are 1 or 2 words long. The keywords by NLP tool only contain one word, thus, they only include a general description of the work. RAKE, another automated keyword extraction tool, is used to extract three-word keywords from the literature. These keywords are then provided in Appendix B.1.

(a)



(b)

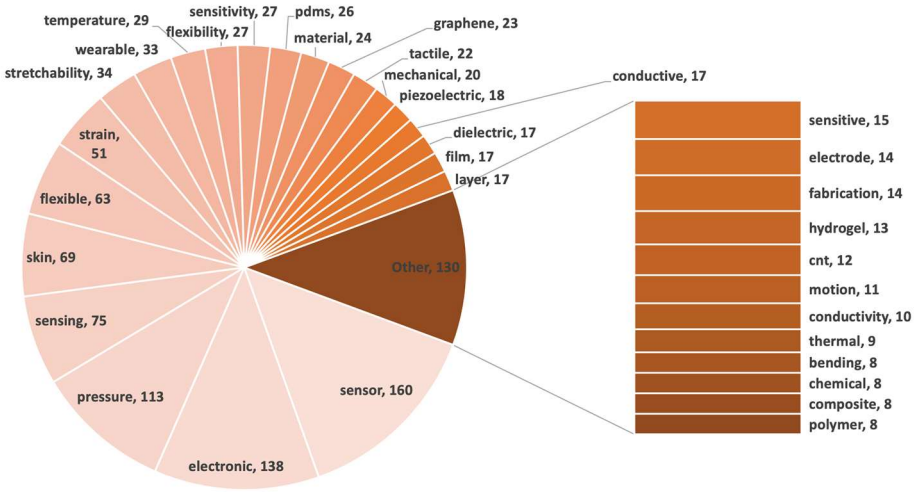


Figure 4.5 The most frequent keywords from 628 publications of flexible electronic sensors. (a) Authors' keywords. (b) Keywords by the Gensim NLP toolkit.

4.2 Strategy

To improve the completeness of review paper by Khan *et. al.* [114], an NLP-assisted survey approach was utilised. The ordinary scientific review depends on the information collected on a database search of relevant literature published in major digital libraries such as Nature, Science, IEEE Xplore, ACM and Google Scholar. The number of recent studies that falls within the scope of a survey are often over a few thousands. Critically analysing such a large

information by human alone could overlook some of the key studies and their findings. To deal with such issue, an NLP-based text summarisation technique was adopted that automatically extracts useful key ideas or most relevant information within the scope of original content. Due to the exponentially increasing availability of documents, the research in NLP-based text summarisation has been demanded and as a result, some of the today's NLP algorithms are capable of producing a concise and fluent summary while preserving key information content and overall meaning [115]. In addition, most publishers today allow the users to access their digital libraries via application programming interface (API).

Zdravevski *et al.* [49] demonstrated the applicability of their NLP toolkit [48] and showed increasing attention from the scientific communities towards enhanced and assisted living environments over the last 10 years and provided new technical trends in the specific research topics. Their NLP toolkit was shown to be useful in expediting the fully automated review process while providing valuable insights from the surveying of relevant articles. The review generated by their NLP toolkit has successfully included informative tables, charts, and graphs. The authors only provided the frequency of keywords from related papers or the yearly changes in keywords, however, this study is enabled to improve an accuracy by compiling a summary of all papers released during the period and using it to assist the writing of review papers.

Recent effective extractive summarisation techniques made use of new machine learning architectures that provided mechanisms through the clustering of deep learning model output embeddings [116]. Large corpora of text data are used to pretrain the model (*e.g.* BERT), which then learns how to represent text units, typically words, in a vector space. These vectorised text representations are capable of accurately capturing a substantial amount of word semantic and syntactic information. The summarisation method based on the BERT

model of Figure 4.6 depicts an example pipeline that includes four steps: pre-processing, text mapping to contextualised embeddings, sentence clustering, and sentence selection.

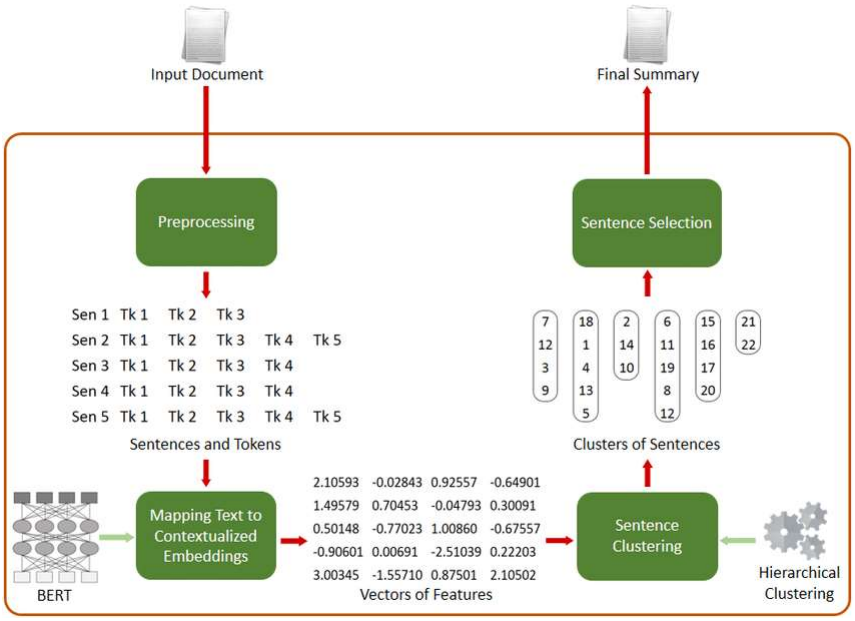


Figure 4.6 The overview of text summarisation process [116].

The concept of shared context is used to describe sentences with close vectors. According to the distance between related sentences' representations in the vector space, the summariser employs a clustering step to group sentences. An agglomerative hierarchical clustering process is used by the summariser to produce groups of phrases while Miller applied the k-means algorithm for summarising lecture transcripts [37]. This study built the framework for our summariser to help with the review of body sensing technologies using the BERT model and k-means algorithm.

As seen in Figure 4.7, the proposed NLP-assisted review strategy has three phases, namely, Keyword Search (KS), Abstract Extraction (AE) and Summary Generation (SG). KS identifies the keywords that correspond to each section of this review. The abstracts which contain any of these keywords are searched against the digital libraries via Elsevier API. At a result of AE, the abstracts that contains each set of keywords are extracted. These keywords

are matched to the sections of the review article that need assistance. The number of extracted abstracts for each set of keywords are shown in the right column of AE. In SG, the extracted abstracts for each set of keywords are fed into the pretrained language model named BERT and k-means clustering for extractive text summarisation. Here, the required number of sentences for each summary could be supplied as a ratio. The supplied ratio parameter for each set of keywords is provided in between the rectangles of AE and SG. BERT built on top of the transformer architecture performs text embedding. Each sentence is vectorised in BERT. k-means clusters similar meaning sentences based on Euclidian distances between these vectorised sentences. Finally, the sentences closest to the centroid of each cluster are identified and a summary for each set of keywords is generated. Appendix 4.2 contains the summary results from all compiled literatures in accordance with the seven subsections of Ahsan and researchers' review article [114].

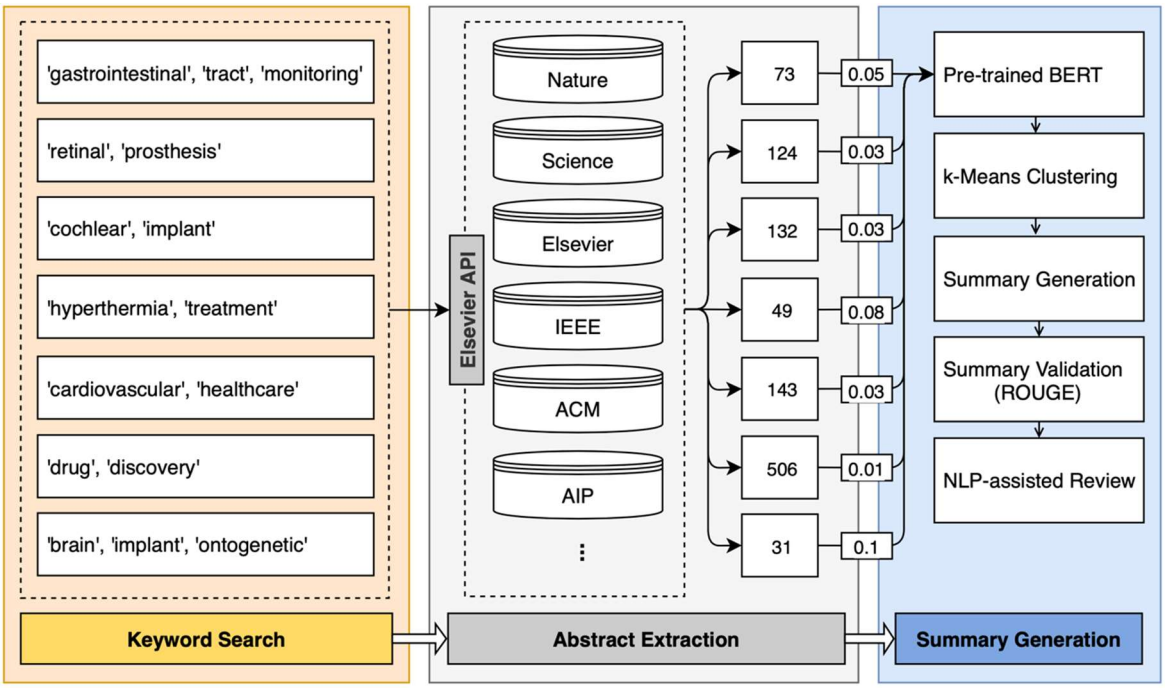


Figure 4.7 The NLP-based summary generation.

In order to evaluate the NLP algorithm-derived summary quantitatively, firstly, this study compared the most frequent keywords from the generated summary and from the subsection of the paper. The most frequently used keywords from the subsection, 'gastrointestinal tract monitoring' of review paper are displayed in Figure 4.8 along with the NLP-generated summary of literatures related to subsection title. As seen in Figure 4.8, although 'capsule' and 'endoscopy' in red boxes are not among the top 20 keywords in the review paper, they are among the top 20 in the NLP-generated summary. These gaps allow human experts to enhance the review article or consider missing subjects since experts are limited in their ability to go through vast amounts of literature and are biased by their own specialisations.

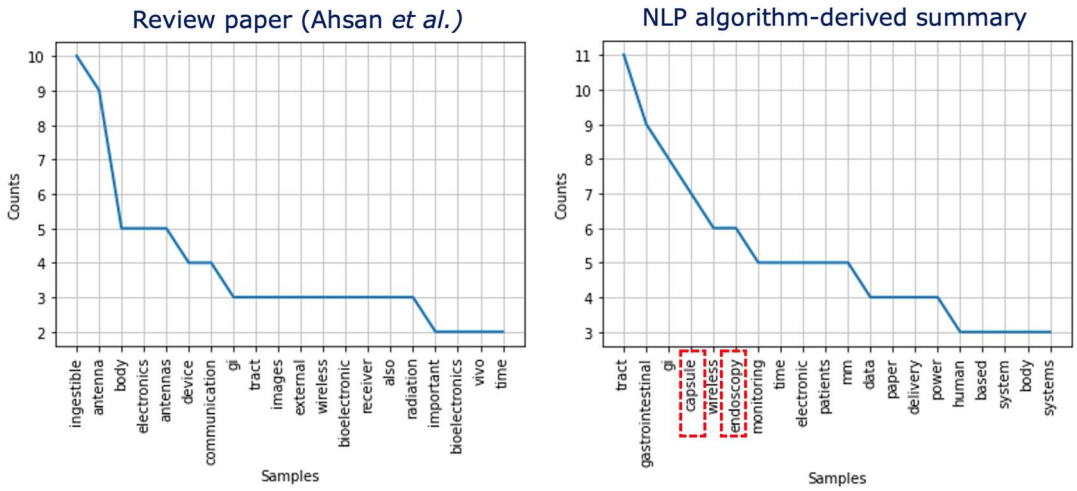


Figure 4.8 A comparison of the most frequent keyword from summaries.

There are two different summaries in Table 4.1, the right one produced by an NLP system and the left one from a review article written by human experts. Although it would be challenging to immediately apply the NLP-produced summary, it might contain some important details that the authors may have missed. These details might then be thoroughly inspected and incorporated to the paper's original section.

Table 4.1 A comparison of the summaries.
The right one shows the NLP-generated summary for the corresponding paragraph, while the left one shows the summary of the original paper.

Review paper (Ahsan *et al.)*	**NLP algorithm-derived summary**
One of the major challenges in ingestible bioelectroincs is the requirement of efficient antenna design for in-vivo wireless communication. Hence, a compromise is required between the antenna dimensions, operating frequency and electromagnetic losses in the body tissues for invivo communication.The impedance matching of implantable antennas is also critically important for their continuous operation as the ingestible electronics traverse through the GI tract. Despite considerable progress in transmitting high resolution images wirelessly from an ingestible bioelectronic, the challenges associated with limited battery capacity are stymieing its operation for longer periods of time in the body.	*The navigation techniques suggested for wireless capsule endoscopy are image-based that are required to transfer and process a significant amount of data in real-time operation. Electronic drug delivery systems such as capsules, on the other side, can be used not only to deliver drugs to a specific site in the gastrointestinal tract but can also record data and report the state of patients' gastrointestinal tract, and after excretion, this information can be studied and used to present them graphically. The wireless capsule endoscopy (WCE) is one of the promising body area networks (BANs) applications that provides a non-invasive way to inspect the entire gastrointestinal (GI) tract.*

4.3 Validation

The Recall-Oriented Understudy for Gisting Evaluation, also known as ROUGE [117], is a software package and a set of metrics for assessing automatic summarisation and machine translation in NLP. The metrics contrast an automatically generated summary or translation with a reference summary or set of references that were created by humans. ROUGE F score gives the harmonic mean of the precision and recall. ROUGE 1 refers to the overlap of unigram (each word) between the generated summaries and the referenced summaries while ROUGE L refers to the overlap of longest common subsequence. The histogram in red box shows the ROUGE scores on public summarisation datasets. To prove the usefulness of the generated summaries, ROUGE matrix is calculated comparing the generated summaries and human-written summaries. As seen in Figure 4.9, the ROUGE scores of our generated summaries are comparable to the state-of-the-art scores on well-reference summarisation datasets. Using the validated summaries, the authors improved the completeness of each section of the review.

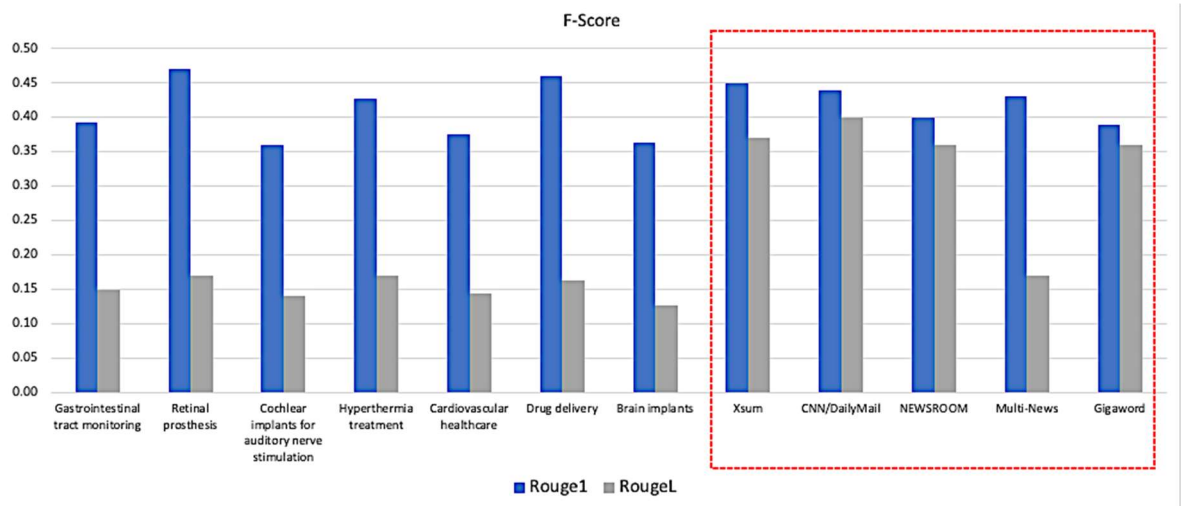


Figure 4.9 The ROUGE validation results.

4.4 Summary

Text summarisation can assist researchers and save time by automatically identifying and presenting the most important ideas within long documents without the need to read the entire text. Early text summarisation research relied on straightforward term frequency characteristics to narrow down a text document's most crucial information. Since then, a variety of features and methodologies have been integrated by numerous summarisation techniques into the process of choosing content. The advancement of powerful language models in NLP has made it possible to boost text summarisation accuracy. The main purpose of this chapter is to provide researchers with writing assistance when using the recently developed text summarisation technique to write review papers. This is as a result that machine-written summaries can have a wider research scope than human summaries, and researchers can take them into consideration because the NLP algorithm used for summarisation can produce fair summaries despite biased thinking. Applying this proposed assistant framework to the review paper which is actually published allows for the writing of a more comprehensive paper, which is a significant accomplishment.

Chapter 5

Review and Prediction of Antenna and Propagation Research from Large-scale Unstructured Data with Machine Learning

The past century has witnessed remarkable progress in antennas and propagation (AP) research, which has made dramatic changes to our society and life and has led to paradigm shifts in engineering and technology. Although the underlying theory of electromagnetics is well established and mature, research on antennas and propagation will continue to play a paramount role in the 4th industrial revolution. This chapter presents an approach based on NLP and ML techniques, to review AP research based on large-scale unstructured data from openly published scientific papers and patents, and, in turn, provides meaningful summative and predictive information. Similar to Chapter 3, 159k research papers published between

1981 and 2021 were carefully screened, and 2,415 significant keywords reflecting important research topics from the past and present in AP were extracted. However, in this chapter, as compared to Chapter 3, a weight for each literature was generated by analysing article attributes in order to produce future predictions that reflect more reality. The weighted time-series data of each keyword then was applied to an encoder-decoder Long Short-Term Memory (LSTM) network with integrated attention mechanism to predict the future trend of AP research in the form of a Gartner's hype cycle.

5.1 Introduction

In a recently published book by URSI on its Centennial anniversary celebration [118], a comprehensive review of electromagnetics theory, antennas and their applications to wireless communication, radar and other modern technologies such as satellites, space and mobile health have been conducted by a group of prominent engineers. Similar reviews can be found and performed regularly by leading experts in their subject areas. Notably, Balanis reviewed the Antenna Theory in [119]; Jensen and Wallace summarised research challenges and opportunities in antennas and propagation for MIMO (multiple-input and multiple-output) wireless communications [120]. In recent years, there have been an increasing number of publications on topical reviews in subject areas such as antennas for energy harvesting and wireless power transfer [121], machine learning for antenna design and optimisation [122], [123], antennas for mobile handsets [124] and flexible wireless sensors [125] and wearable technologies [126], [127].

With the advent of AI and ML technologies, it is widely accepted that mining and extracting useful information from big data of scientific publications will help us analyse the past and current trend of research, and predict future directions assisted by ML algorithms.

Among all approaches developed in ML, NLP has been successfully applied for voice recognition and semantic information extraction, both of which have played an important role in conducting new scientific research [128], [129], [130], [131]. For example, Kuniyoshi *et al.* [53] proposed the label definitions for material names and properties and built a corpus containing 836 annotated paragraphs for training a named entity recognition (NER) model [55]. They achieved a micro-F1 score of 78.1% for NER model. This model was then applied to analyse 12,895 material research papers, and the trend of inorganic material research was captured by investigating the change of keyword frequencies by year and country. Elton *et al.* [66] collected textual data from various sources (BERT journal articles, conference proceedings, the US Patent & Trademark Office, and the Defense Technical Information Centre archive on archive.org) and successfully extracted meaningful chemical-chemical and application-chemical relationships by doing computation with word vectors and without hand-labelling.

In this chapter, a fully automate framework that extracts meaningful keywords from a large number of publications related to antennas and propagation research was proposed with an architecture as shown in Figure 5.1.

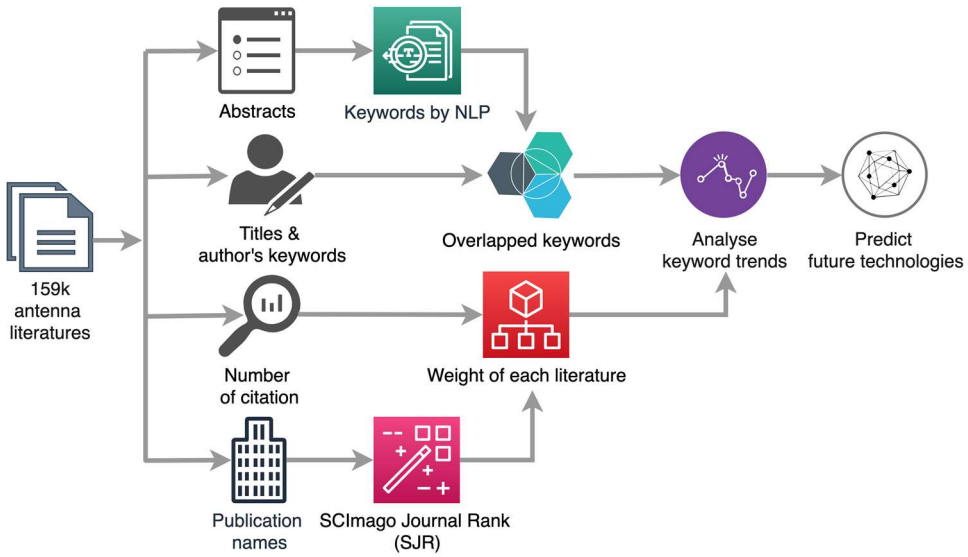


Figure 5.1 An overview of keyword prediction framework.

From the abstracts of 159k antenna literatures, RAKE (Rapid Automatic Keyword Extraction) algorithm extracts meaning keywords. The weight of each abstract is calculated using number of citation and SJR (SCImago Journal Rank) index which is applied to the extracted keywords of the abstract. The weighted keywords are analysed by number of word occurrences (*i.e.* frequency) and finally, the future trends of antenna research using encoder-decoder LSTM with attention layer are predicted.

In particular, a structured database which has the details of scholarly article attributes (*e.g.* title, authors, affiliation, author's keywords, journal name, number of citation and abstracts) was built from the literatures using Scopus Application Programming Interface (API). 167k papers between 1906 and 2021 were retrieved, however, for some of early papers, only limited information (*e.g.* title) was searchable due to lack of digitisation. Therefore, scholarly article attribute data from 159k papers published since the year of 1981 was utilised. This large-scale literature data was then used as an input for extracting keywords and weighting each literature. Here, this chapter employ one of the NLP techniques, a domain-independent keyword extraction algorithm which determines key phrases in a body of text by analysing the frequency of word appearance and its co-occurrence with other words in the text [132]. By verifying the obtained keywords of the automatic keyword extractor (*i.e.* RAKE) using titles and author's keywords, it significantly reduces the data complexity and enhances the accuracy of selecting meaningful keywords. Figure 5.2 displays the collection of 159k papers on metamaterial from 2001 to 2020 along with a record of all article attributes, including titles, author keywords, and abstracts. From each abstract, the RAKE algorithm generated the keywords. In order to select the most appropriate keywords, RAKE first generates keyword pairs, and if these RAKE-generated keywords appear in the paper's title or author keywords, they are chosen as keywords. This method, which was implemented using Python code, created a total of 2,415 meaningful keywords. These keywords are taken into account throughout all years and papers. This is a fully automated process using NLP; no assistance from humans is required.

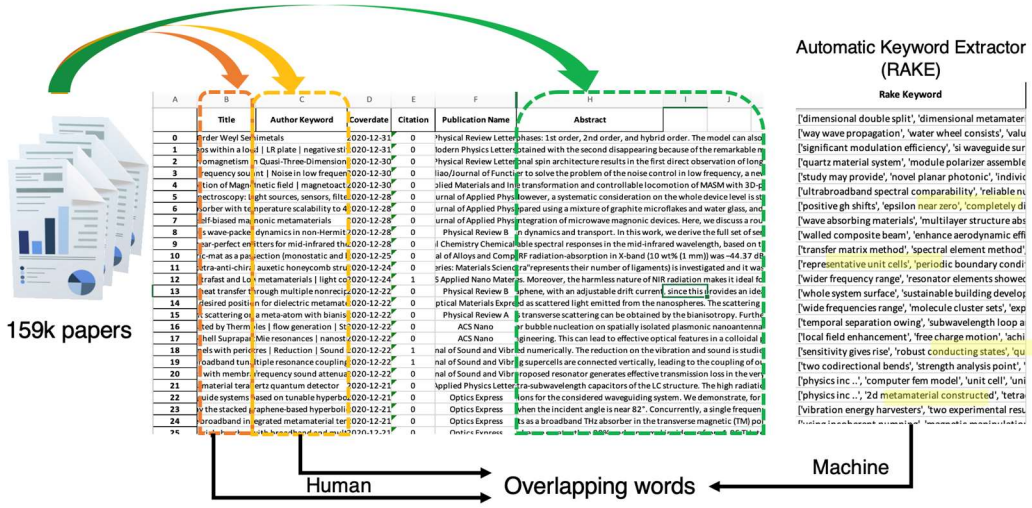


Figure 5.2 A methodology for validating keywords.
Titles and authors' keywords have been used to validate keywords from an NLP algorithm.

Besides, a weight metric was assigned to each literature based on some scholarly attributes (*i.e.* number of citations and publication name). This new weighting scheme allows going beyond simple numerical analysis to assessing the actual influence of each paper to its research field. The number of occurrences of each keyword per year over the period of 1981–2021 is counted reflecting the weight of each paper. Analysing the change in frequency of each keyword in the form of time-series helps us to comprehend the past antenna research trend. Then, an encoder-decoder LSTM with attention layers was utilised to predict the future trend of antennas and propagation research in the form of Gartner's Hype Cycle.

Within this framework, the methods for indicating article performance are as follows. 1) an average citation for each paper as a measure of the usefulness, impact, or influence of a publication; and 2) SCImago Journal Rank (SJR) indicator [133], developed by SCImago from the widely known algorithm Google PageRank [134], [135]. The former is an article level metric calculated by dividing the total number of citations by the number of years since the article has been published. Average citation is a simple metric that can compensate for the time an academic/journal has been active to some extent, which provides a fair comparison for both junior and senior researchers. The latter is a journal level metric obtained as the

average number of weighted citations received per document published in that journal over the previous three-year time window, as indexed by Scopus. This SJR indicator could represent the scientific influence of scholarly journals that accounts for both the number of citations received by a journal as well as the importance or prestige where the citations come from. Higher SJR indicator values are meant to indicate greater journal prestige. It should be noted that above indicators have been designed to address the limitations of the well-known Journal Impact Factor (JIF) [136], [137]. Supporting the recent statement by San Francisco Declaration on Research Assessment (DORA) [138] that the research impact should be shown with other journal-based indicators (Eigen index, SCImago, h-index, publication period, etc.), this study adopted both article-level and journal-level metrics, which provide a multi-faceted view of research impact without simply relying on JIF to reflect the influence of individual articles or scientists. In this way, every paper in the database is given a weight to indicate its overall influence other than the only number of citations.

5.2 Information Collection and Retrieving

In the post-industrial society (*a.k.a.* the third wave), large scientific databases have been built on the World Wide Web (WWW) offering search facilities on a particular subject and access to the massive amount of data simultaneously. Elsevier's Scopus is one of the abstracts and citation databases. Scopus allows access to over 36,377 titles from approximately 11,678 publishers with over 20% more coverage than Web of Science [139].

The proposed framework utilises a Python package, pybliometrics [140], to access published papers related to antennas and propagation via the Scopus' RESTful API using Hyper-Text Transfer Protocol (HTTP) requests. Using this package, all the scholarly article attributes (*e.g.* title, abstracts, publication name, author's keywords, number of citations and

affiliation name) could be collected. Among different literature types, namely, article, review, conference, and book, only the data from the article type of literature, which returned 167k antenna-related articles published since 1906, was collected. Figure 5.3 gives a statistical overview on each of the attributes. Figure 5.3(a) shows that majority of antennas research has been published in two flagship journals in the IEEE Antennas and Propagation (A&P) Society and Microwave and Optics Letters. Antennas and propagation research has longstanding associations with applications such as remote sensing, wireless communications, and electromagnetics, reflected by the number of papers published in journals from sister societies.

Over the years, A&P research has attracted a significant amount of financial support from respective national funding bodies, with China and USA leading the way, followed by Japan, EU and UK (Figure 5.3(b)). Figure 5.3(c) presents a network graph visualising key research topics in A&P and their inter-relations. Clearly, despite recent boom of 5G/6G development, millimetre wave and THz antennas have been a topic of interest for some time. Microstrip patch antennas remain to be a strong research hotspot, due to their ease of fabrication, cost-effectiveness and being low profile. Some new technologies have spun out and led to development of Frequency Selective Surfaces (FSS), Defective Ground Planes (DGS) and Surface Integrated Waveguides (SIW) etc. The latter has strong links with the development of millimetre wave technologies, equally, studies on lens antennas and phase shifters have made their mark and attracted renewed interests due to the need of hybrid analogue and digital beamforming network for 5G/6G wireless communications.

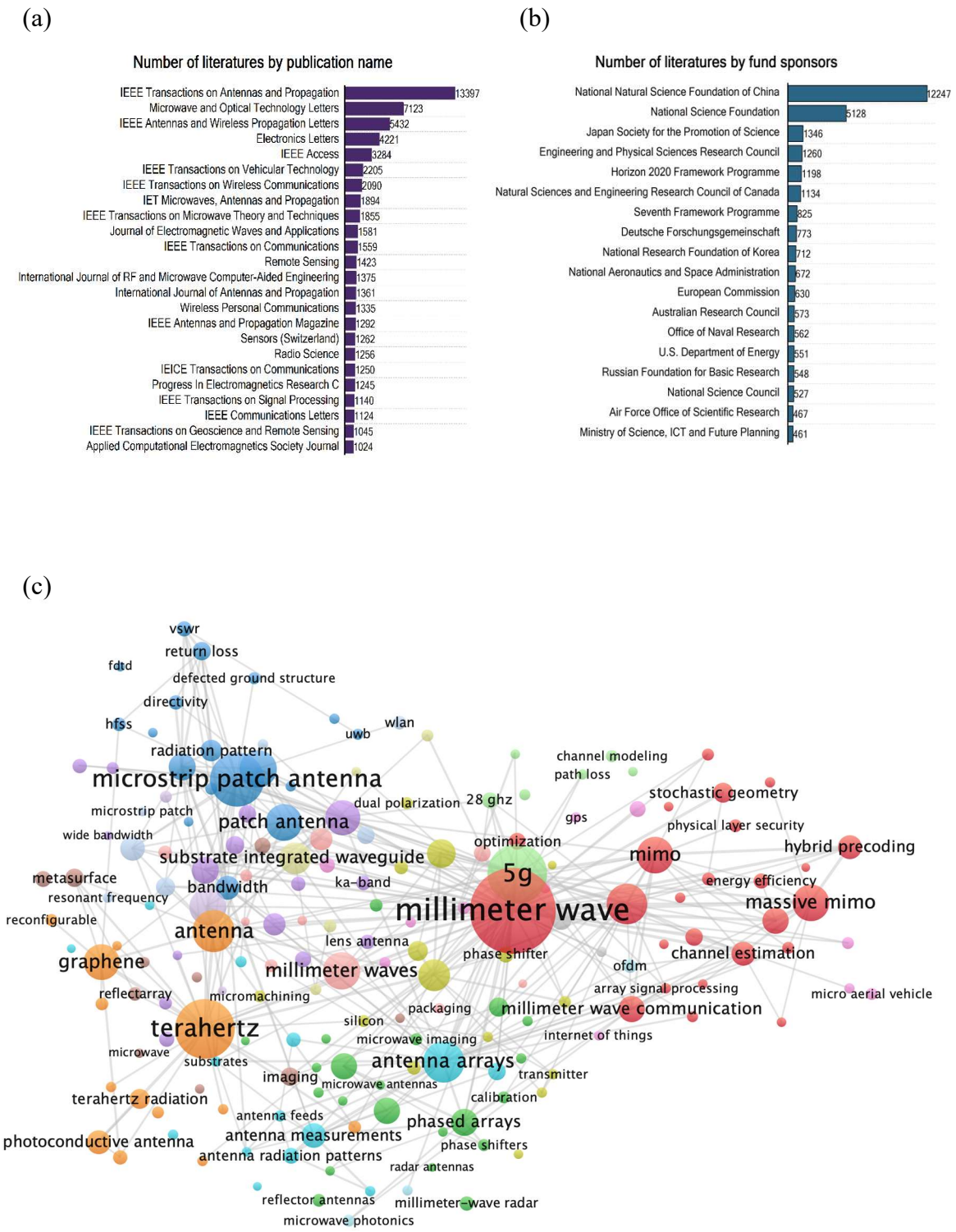


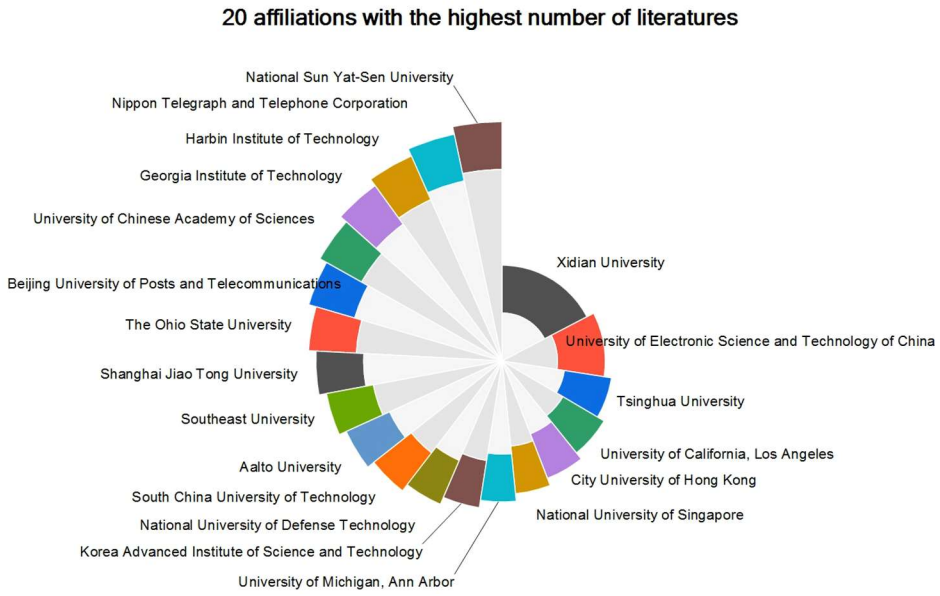
Figure 5.3 A statistical overview on A&P research.
(a, b) These figures show number of publications by publication names, and fund sponsors. (c) This figure visualises co-occurrences of author's keywords focused on millimetre, terahertz, and micro antenna. Here, the size of each circle represents a number of co-occurrences.

Antenna arrays including phased arrays have direct applications to radars, imaging and wireless communications. Significant research has been conducted in designing feed antennas, reflector antennas and arrays, and microwave photonics. Notably, antenna measurement techniques have been studied in conjunction with antenna arrays. THz development is supported by new emerging materials such as graphene for photoconductive antennas, silicon technologies with micromachining, and reconfigurable reflectarray and metasufaces. Technologies such as massive multiple-input and multiple-output (MIMO), channel modelling, array signal processing, channel estimation and new coding/modulation schemes have been highlighted for wireless propagation research with applications related to micro-aerial vehicles, internet of things and Global Positioning System (GPS) etc.

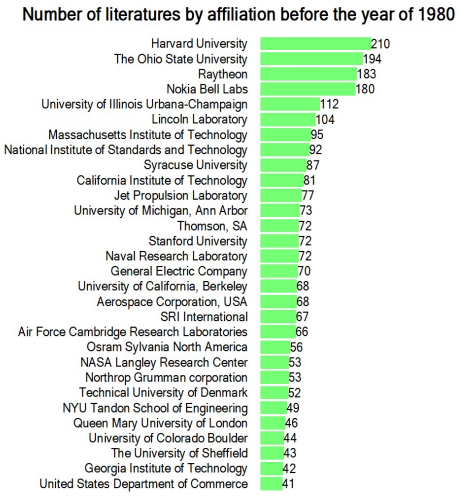
It should be noted that in general the articles which were published before 1980s are available as a PDF or image file while the articles published after this time are mostly available in Hyper-Text Markup Language (HTML) format. This means that the articles in HTML format can be recognised, and each scholarly article attribute data can be downloaded in an automated manner. Limited attribute data in HTML format, including all the title, author, affiliation, and journal information, was extracted for literature that was published before 1980. As expected, since around 1980, the number of articles published increased significantly. In the period 1981–2021, remarkable changes are observed in ranking in particular, number of literatures by affiliation and number of literatures by country. Figure 5.4 compares how each country/institution's research activity changes over time based on number of published papers before and after the year of 1980. Top 20 affiliations have produced the most publications throughout the whole period in A&P research are shown in Figure 5.4(a). Figure 5.4(b) presents top contributors to the research area, dominated by US and UK universities and companies, with Harvard University and the Ohio State University leading. Post-1980s have witnessed the surge of research from other countries and areas

including France, Sweden, Finland, Russia, Japan and Hong Kong with China dominating the chart in terms of research publications in A&P (Figure 5.4(c)). Figure 5.4(d) presents a geographic distribution of research papers around the world before and after 1980. It is evident that A&P is now researched globally with increasing papers coming from Latin America, Africa and the Mideast. A&P research is continuously growing in countries such as China, India, Iran, Spain, and South Korea, and remains strong in Australia, Canada, France, Germany, Italy, Netherlands, and Sweden (Figure 5.4(e, f)).

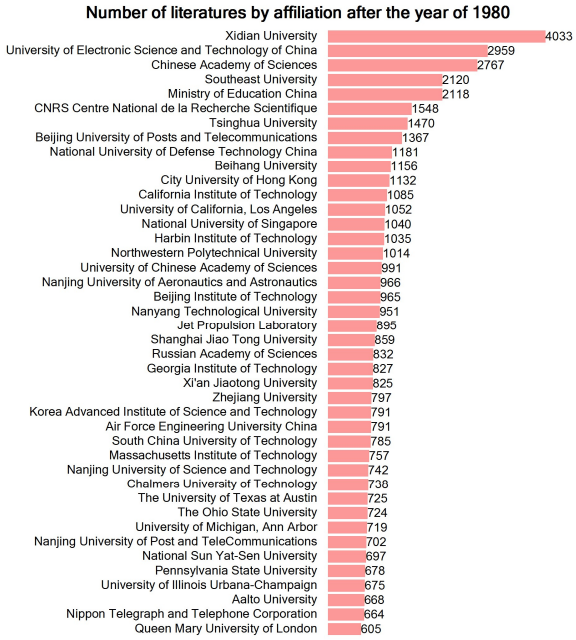
(a)



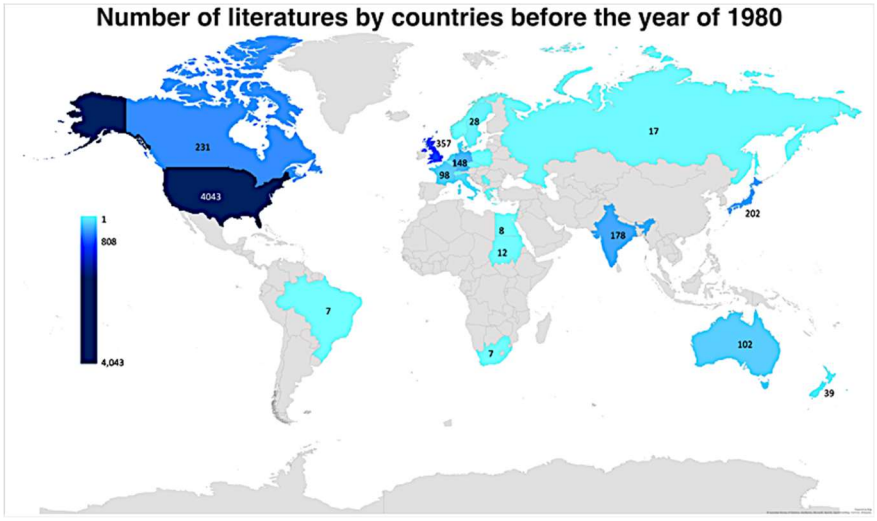
(b)

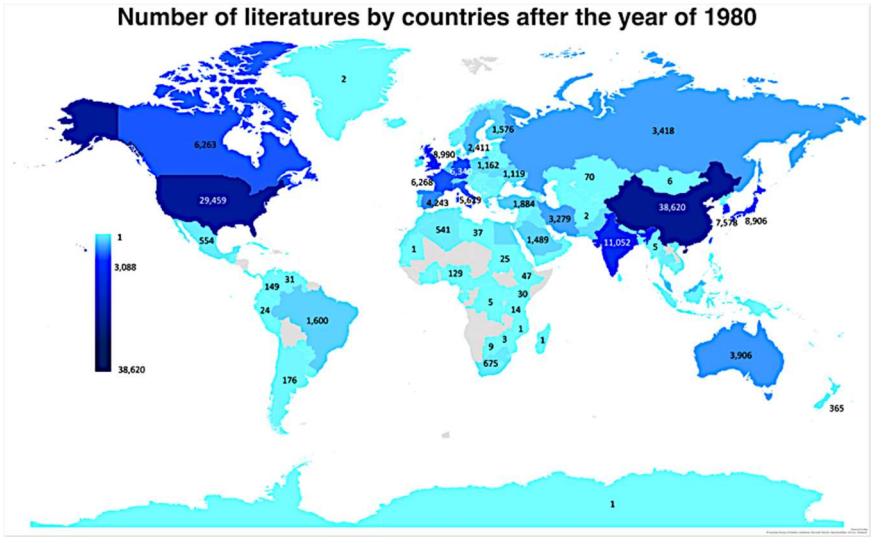

Number of literatures by affiliation before the year of 1980
Harvard University 210
The Ohio State University 194
Raytheon 183
Nokia Bell Labs 180
University of Illinois Urbana-Champaign 112
Lincoln Laboratory 104
Massachusetts Institute of Technology 95
National Institute of Standards and Technology 92
Syracuse University 87
California Institute of Technology 81
Jet Propulsion Laboratory 77
University of Michigan, Ann Arbor 73
Thomson, SA 72
Stanford University 72
Naval Research Laboratory 72
General Electric Company 70
University of California, Berkeley 68
Aerospace Corporation, USA 68
SRI International 67
Air Force Cambridge Research Laboratories 66
Osram Sylvania North America 56
NASA Langley Research Center 53
Northrop Grumman corporation 53
Technical University of Denmark 52
NYU Tandon School of Engineering 49
Queen Mary University of London 46
University of Colorado Boulder 44
The University of Sheffield 43
Georgia Institute of Technology 42
United States Department of Commerce 41

(c)

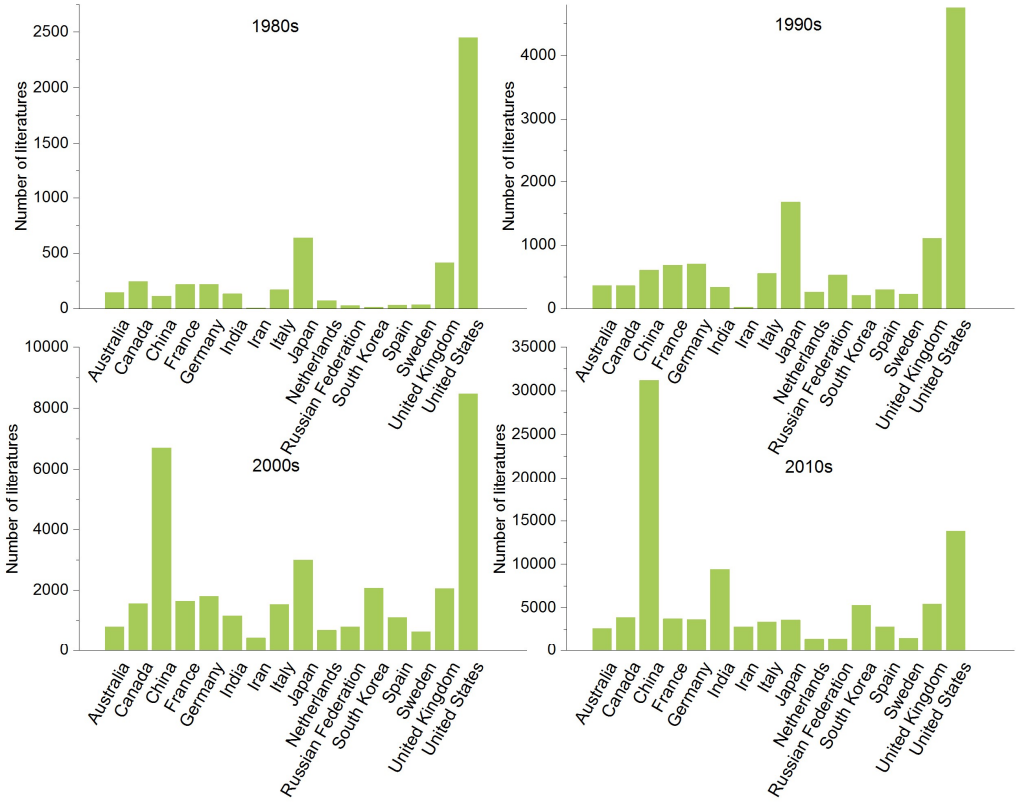

Number of literatures by affiliation after the year of 1980
Xidian University 4033
University of Electronic Science and Technology of China 2959
Chinese Academy of Sciences 2767
Southeast University 2120
Ministry of Education China 2118
CNRS Centre National de la Recherche Scientifique 1548
Tsinghua University 1470
Beijing University of Posts and Telecommunications 1367
National University of Defense Technology China 1181
Beihang University 1156
City University of Hong Kong 1132
California Institute of Technology 1085
University of California, Los Angeles 1052
National University of Singapore 1040
Harbin Institute of Technology 1035
Northwestern Polytechnical University 1014
University of Chinese Academy of Sciences 991
Nanjing University of Aeronautics and Astronautics 966
Beijing Institute of Technology 965
Nanyang Technological University 951
Jet Propulsion Laboratory 895
Shanghai Jiao Tong University 859
Russian Academy of Sciences 832
Georgia Institute of Technology 827
Xi'an Jiaotong University 825
Zhejiang University 797
Korea Advanced Institute of Science and Technology 791
Air Force Engineering University China 791
South China University of Technology 785
Massachusetts Institute of Technology 757
Nanjing University of Science and Technology 742
Chalmers University of Technology 738
The University of Texas at Austin 725
The Ohio State University 724
University of Michigan, Ann Arbor 719
Nanjing University of Post and TeleCommunications 702
National Sun Yat-Sen University 697
Pennsylvania State University 678
University of Illinois Urbana-Champaign 675
Aalto University 668
Nippon Telegraph and Telephone Corporation 664
Queen Mary University of London 605

(d)


Number of literatures by countries before the year of 1980
1
808
4,043
231
4043
28
357
148
98
17
202
178
8
12
7
7
102
39


Number of literatures by countries after the year of 1980
1
3,088
38,620
6,263
29,459
554
149
31
24
1,600
176
2
1,576
3,418
2,411
1,162
1,119
70
6
6,268
38,620
1,884
3,279
2
7,578
8,906
541
37
1,489
1
11,052
5
129
25
47
30
5
14
1
3
9
675
1
3,906
365
1

(e)



(f)

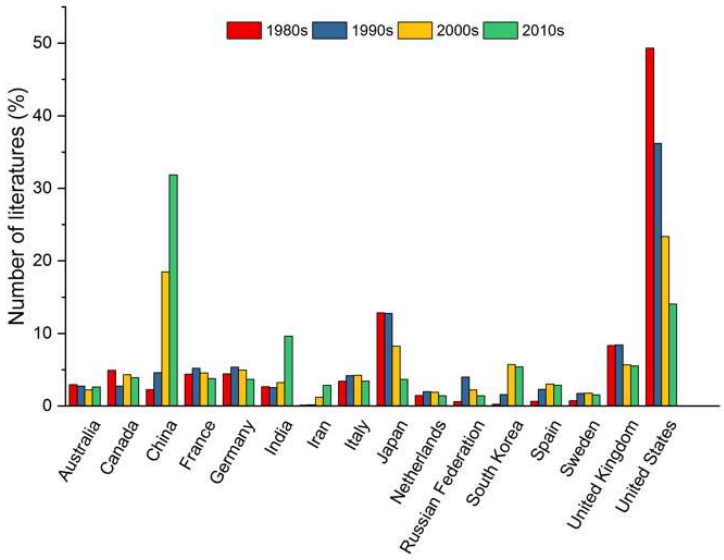


Figure 5.4 Each country/institution's research activity changes over time based on number of published papers.

(a) 20 affiliations with the most publications over the entire time period. (b, c) Ranked affiliations by number of literatures before and after the year of 1980. The total number of literatures is 167k and the period of 1981–2021 gives 159k literatures only. (d) These map graphs show the changes of number of literatures by countries before and after the year of 1980. Dark colour indicates a greater number of papers. (e) number of literatures per countries is compared every decade during the period. (f)The figure represents that number of literatures is converted into percentage.

5.3 Abstract Weighting and Keyword Extraction

To build a comparable analysis and prediction, each abstract is given a different weight. Each paper's normalised citation is used to reduce its time bias, and the SJR index is used to measure its overall scientific influence in a broader level. In consequence, each abstract's publication name is replaced with a corresponding SJR index. Weight of each abstract (or paper) is calculated using equation (5.3.1), with two hyper-parameters, α and β.

$$weight = \alpha \cdot normalised\ citation + \beta \cdot normalised\ SJR\ Index$$
$$where,\ \alpha + \beta = 1 \qquad (5.3.1)$$

The example using cut-off is Figure 5.5. While Figure 5.5(a) shows top 20 affiliations based on the weighted abstracts, Figure 5.5(b) re-orders the ranking of the affiliations using the weighted abstracts with a cut-off (*i.e.* 0.02) of the weight. Since the number of papers with weights below 0.02 is very large, we removed the papers with weights below 0.02, and the results are shown in Figure 5.5(b). Cut-off was applied, and the order significantly changed. The top institution on Figure 5.5(a) is now listed as being in position 14. That is as a result of the specific institutions publishing a large number of papers with little impact. Additionally, Table 1 provides the details of seven papers with the highest weight values.

(a) (b)

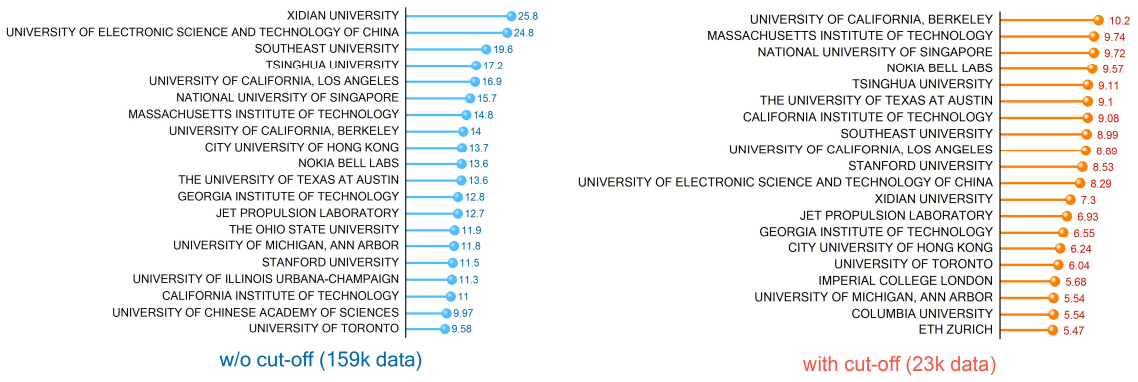


Figure 5.5 Top 20 affiliations based on the weighted abstracts.

These are generated by considering up to third affiliations on each paper with each paper weighted using equation (5.3.1) which has two parameters whose values need to be set (alpha=0.3 and beta=0.7). In this experiment, a higher weight is given to the prestige of the journal in which the paper is published than the average number of citations received by the paper as suggested by DORA [138]. The numbers shown in figures (a) and (b) indicate the mean of the calculated weight of individual papers published by each institution.

Table 5.1 Top 7 highly weighted papers

Title	Journal	Year	Ref.	Weight
What will 5G be?	IEEE Journal on Selected Areas in Communications	2014	[141]	53.98
Light propagation with phase discontinuities: Generalized laws of reflection and refraction	Science	2011	[142]	50.44
Cooperative diversity in wireless networks: Efficient protocols and outage behavior	IEEE Transactions on Information Theory	2004	[143]	46.31
Massive MIMO for next generation wireless systems	IEEE Communications Magazine	2014	[144]	41.56
A simple transmit diversity technique for wireless communications	IEEE Journal on Selected Areas in Communications	1998	[145]	38.10
Millimeter wave mobile communications for 5G cellular: It will work!	IEEE Access	2013	[146]	37.78
Scaling up MIMO: Opportunities and challenges with very large arrays	IEEE Signal Processing Magazine	2013	[147]	34.70

To predict the future trend of past and current antenna technologies, extracting relevant features (*e.g.* keywords) from all available abstracts is the first and foremost step. Extracting right keywords with significant information not only leads to achieving high accuracy in prediction but also reduces the time required to search over the large unstructured abstract data. Thus, the well-referenced Rapid Automatic Keyword Extraction (RAKE) algorithm which automatically calculates a sum of the number of co-occurrences and then divided them by their occurrence frequency [148] was applied. Although meaningful keyword phrases of each paper were extracted confidently based on our previous research using RAKE [149], due to its volume and complexity, it is not possible to use all of these phrases for a prediction task. To address this problem, only keywords of each abstract which are overlapped with the RAKE's keyword phrases and author's keywords or the words in the title of the paper are selected. From a total of 159k abstracts, 2,415 keywords are extracted, and these are now embedded into a 200-dimensional vector space via word embedding to analyse their relationship. For vectorisation of words, a pre-trained word embedding model, Mat2Vec, which is trained on 3.3 million scientific abstracts [40] was utilised. By combining

words with similar meanings, keyword clusters were formed by detecting inherent similarity relationships, which aimed to simplify perception and cognition related to complex keywords. Clustering analysis performs an unsupervised mapping and labelling for each keyword based on pre-identified clusters in the vector space. To facilitate this process, all 200-dimensional 2,415 keywords are compressed into 3-dimensional spaces using Principal Component Analysis (PCA).

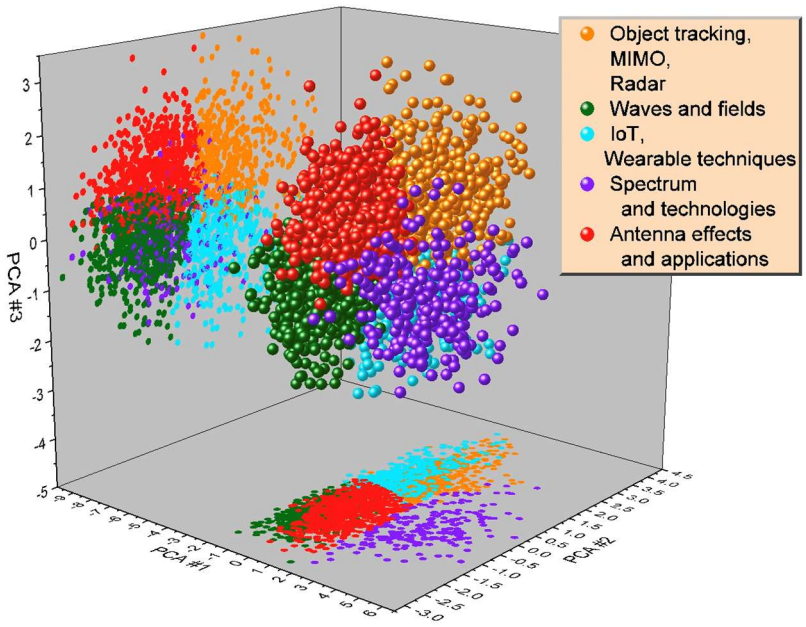


Figure 5.6 3D plot of *k*-means clustering results.

All 2,415 keywords in 200-dimensional spaces are plotting into 3-dimensional space using PCA. *k*-means algorithm makes all 3D keywords categorise into 5 groups.

Figure 5.6 visualises the embedding keywords using PCA and shows that all the keywords are clustered into 5 groups (or categories) using the *k*-means algorithm [37], [38]. These clusters are the collections of keywords with relatively similar meanings or high number of co-occurrences. Here, the representative attributes of each cluster were listed. Table 5.2 shows 5 clusters and their attributes with 10 frequent keywords in each category. Based on these keywords, human experts have attempted to identify a broad topic for each cluster as shown in Figure 5.6 and Table 5.2.

Table 5.2 The 10 most frequent keywords from 5 clusters

Cluster 1	*Cluster 2*	*Cluster 3*	*Cluster 4*	*Cluster 5*
massive MIMO	energy efficiency	3D printing	antenna array	phase shifter
hybrid beamforming	dielectric constant	smart city	microstrip antenna	beam steering
5G communication	quality factor	cellular communication	patch antenna	impedance matching
satellite image	compressive sensing	mmWave communication	dipole antenna	mutual coupling
software radio	thermal emission	wearable application	radio frequency	electromagnetic scattering
small UAV	optical fibre	invasive specie	ultra-wideband	doppler effect
passive radar	microwave discharge	Shannon capacity	rectangular waveguide	magnetic resonance
machine learning	RCS reduction	blood pressure	RFID technology	spin wave
artificial intelligence	Wishart matrix	sensor network	THz communication	time switching
object tracking	gold nanorods	smart agriculture	wireless charging	peak gain
Object tracking, MIMO, radar	*Waves and fields*	*IoT, wearable techniques*	*Spectrum and technologies*	*Antenna effects & applications*

5.4 Analysing Past Trends

Quantifying how the frequency of each keyword changes over time helps us to understand its trend intuitively. To do this, we build a dataset that tracks the occurrences of each keyword from 1981 to 2021 in the form of time series. These are occurrences based on the weight of the published paper where a particular keyword belongs to. The frequency dataset is then normalised using equation (5.4.1) which transforms the values of each point in the time series to a common scale. As in equation (5.4.1), NF is a normalised value that is computed as the frequency of the given word frequency (WF) divided by the total number of literatures from a certain time period, k.

$$NF_k = \frac{WF_k}{total\ number\ of\ literatures} \times 100 \tag{5.4.1}$$

Figure 5.7 shows how the frequencies change with respect to the number of literatures per year quantifying the rate of change for each keyword over time, for example, ‘massive MIMO’ and ‘transmit power’. As seen this figure, at some point after applying the weight,

frequency data are amplified. All keywords' weighted frequency data were calculated by the published years' sequence, and from the sequence data, each keyword was then divided into three groups according to the trend, which is, respectively, increasing, decreasing, and emerging. The weight applying results of some keywords from each trend are provided in Appendix C.1. 'transmit power' as in Figure 5.8(a) shows a steady increasing trend since its first occurrence while 'massive MIMO' shows a sharp increase around the year of 2010 which makes it an overall increasing trend as in Figure 5.8(b). Notably, 'transmit power' is an increasing keyword while 'massive MIMO' is an emerging keyword (Figure 5.8(c)). Figure 5.8 illustrates top 30 increasing, decreasing and emerging keywords.

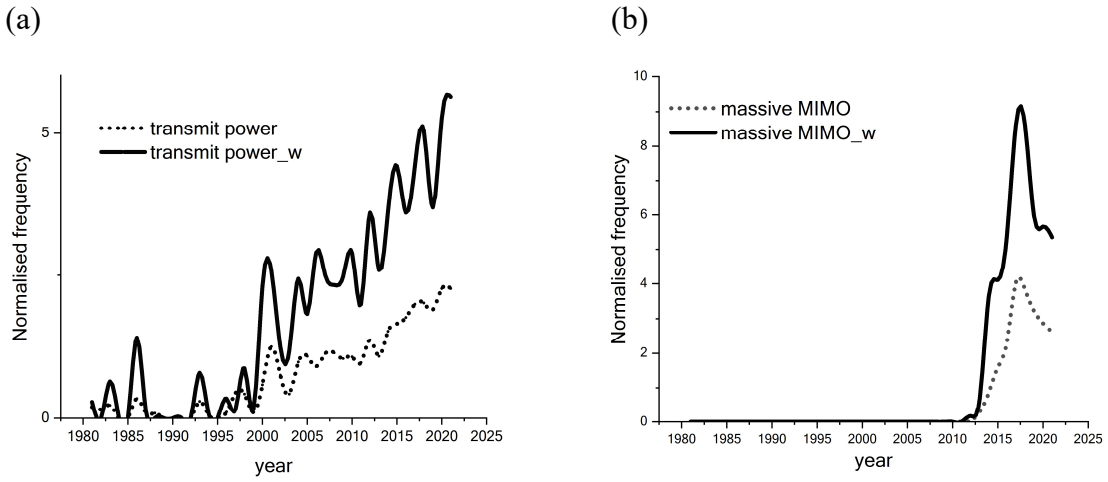


Figure 5.7 Normalised frequency changes of 'transmit power' and 'massive MIMO' by applying weights over time.
A plain line indicates weighted data, while a dotted line indicates time-series data without a weight.

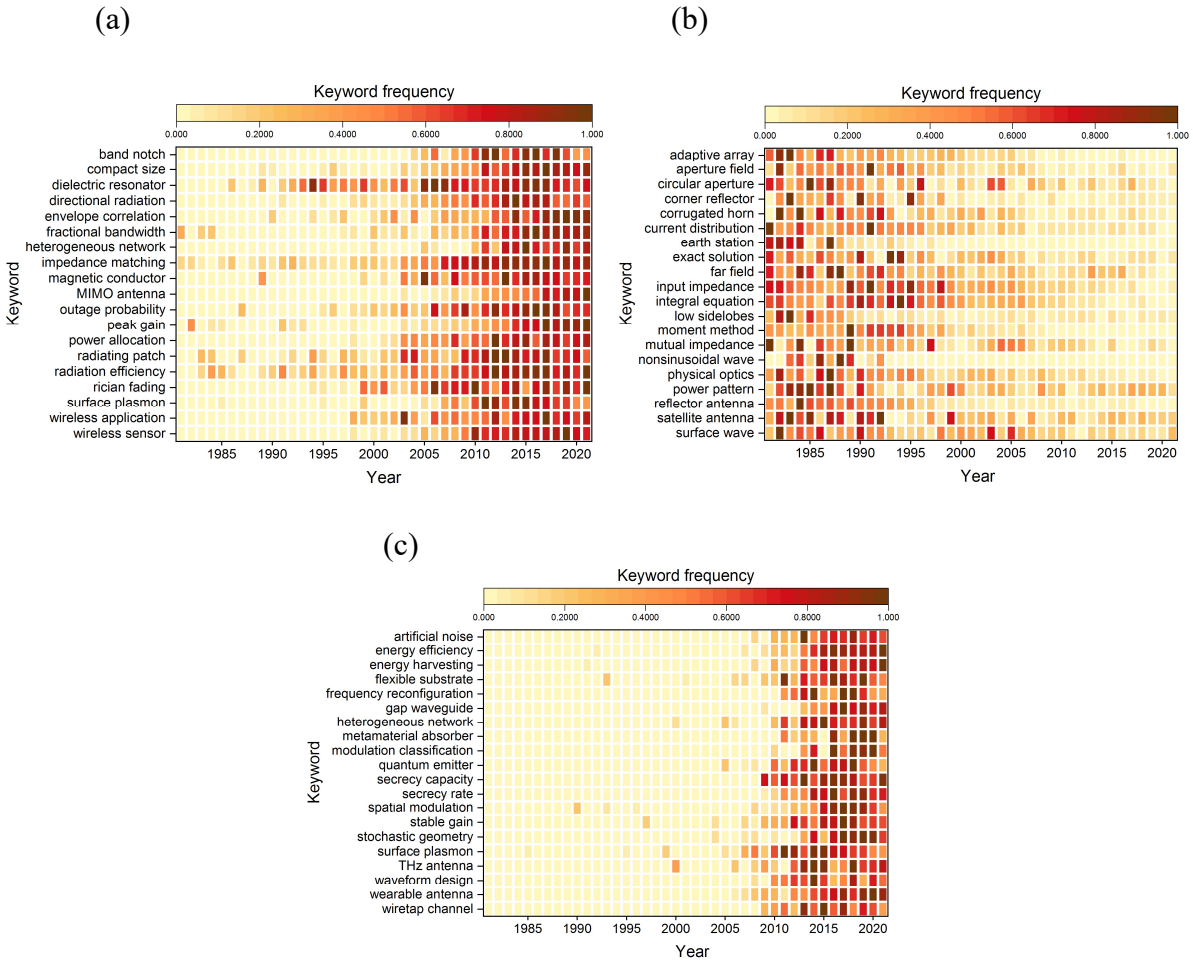


Figure 5.8 Heatmaps for Increasing, Decreasing, and Emerging trend keywords.

(a) Each keyword's frequency has been gradually rising since 1980. (b) Prior to the year 2000, each keyword was used frequently, but this has declined. (c) There is no data prior to the year 2005, but over the last 10 years, each keyword's frequency has increased significantly. Deeper colour indicates a higher frequency.

5.5 Learning with Attention Mechanism

The time-series data illustrated in Section 5.4 is now fed into an encoder-decoder LSTM network for the prediction of the future trend of keywords. The encoder-decoder architecture of recurrent neural networks (RNNs) has proven to be powerful for sequence-to-sequence-based prediction problems in various fields such as natural language processing, neural machine translation and image caption generation [17], [150], [151], [152]. The encoder-decoder LSTM network consists of three main components, namely, encoder, intermediate (encoder) vector and decoder. Encoder and decoder use a multi-layered LSTM unit to map

the input sequence to a vector of a fixed dimensionality, and then another LSTM is used to decode the target sequence from the vector [43]. These LSTM units are trained using the input data while maximising the conditional probability of the target sequence for a given sequence. One of the main drawbacks of this network is its incapability to extract strong contextual relations from long input series, that is if a particular piece of long time series data has some context or relations within its substrings, then a basic encoder-decoder model cannot identify those fixed-length contexts. This limitation could significantly deteriorate the predictive performance of the model. Attention mechanism [153], which is integrated with the LSTM, however, can address this problem. Adding the attention component to the network permits the decoder to utilise the most relevant parts of the input sequence in a flexible manner, by a weighted combination of all the encoded input vectors, with the most relevant vectors being attributed the highest weights [17], [28]. There are two most well-referenced attention mechanisms (by Bahdanua *et al.* [153] and Loung *et al.* [131]) in the literature and as there is a difference in the way in calculating attention scores. Bahdanua's attention mechanism is adopted in this study. Figure 5.9 illustrates the encoder-decoder LSTM with attention layer.

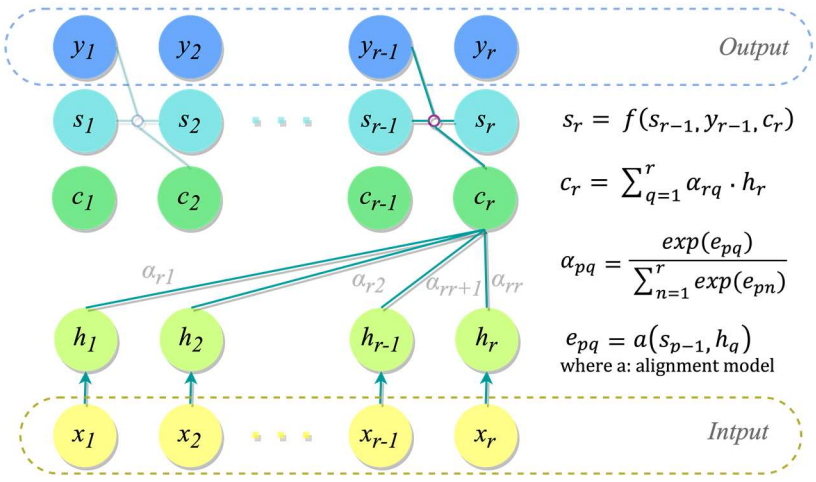


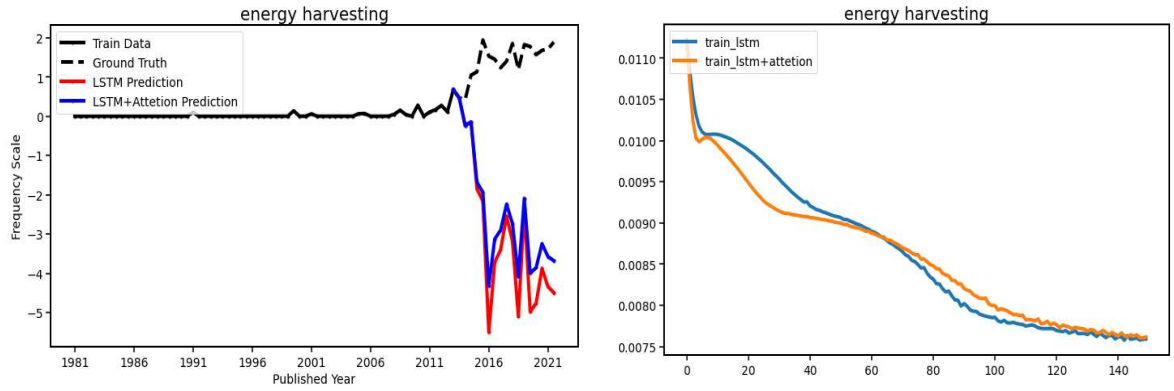
Figure 5.9 An explanation of the encoder-decoder LSTM with attention layer.

The input sequence (x_r) is fed into the encoder whose hidden states h_r) are exposed to the decoder via the attention layer. These states are weighted to give a context vector (c_r) that's used by the decoder. Attention weights (α) are calculated by aligning the decoder's last hidden state with the encoder hidden states. The decoder's current hidden state is a function of its previous hidden state, previous output word and the context vector. Attention is passed via the context vector, which itself is based on the alignment of encoder and decoder states. An alignment score quantifies how well output at position p is aligned to the input at position q. The context vector that goes to the decoder is based on the weighted sum of the encoder's LSTM hidden states h_q. These weights come from the alignment. The decoder's hidden state is based on its previous hidden state s_{r-1}, the previous predicted word and the current context vector. At each time step, the context vector is adjusted via the alignment model and attention. Thus, the decoder selectively attends to the input sequence via the encoder hidden states.

5.6 Model Validation and Prediction

In the previous section, the models are developed and trained to maximise the conditional probability of the target sequence for a given sequence. To achieve the maximum probability, the most data is required, however if annual data points from 1981 to 2021 are utilised, only 41 data points can be obtained. This amount is higher than the number used in the Chapter 3 study. However, especially for analysis of emerging keywords, as can be seen for the keyword 'energy harvesting' in Figure 5.10(a), the ground truth increases with positive frequency scales but makes negative predictions. The frequency scale is unable to produce negative results. There is no doubt that this is incorrect. As you can see in Figure 5.10(a), the majority of the data has values that are almost zero up until 2011, which means that these two thirds of the data are unusable for training. This incorrect validation results can be seen in the right of graph in Figure 5.10(a). Half-yearly data is used to avoid such data scarcity in training, particularly for keywords that have an emerging trend. There is a slight increase in noise with each doubling of the number of data points. A total of 82 data points is thus provided, which is then fed into the many-to-one encoder-decoder LSTM. This model which was trained using 82 data points gives better results, as shown in Figure 5.10(b).

(a)



(b)

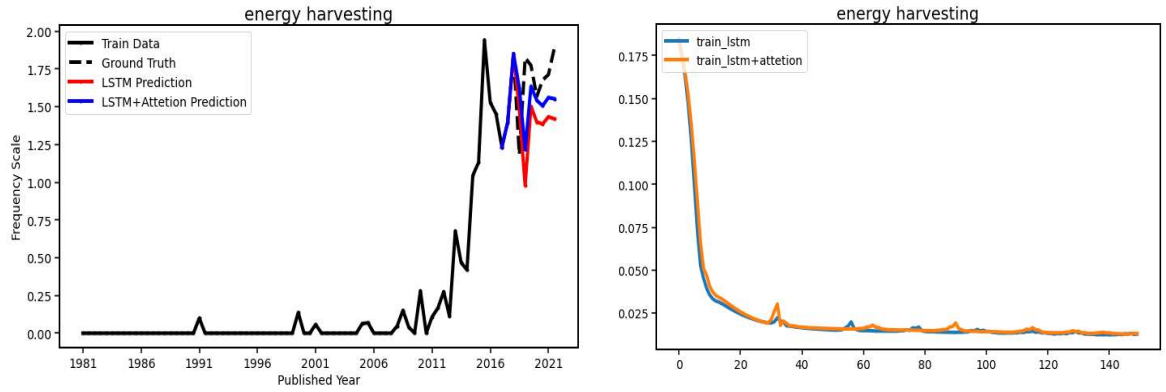


Figure 5.10 The results of training the encoder-decoder LSTM model by changing a number of data points.

(a) The 41 data points (annual data) were used. Since the majority of the data is zero, training the model did not go as expected. (b) Using the 82 data points from the half-yearly data, a model was properly trained.

To predict the research trend in the next 4 years, the time series data was fed into the many-to-one encoder-decoder LSTM, which is modified to have two LSTM units (*e.g.* encoder and decoder) and each unit is composed of 300 layers with rectified linear unit (ReLU). A series of every three frequencies is loaded into the trained model in the form of a sliding window in order to predict the following data points.

The validation result uses an average of Mean Square Errors (MSEs) less than 0.058, which is computed over 153 selected keywords. As shown in Figure 5.11, the attention layer we adopted finds a positive effect however with minor magnitudes. It is presumed that this magnitude could increase as the data scarcity is overcome. Figure 5.11 shows the validation

result of the keyword ‘reflectarray antenna’. Figure 5.12(a) and (b) depict the prediction result of the validated models with 8 steps ahead for the period of 2022 to 2026. Here, 2 steps ahead would mean 1 year, 8 steps mean 4 years. This applies to individual 153 keywords. In Appendix C.3, some prediction results for increasing trend keywords are given. The prediction of each of these antenna research keywords is then analysed and plotted using Gartner’s Hype Cycle [154], [155] in order to provide a holistic view of antenna research representing the maturity of new technologies in a simple and graphical way. This representation can give the readers strong hints and insights of which antenna technologies are potentially relevant to solving real problems and exploiting new opportunities.

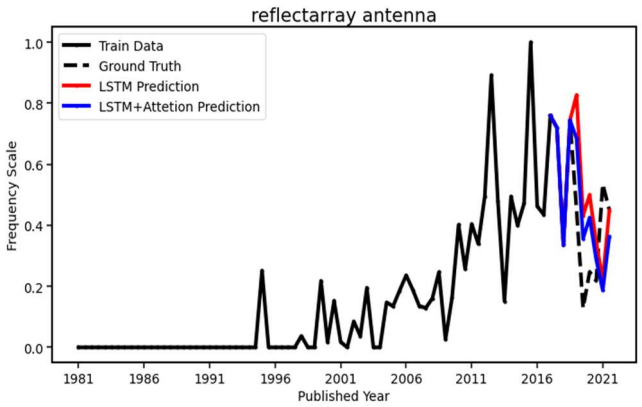


Figure 5.11 The validation results of encoder-decoder LSTM with attention layer models.

The Hype Cycle graph is divided into 5 key phases: Technology Trigger (TT), Peak of Inflated Expectations (PI), Trough of Disillusionment (TD), Slope of Enlightenment (SE) and Plateau of Productivity (PP). TT phase is the moment when new technologies become apparent or prominent. If a particular technology receives major attention with some success stories, it moves to the PI phase, despite doubts or reservations from the community. Some activities beyond early adopters can be found soon after a technology innovation enters the PI phase. Frequently, some negative press surfaces in the TD phase. Following the SE phase is that the technology innovation has now undergone a plenty of scrutiny with both failures and

successes, updates, and improvements for the industry to understand an optimal path of growth trajectory. Finally, the technology is readily produced and available as off-the-shelf solutions in the PP phase. From our analysis, however, in antenna research, some keywords do not slide into SE or PP phases, however, rather disappear. When the inflated expectations begin to die down via the phase of TD, they start to decrease, and this trend continues over the phases of SE and PP as indicated with a red line in Figure 5.12(c).

Among the topics which are triggered by technology, ‘secrecy capacity’, ‘super wideband’, ‘power management’, ‘heterogeneous network’, ‘secrecy rate’, ‘spatial modulation’, ‘stochastic geometry’, ‘firefly algorithms’ are directly linked to wireless communications. Many of these technologies require antennas to be directive and frequency agile and thus poise changes to the design, which novel materials and array architecture need to be developed in order to improve both power and spectral efficiency, as well as the security in wireless communications. The concurrent operation of Macro-, micro-, pico- and femto-cells is termed as heterogeneous networks (HetNets) [156], and it will require super-wideband antennas, and perhaps new architectures to enable emerging cell-free wireless communications. As for operating frequencies, the need for THz and optical antennas is still demanding, so is some fundamental requirement such as impedance matching and radiation efficiency etc.

Technologies related to ‘metamaterials’, ‘flexible materials’ and ‘wearable antennas’ may have reached their peaks, although challenges remain in upscaling and industrial uptake. Secured communications will require novel concepts and designs from a system aspect, such as ‘artificial noise’ and ‘envelope correlation’ built into physical layers, antennas, and RF systems. Interestingly, ‘effective permeability’ and ‘landau damping’ are two concepts respectively related to functional electromagnetic materials and plasma physics and are reaching the plateau of productivity, together well-known technologies of ‘reflector

antennas'. Technologies presented in the red curve have started disappearing, while those in the blue may also reach technology maturity, it remains to be key in A&P research community.

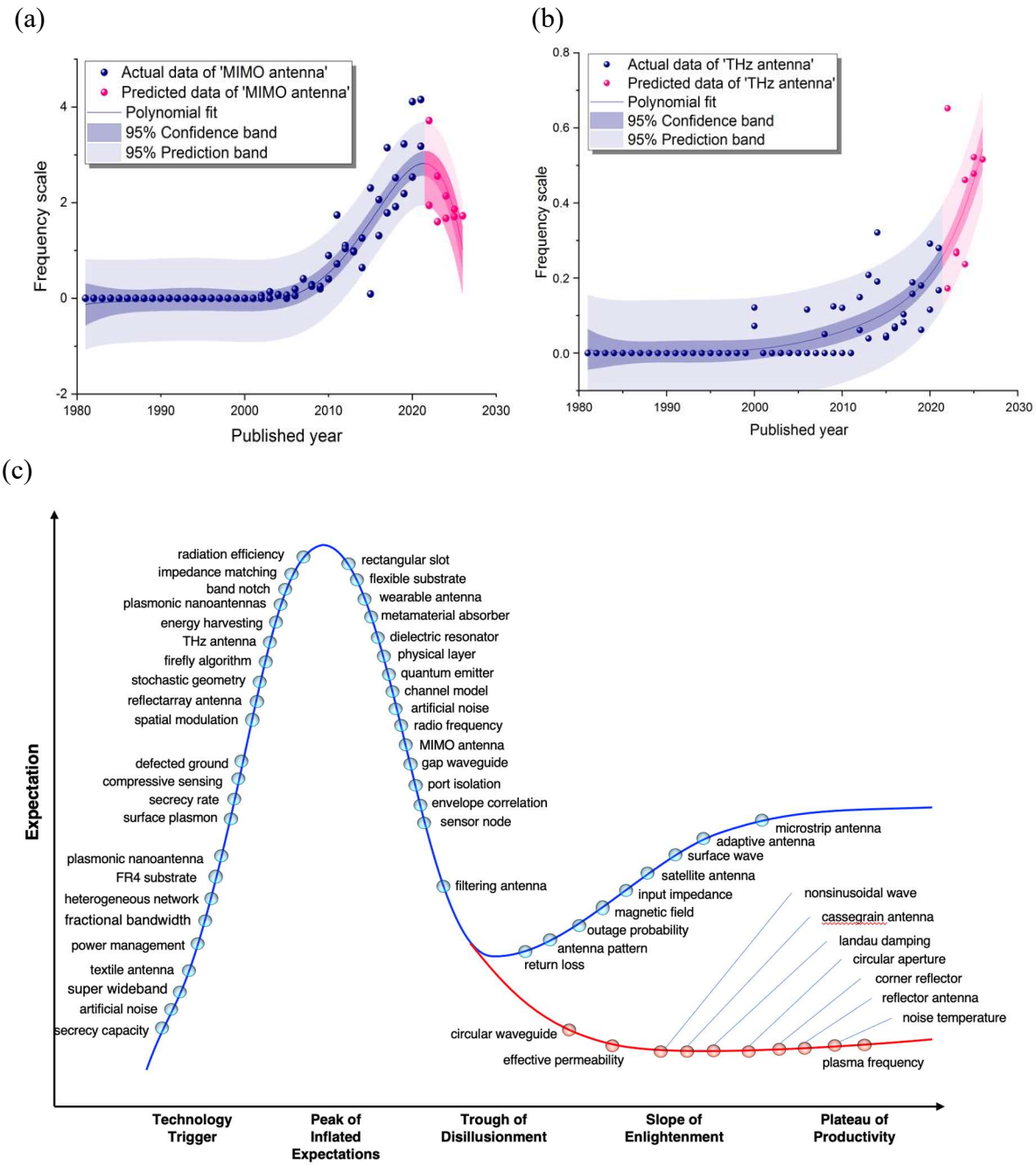


Figure 5.12 The results of prediction for next 4 years using encoder-decoder LSTM with attention models and the hype cycle in year of 2026.

5.7 Summary

To facilitate the advancement of antenna research, this chapter has successfully carried out a study of review and predict antennas and propagation research from large-scale unstructured data with machine learning. Geographic shifts in A&P publications have been observed, partially as a result of government investment, effort, and societal leaders' strategic goal to advance A&P research in developing and under-presented worlds. Meanwhile, this chapter provides A&P researchers with the provision of the information on how future research direction would change in a fully automated and unbiased manner based on openly published papers/books. In this chapter, recent advances in big data analytics, natural language processing and machine learning play key roles and they may benefit A&P researchers with the introduction of new subjects and cross-disciplinary knowledges. To the best of our knowledge, this is the first study in the antenna research utilising a large amount of unstructured data extracted from over 159k abstracts related to antenna research published between the years of 1981 and 2021, extracting informative patterns via keywords and using customised weighting via citation and SJR index. In a collective manner, the predictions by the proposed framework show a sensible and practical idea of what could be expected while representing the wholistic idea of the science literature. The importance of understanding the past and current technological trend and having an insight into future cannot be overemphasised. The predictive results visualised in the form of a Gartner's hype cycle successfully show what the technical trends will be in the next four years. This scientific attempt well constitutes quantitively addressing one of the important challenges within A&P community, the provision of the future research direction for the scientists in antenna research.

Chapter 6

Automated Tunable Materials Database Building for Reconfigurable Electromagnetics

Establishing a database is crucial for the discovery of new materials. Although data-driven approaches have been introduced, they come with their set of challenges. Chemical descriptors can often be vague, and in the absence of the right context, an automated system might misinterpret the information. Despite the availability of vast data sources, there's a notable shortage of annotated data. The efficiency of automated information extraction tools largely hinges on their training and the pre-defined chemical entities they recognise. If a novel or rare compound is introduced, the tool may fail to identify it. Taking these factors into account, this chapter highlights how an effective tool like ChemDataExtractor (CDE) [42] is enhanced by creating functions specifically designed to extract data on phase changing

materials (PCMs) and updating its model accordingly. To address the ambiguities in chemical language, domain-specific ontologies and knowledge bases are integrated to furnish more context.

For more than a century, many researchers in antenna and propagation (A&P) have endeavoured to create smaller, more affordable, and functional antennas that may be used for practical issues like 5G and 6G communications, satellite technology, and rising space technologies. New opportunities are presented by newly developed materials such as nanomaterials, graphene, and tunable materials. The development of tunability is crucial for the practical applications of antennas, thus materials capable of achieving tunability have garnered a significant amount of interest. The objective of this chapter is to develop a comprehensive database of tunable materials, including those utilising ferroelectrics and other PCMs. The database will serve as the basis for training a surrogate machine learning model that will enable efficient and effective exploration of the parameters associated with various material properties. The model will facilitate the identification of optimal combinations of parameters, accelerating the discovery and optimisation of materials with desirable properties. To achieve this, the state-of-the-art 'materials-aware' text-mining toolkit, CDE is modified that can extract target information from the literature. This database constitutes a valuable resource for identifying novel materials characterised by high tunability and low loss tangent. Furthermore, it has the potential to suggest designs for reconfigurable antennas and electromagnetics.

6.1 Introduction

Recently, there has been a greater demand for more compact antenna systems with reconfigurable features (frequency, radiation pattern, and/or polarisation) due to the increase in internet usage and fast development of wireless communication technologies [157], [158].

The development of reconfigurable features is a growing area of interest, highlighting the need for efficient discovery of tunable materials. By leveraging the power of artificial intelligence (AI), the discovery of tunable materials can be expedited through predictive synthesis and characterisation. This approach will facilitate the development of bespoke ferroelectrics and phase-change materials, while also enriching the dataset of tunable materials.

The first step in achieving the efficient discovery of tunable materials is to establish a comprehensive database by exhaustively querying the literature. Scientific articles related to tunable materials were collected from Elsevier and Springer sources and their full text was processed by converting it to plain text using the CDE. This plain text was then subjected to parsing and named-entity recognition (NER), a subtask of information extraction, to construct a database comprising material formulas, frequencies, and tunability information.

In version 2.0 of CDE, common user models for extracting the attributes are already defined and can be easily utilised by researchers, as depicted in Figure 6.1 [38]. Alternatively, researchers can define and apply custom models that are suitable for their specific project needs by inheriting these properties. The temperature model plays a significant role in the field of tunable materials research. It offers a valuable tool for acquiring vital data concerning the material's transition temperature and the specific testing conditions utilised during the experimental procedure. This information proves essential in the development and optimisation of tunable materials.

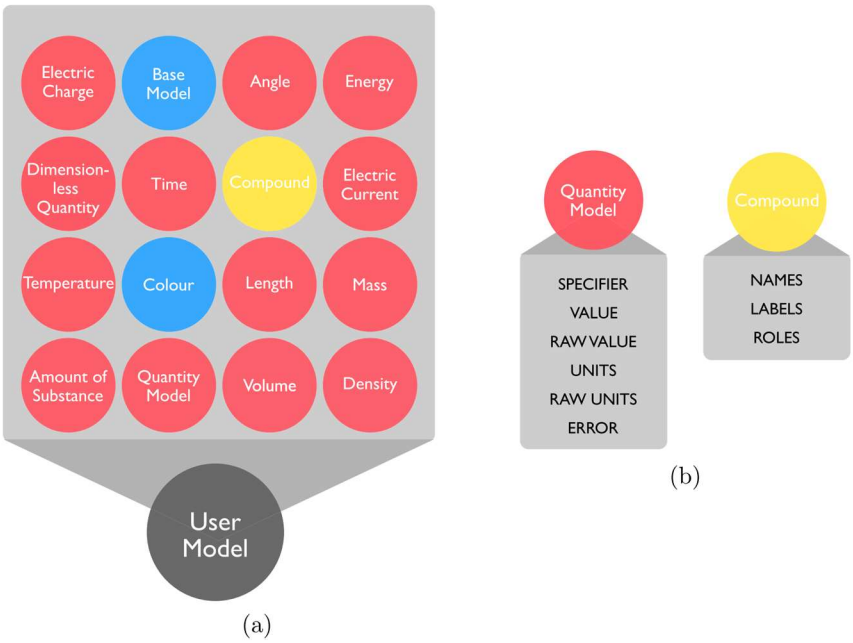


Figure 6.1 Model concept in ChemDataExractor 2.0 [47].

This database aims to compile data on microwave tunable materials, specifically those utilising barium strontium titanate (BST) and lead titanate (PT) ceramics, as well as phase change materials (PCM) with associated dopants. The information included in the database covers critical temperature (T_c), bandgap in the insulating or semiconducting phase, temperature-dependent electrical conductivity or resistivity near T_c, crystal structure details, and relevant references.

The modified CDE is to automatically extract and organise material compositions, along with their respective microwave tunability, loss tangent, and permittivity under specific conditions, including temperature, frequency, and biasing field. In addition, the NLP algorithm is provided with specific keywords such as ‘tunable’, ‘microwave’, ‘loss tangent’, ‘ferroelectric’, and ‘dielectric’, which are combined with logical conditions to facilitate efficient data extraction. Furthermore, to capture the properties of PCMs, keywords such as ‘VO_2’ and ‘GeSbTe’ are provided, taking into account the effects of dopants or substitutes.

All relevant references are meticulously recorded to enable subsequent manual validation of the extracted data, thereby ensuring accuracy and reliability.

PCMs are a particularly promising candidate to achieve modulation functionality because they allow high-speed reversible switching between its temperature stable structural phases. [159]. Moreover, PCMs have made a huge range of application possibilities feasible and have been deemed the most effective method for obtaining optical programmability[160], [161]. The optical properties of PCMs significantly change during the phase transition, which may be electrically or optically controlled. This has enabled the development of various types of programmable optical couplers, lenses, and metamaterials. [162], [163].

The most desirable PCMs are chalcogenide compounds (*e.g.* $Ge_2Sb_2Te_5$, GeTe, and Ge–Sb–Se–Te) and transition-metal oxides (*e.g.* VO_2 and NbO_2) [164]. Ge-Sb-Te alloys in particular, which belong to the chalcogenide family, are suitable for use in materials for optical data storage devices like digital videodiscs (DVDs) because they are stable in nature and don't require a power source to maintain their programmed state or recorded information. [165]–[167]. Meanwhile, vanadium dioxide (VO_2) is a potential reconfigurable and reprogrammable active optical PCMs due to its metal-insulator transition (MIT) [159]. The features of VO_2 have been extensively studied because of its adaptability. Examples of these applications include optical and electrical switches, variable attenuators, modulators, optical filters, infrared (IR) bolometers, and smart thermal radiator systems for spacecraft [168].

6.2 Data Collection

In contrast to previous chapters where only the abstracts of the scientific literature were required, this chapter calls for the entire text of the articles because the experimental details and results are typically reported in the methodology and results section of scientific articles.

CDE uses the ScienceDirect APIs to conduct keyword searches to retrieve and incorporate full-text content from ScienceDirect publications.

As seen in Figure 6.2, A DOI (Digital Object Identifier) list is then obtained as the search results and each relevant research paper from the DOI list can be downloaded as XML (Extensible Markup Language) files. XML is a set of standards for encoding documents in a format that is both human and machine readable.

Using the keyword ‘phase change materials’ as a search term, almost 2 million publications are found. It would be impossible to retrieve all of those works, so it was necessary to narrow the search area to increase accuracy. Using searched term ‘GST phase change materials tunable’, 455 xml files were collected. To determine whether the collected files were appropriate, only the sentences containing the word ‘transition temperature’ were extracted from the 455 files. Table 6.1 shows the results. Moreover, the names of compounds and successfully revealed temperature values (given in orange and red, respectively). However, as can be seen in Table 6.1, the extracted sentences contain both general descriptions of transition temperature and actual experimental data about transition temperature.

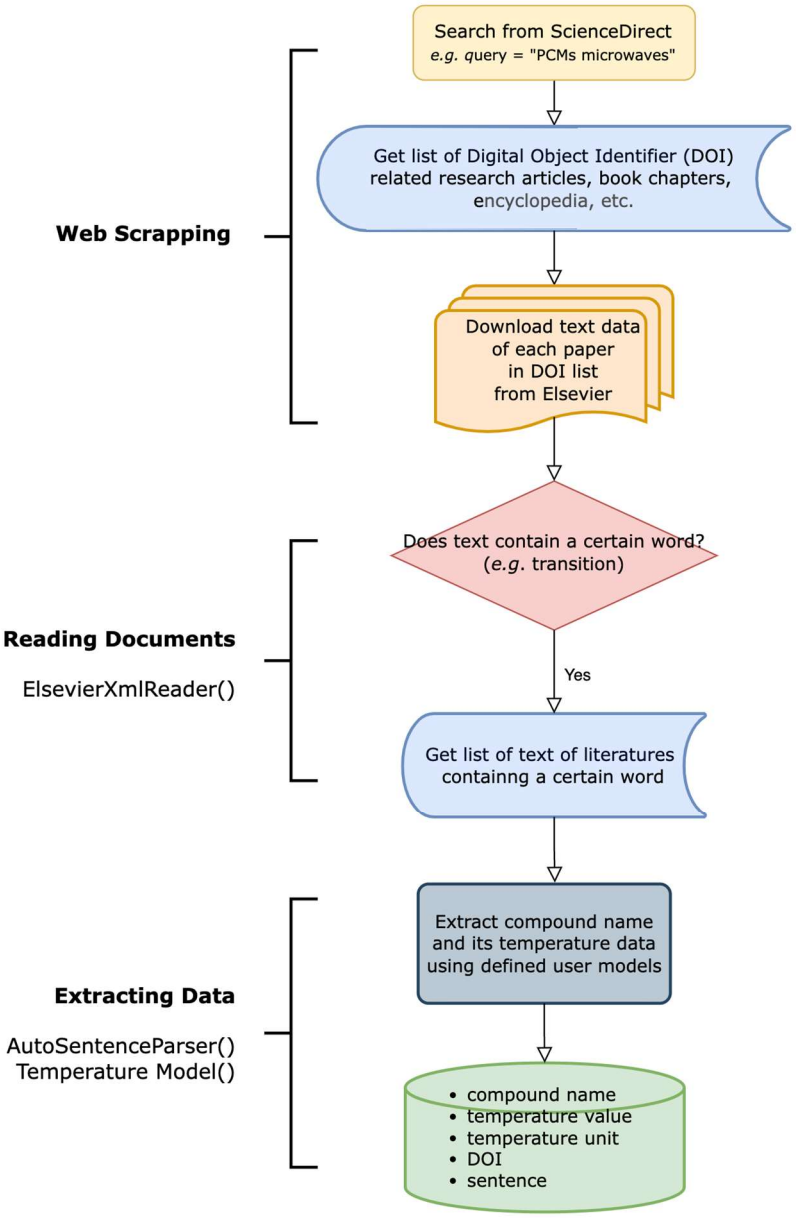


Figure 6.2 The flow chart of proposed framework with modified CDE [47].

When another search term 'GST phase change material microwave' was utilised, 658 papers were found. Figure 6.3 lists some of the results of the CDE parser related to the compound of VO_2 for dividing and classifying the compound names, their temperature data, specifier, and DOI. CDE uses a multi-step process to extract the chemical information from each document element before combining the data to create a single record for each chemical entity. Figure 6.3 demonstrates the effectiveness of CNE (Chemical Named Entity), and the

Chemical Records feature of CDE. However, this figure indicates that CDE needs to be modified in order to adapt this study and obtain the proper temperature data and its condition together.

Table 6.1 The results of search and collect text from relevant scientific literatures.

Once the metamaterial is heated above the **transition temperature** of VO_2, MP cannot be excited within the metallic VO_2 film, resulting in disappearance or "switch-off" of the absorption peak.
With a **transition temperature** of **68 °C**, thermochromic VO_2 undergoes a reversible phase change between an insulating state and a metallic state, accompanied by a large change in optical properties.
However, if the VO_2 becomes metallic above phase **transition temperature** upon heating, it will become a conductor instead of a capacitor between the Al disk and substrate.
The absorptance peak can be "switched off" by heating the metamaterial above the phase **transition temperature** of VO_2, allowing the metamaterials to act as a switchable absorber.
ELPs phase separate above a **transition temperature** (Tt), while remaining highly soluble below Tt.
Here, the material used for top layer was **perovskite manganite**, whose phase **transition temperature** can be customized without sacrificing optical tenability
Comparing with the existed **thermochromic DTEs, such as GST- and VO_2-based devices**, our deformable manganite-based resonator allows for the easy adjustment of the **transition temperature**
In order to examine the customizability of **transition temperature**, we prepared three samples by using the different perovskite layer, *i.e.* **$La_{0.7}Ca_{0.12}Sr_{0.18}MnO_3$ (LCSMO), $La_{0.7}Ca_{0.2}K_{0.1}MnO_3$ (LCKMO), and $La_{0.7}Sr_{0.3}MnO_3$ (LSMO).**
The **transition temperature** is weakly cell type and cooling rate dependent and ranges from **- 10 ~ for ND-1 through - 15 ~ for DU 145 to - 20 ~ for PC-3.**
SL-SPB and SL-SPC also showed clear differences in their preferred rotation axis in both lipids and at temperatures both above (**50'Â°C and below 125Â°C**) the gel-to-fluid **transition temperature** of DPPC or DPPG (410C).
In contrast to the case of the **Nb/ZnO/Pt device**, the observed superconductive **transition temperature** (Tc) of ~**5.8 K** is inconsistent with the Tc of **4.3 K** for the electrode element, but it agrees well with the **metallic Bi nanowire**.

No.	compound name	raw_value	raw_units	value	units	specifier	DOI	error	labels	text
1316	VOS	90	°C	[90.0]	Celsius^(1.0)	at	doi/10.1016/j.carbpol.2009.10.041			'The thermal treatment was processed in a programmec
144	VO2	68	°C	[68.0]	Celsius^(1.0)		doi/10.1016/j.ijthermalsci.2020.106754			'Therefore, considering the optical phase change temper
835	VO2	341	K	[341.0]	Kelvin^(1.0)		doi/10.1016/j.optlastec.2022.108245			'During the heating process, when the temperature reac
1155	VO2	300	K	[300.0]	Kelvin^(1.0)	at	doi/10.1016/j.optmat.2021.111745			'We verify the polarization conversion properties of the c
1156	VO2	25	°C	[25.0]	Celsius^(1.0)	at	doi/10.1016/j.optmat.2021.111745			'More specially, the VO2 film exhibits an insulating state
1264	VO2	300	K	[300.0]	Kelvin^(1.0)	at	doi/10.1016/j.spmi.2020.106653			'When the metasurface operates at T\xa0=\xa0300\xa0K
1265	VO2	400	K	[400.0]	Kelvin^(1.0)	at	doi/10.1016/j.spmi.2020.106653			'When the metasurface operates at T\xa0=\xa0400\xa0K
1266	VO2	300	K	[300.0]	Kelvin^(1.0)	At	doi/10.1016/j.spmi.2020.106653			'At T\xa0=\xa0300\xa0K, VO2 is insulator and thus the ef
1267	VO2	400	K	[400.0]	Kelvin^(1.0)	T	doi/10.1016/j.spmi.2020.106653			'The reflected half-wave plate is applied to transform th
1349	VO2	~67	°C	[67.0]	Celsius^(1.0)	at	doi/10.1016/j.mattod.2020.11.013			'VO2 breaks the confinement of common materials due t
1350	VO2	295	K(	[295.0]	Kelvin^(1.0)	at	doi/10.1016/j.mattod.2020.11.013			'The IR-coded Albert Einstein image is clearly seen for at
1386	VO2	300	K	[300.0]	Kelvin^(1.0)	T	doi/10.1016/j.infrared.2020.103440			'For T\xa0=\xa0300\xa0K, weak electric field resonance
1387	VO2	320	K(	[320.0]	Kelvin^(1.0)	T	doi/10.1016/j.infrared.2020.103440			'For T\xa0=\xa0320\xa0K (still lower than the phase char
1388	VO2	340	K	[340.0]	Kelvin^(1.0)	T	doi/10.1016/j.infrared.2020.103440			'When the temperature is increased to between T\xa0=\
1389	VO2	380	K	[380.0]	Kelvin^(1.0)	T	doi/10.1016/j.infrared.2020.103440			'For the temperatures T\xa0=\xa0380\xa0K and T\xa0=\v
1390	VO2	300	K	[300.0]	Kelvin^(1.0)	at	doi/10.1016/j.optcom.2018.05.085			'The absorption band from 0.76 THz to 0.86 THz at 300 K
1391	VO2	68	°C	[68.0]	Celsius^(1.0)		doi/10.1016/j.optcom.2018.05.085			'When the temperature is higher than the phase change
1392	VO2	350	K	[350.0]	Kelvin^(1.0)	T	doi/10.1016/j.optcom.2018.05.085			'Conductivity of VO2 is set to be frequency independent
1393	VO2	300	K	[300.0]	Kelvin^(1.0)	at	doi/10.1016/j.optcom.2018.05.085			'However, at 300 K no magnetic response can be observe
1394	VO2	350	K	[350.0]	Kelvin^(1.0)	at	doi/10.1016/j.optcom.2018.05.085			' (c) and (d) shows that at 350 K the resonance at 1.15 TH
1796	VO2	68	°C	[68.0]	Celsius^(1.0)		doi/10.1016/j.optcom.2017.06.068			'During thermal phase transition process, when the outsi
1797	VO2	50	°C	[50.0]	Celsius^(1.0)		doi/10.1016/j.optcom.2017.06.068			'When the environment temperature is 50\xa0°C below t
1798	VO2	80	°C	[80.0]	Celsius^(1.0)		doi/10.1016/j.optcom.2017.06.068			'When the operating temperature is 80\xa0°C which is hi
1799	VO2	68	°C	[68.0]	Celsius^(1.0)		doi/10.1016/j.optcom.2017.06.068			'When the operating temperature is below 68\xa0°C, the
1800	VO2	50	°C	[50.0]	Celsius^(1.0)		doi/10.1016/j.optcom.2017.06.068			'And When the environment temperature is 50\xa0°C be
1993	VO2	300	K	[300.0]	Kelvin^(1.0)	at	doi/10.1016/j.physe.2021.114630			'In this set of simulations, the conductivity of VO2 particl

Figure 6.3 The example of extracting results with search term 'GST phase change material microwave'.

6.3 Results

To further narrow the focus of the search and reduce the risk of retrieving erroneous data, the words 'ferroelectric & tunable & microwave & loss & tangent' were set to the search terms. Subsequently, 'GeSbTe phase change' and 'VO_2 phase microwave' were chosen as the keywords and connected by the 'AND' operation. This enabled a more targeted search, allowing for the identification of pertinent data without needing to review a large number of sources.

The results from the ScienceDirect API search are outlined in Table 6.2, while the information extracted using the modified CDE are presented in Table 6.3. Temperature values, units, compounds, texts and DOI were all automatically collected. However, as indicated in Table 6.3, different types of temperatures were combined, thus necessitating further post-processing. The user must then manually go through all the selected texts containing the temperature data in order to select the relevant information.

Table 6.2 The number of literatures and temperatures data of colleting papers.

Search keyword	No. of literatures	No. of temperature data
GeSbTe phase change	298	745
VO_2 phase microwave	310	651

To obtain sufficient meaningful data, identifying out the compositions of the materials should be the starting point. Regarding one of the promising PCMs, VO_2, it is desirable to have different VO_2 transition temperatures for practical use in to satisfy the various demands based on the applications. As an example, thermochromic smart windows are suitable for phase transition temperatures (T_c) of around 30 °C [169], whereas solar thermal collector optimisation demands T_c above 110 °C [170]. Hence, modulating T_c has been the object of intense research and doping is the most popular and effective method for achieving this goal

[171][172]. In order to find information about doping in the VO_2 keyword search, it was decided to add the word 'doped' in this study.

Table 6.3 The number of literatures correspond to the search terms.

Search keyword	*No. of literature*	*No. of sentence*
VO_2 phase change doped	963	
VO_2 phase change doped + (mol or vol) %	165	648
VO_2 phase change doped + (mol or vol) % + T_c (transition temperature)	49	95

Table 6.4 The 14 extracting sentences of PCMs and their transition temperature data.

0	doi_10.1016_j.ceramint.2013.04.016.xml	shows the XRD patterns of W-doped VO_2 prepared at different temperatures for 48 h using 0.91 g of V_2O_5, 1.26 g of $H_2C_2O_4{\cdot}2H_2O$ and 1.0 at% of tungstic acid.
1	doi_10.1016_j.solmat.2021.111055.xml	The temperature-dependent transmittance of VO_2 (M) doped with 0, 0.25, 0.5, and 0.75 mol% of tungsten ions was monitored at a wavelength of 1350 nm (
2	doi_10.1016_j.solmat.2021.111055.xml	The 0.25 mol% W-doped VO_2 (M) NCs exhibited similar transmittance and phase transition behaviors to those of undoped VO_2(M) NCs, while the phase transition temperature was reduced to 50 °C.
3	doi_10.1016_j.solmat.2021.111055.xml	The 0.5 mol% W-doped VO_2 (M) NC films showed exceptional optical transition properties: a Tlum of 59.57% at 20 °C and a ΔTsol of 15.31%, along with a reduced phase transition temperature of 37 °C.
4	doi_10.1016_j.solmat.2021.111055.xml	The 0.75 mol% W-doped VO_2 (M) NC films also showed high Tlum (61.31% at 20 °C) and ΔTsol (10.18%) values, while their phase transition temperature was reduced to 32 °C.
5	doi_10.1016_j.ceramint.2014.04.113.xml	The phase transition temperatures of 3 at% Mo and 4 at% W co-doped VO_2 (R), 7 at% Mo and 8 at% W co-doped VO_2 (R) as well as 9 at% Mo and 10 at% W co-doped VO_2 (R) are 36.25 °C, 19.50 °C and 41.27 °C in the heating cycles and 32.44 °C, 15.69 °C and 35.878
6	doi_10.1016_j.jmst.2019.10.022.xml	The metal-insulator transition (MIT) temperature increases to above 380 K when the TiO_2 ratio of the source is 5 at.%, although the Ti source is not physically doped into VO_2 nanobeams.
7	doi_10.1016_j.scriptamat.2019.09.019.xml	The temperature dependence of the electrical resistivity of the 0.8 at.%B and 1.6 at.%B doped VO_2 films is shown in
8	doi_10.1016_j.apsusc.2022.154519.xml	The difference is that the solid–liquid endothermic phase transition temperature decreases by ~1 °C, while the liquid–solid exothermic phase transition temperature increases by ~1 °C compared to those of pure paraffin due to the presence of W-doped VO_2 nanoparticles and chitosan when the W-doped VO_2 nanoparticle-to-paraffin mass ratio reaches 1.25%.
9	doi_10.1016_j.jallcom.2012.07.093.xml	The optical switching properties of W-doped VO_2 (M) (1.5% of W) were first investigated by a series of variable-temperature infrared spectra of heating and cooling, as shown in
10	doi_10.1016_j.apsusc.2021.148937.xml	Particularly, the 1.3 at% W-doped VO_2 thin films show Tlum and ΔTsol values of 53.0% and 10.0%, respectively, when the phase transition temperature is reduced to 29 °C; these are the highest values reported so far for flexible VO_2(M) thin films with the Tc at a near-ambient temperature.
11	doi_10.1016_j.apsusc.2021.148937.xml	In particular, the 1.3-at% W-doped VO_2 thin film showed a **Tc** of 29 °C with Tlum and ΔTsol values of 53.1% and 10.0%, respectively, which are the highest Tlum and ΔTsol values reported for W-doped VO_2 thin films with a Tc near ambient temperature.
12	doi_10.1016_j.solmat.2019.04.022.xml	The nominally 3 at.% W-doped VO_2/PVP coating displayed an acceptable transition temperature of 53.6 °C, which moves the material closer to an application in smart windows where a **Tc** closer to 30 °C is required.
13	doi_10.1016_j.ceramint.2021.02.133.xml	It is found that Zn-doped VO_2 not only exhibits excellent solar modulation ability (ΔTsol = 15.27%) but also can reduce the phase transition temperature and increase the visible light transmittance after the heat-induced phase transition (ΔTlum=+5.78%).
14	doi_10.1016_j.ceramint.2021.02.133.xml	Moreover, it has been clarified that Zn doping causes lattice distortion and generation of oxygen vacancies, resulting in a linear decrease in phase transition temperature Tc (~-0.49 K/at% Zn) with the increase of Zn atom doped concentration and also a decrease in the width of the thermal hysteresis loop (ΔT).

The keyword 'VO_2 phase change doped' was used, and 963 papers were collected, as can be seen in the first row of Table 6.3. If there are experiment volume percentage details in any of the 963 papers, that one is chosen. 648 sentences are then taken from 165 papers. Then, 49 papers with 95 sentences were returned after the transition temperatures were added to the search. In regard to the amount of literature found, this method is comparable to the manual search. Table 6.4 displays 14 of the final 95 sentences that were chosen, each with the information 'VO_2 phase change doped % T_c'.

This chapter has laid down some solid foundations for discovery because it can collect data on tunable materials, crucial materials, even though the results obtained have not yet been fully automatically extracted from the results of the experiment or the conditions of the experiment. Additionally, as this study is still only a test to confirm the usefulness of CDE, if it is used to further modify CDE in the future, this database enables the prediction of key properties of interest, such as critical temperature (T_c) and bandgap. Moreover, through the use of machine learning (ML), it becomes possible to perform a combinatorial analysis of a vast array of dopants and their respective ratios, thereby allowing for the customisation of PCMs with a desired bandgap and a sharp phase transition.

6.4 Summary

This chapter has explored the use of a modified information extraction tool to build a comprehensive database of tunable materials and their properties. Using the ScienceDirect API, relevant scientific papers were collected, and full-text content was retrieved. XML files were then downloaded, and the modified CDE was used to extract information.

As a result, the database was able to provide a range of materials and their properties, allowing for a more comprehensive understanding of the features of tunable materials. Additionally, the addition of the keyword 'doped' enabled the identification of papers related

to the doping of VO_2, leading to the gathering of pertinent data in relation to the manipulation of transition temperature.

This study has demonstrated the effectiveness of automated information extraction tools, such as CDE, in the field of tunable materials and their properties. Moving forward, it is imperative to expand the scope of research to encompass not only textual data but also tabular and image data. This will enable the extraction of valuable information from a wider range of sources, leading to a more comprehensive and insightful database.

Chapter 7

Summary, Challenges and Future Work

7.1 Summary

Natural language processing (NLP), a sub-field of artificial intelligence (AI), has made significant advances in a range of applications such as machine translation, question answering, sentiment analysis, and information retrieval. Material science researchers are beginning to recognise the value of AI approaches for discovering unknown properties of materials or for discovering new materials. Utilising such techniques not only saves time and resources, but also eliminates the potential for biased insight that comes with human judgement. The success of these AI-based methods, however, is highly dependent on the quality of input data. It is well established that the collection of accurate input data is of primary importance in both traditional experimental and computational methods.

Furthermore, a significant amount of input data is required in order to train AI models

effectively. Big data, or data with a greater variety and faster velocity, has recently gained attention for solving various issues in material science using AI tools. Scientists can now extract high-quality data from vast amounts of freely accessible unstructured information thanks to recent developments in NLP. Scientific journal publishers have also digitised their collections and resources, making them accessible in application programming interfaces (APIs), computer-readable html/xml format for users and developers. Despite being a reliable source of data, peer-reviewed scientific literature is unstructured and heterogeneous, which makes it difficult to analyse the data on a large scale.

Even though several projects have already examined the potential of NLP in this domain, its implementation is still in its formative stages, particularly when it comes to extracting data from unstructured sources (*e.g.* academic papers), processing it, and generating new knowledge. This thesis incorporates the current NLP techniques to facilitate automated document analysis, information extraction, and the creation of a database for new material discovery. The following is a list of the innovative contributions made by this thesis.

- The author proposed a state-of-the-art NLP method that enables an automatic keywords extraction from unstructured texts, analyses of publicly available literatures in metamaterial research. This approach can generate a comprehensive pool of keywords and uncover trends and relationships between each keyword in the pool.
- The author developed a modified recurrent neural network (*e.g.* encoder-decoder LSTM) to forecast future research directions in metamaterials science. This method can not only help to gain a greater insight into the influence of metamaterials research, but also laying a strong groundwork for the formation of a Gartner hype cycle-based metamaterials research roadmap.

- The authors suggested a framework which leverages an advanced text summarisation technique to increase the thoroughness of review papers. To validate this approach, experiments were conducted, and the results were compared to well-referenced datasets for text summarisation.
- The author designed a novel weighting system to evaluate article attributes for antenna and propagation research, enabling predictions of future trends. This approach extends beyond basic numerical analysis, enabling a more accurate assessment of each paper's actual impact on the field.
- The author presented the effectiveness of automated information extraction tools, such as ChemDataExtractor, in the domain of tunable materials and their properties. This automated database building approach can provide an extensive and precise overview of the tunable materials research field, which holds potential to enable tunability in antennas design for various applications.

7.2 Challenges and Overcomes

Despite adopting advanced NLP technology has provided numerous advantages in the field of material science, there are still many obstacles to be overcome before establishing a fully automated framework. Generally, the terminology used in the field of materials science is very intricate and technical, making it challenging for NLP models to understand. This is due to the fact that the vocabulary frequently consists of specialised terms and concepts that many NLP models might not be familiar with. Inconsistencies in terms and definitions could result from the fact that the data may come from various sources. As a result, using NLP to interpret and comprehend the vocabulary of materials science can be difficult. Also, there are unstructured data and numerous formats such as text, images and tables for which data in the

field of materials science may be accessible. So, tabular or image data should be converted and utilised to train the NLP model in the same way that text data is. In chapter 3 and 5 of the thesis, a methodology is proposed to address this limitation. Specifically, unstructured keyword data is transformed into a vectorised format to facilitate visual analysis. The resulting numeric representation, consisting of normalised frequencies of each keyword, is then utilised as input to a sequential neural network model for prediction.

Data labelling and annotation, which are steps in the construction of NLP models, constitute another problem. Manually labelling data for materials science can be challenging and time-consuming, though. This is because the data is frequently technical and extremely specialised, necessitating an in-depth knowledge of the subject in to correctly tag it. This study faced certain limitations, particularly in the mixed-level nature of the numerous extracted keywords, which included both the object and its associated properties, making it difficult for a human expert to label which one is object or property. To address this challenge, the study employed the RAKE algorithm to extract key phrases rather than isolated keywords. These key phrases typically consisted of the form ‘object + property’. The extracted key phrases were then validated by comparing them to the title and author's keywords reported in the literature. This approach resulted in a more precise, albeit smaller set of keywords. To provide a high-level overview without manual labelling, the study employed a clustering method.

Embeddings are important for NLP models, but there is currently limited availability for embeddings that are specific to materials science. This is because embeddings are typically created using large amounts of data, and there is not yet enough materials science data to create embeddings that are specifically tailored to the subject. Additionally, most existing embeddings are too general to accurately capture the nuances of materials science. As such, the limited availability of domain-specific embeddings can be a challenge when using NLP in

materials science. The challenge identified in the aforementioned study pertains to the inadequate embedding of two-word keywords that hinders capturing their complete contextual significance. Unlike one-word keywords that can be readily vectorised, the lack of appropriate embeddings for two-word keywords posed a challenge. To tackle this, the study proposed a solution that involves individual vectorisation of each word in the two-word keyword, followed by their addition utilising the word embedding analogy. This approach enables vector-oriented reasoning based on the offsets between the pair of words and can overcome the challenge of inadequate embedding for two-word keywords.

7.3 Future Work

7.3.1 Technology Trend Prediction

In this thesis, keywords were extracted from published papers, and their frequency was transformed into a numerical sequence data format to make predictions several years later. To validate the accuracy of the predictions, further research is required to confirm their actual outcomes at the predicted point in time. In the event of inaccurate predictions, further investigations are warranted to identify potential areas of improvement that could enhance the precision of future predictions.

The thesis predicted technological trends solely based on the occurrence of keywords. However, an analysis of how keywords have evolved over time, such as examining the properties with which the object keyword is combined or separated from, can provide valuable insights for predicting trends. As illustrated in Figure 7.1, conducting a detailed analysis of keywords using Neo4j [173] and organising the relationships between objects and their properties over time would be beneficial.

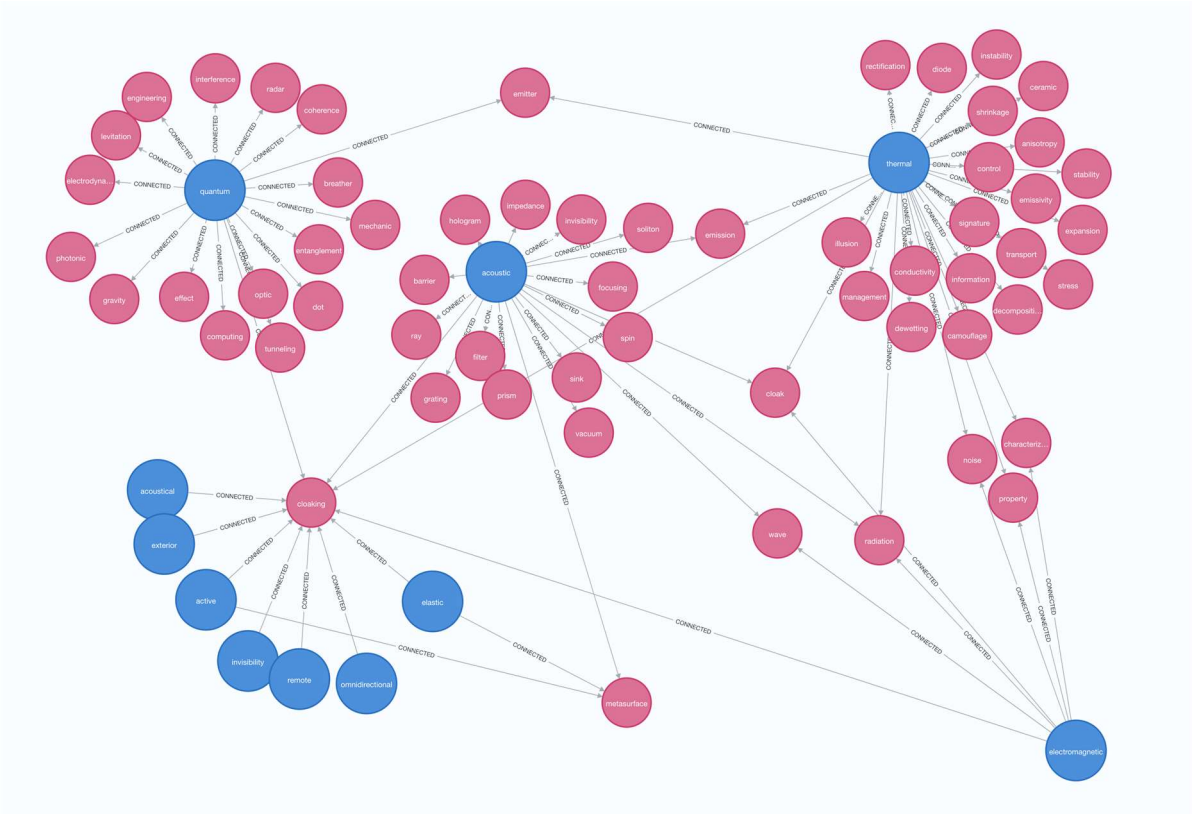


Figure 7.1 Visualising the relationship of keywords in metamaterials research using Neo4j.

This visualisation depicts the connections between individual words in a compound word, with the blue node representing the first word and the red node representing the second word. For instance, in the top-left cluster, ‘quantum emitter’ is shown. The visualisation reveals that ‘emitter’, the node in the top-middle position, is also linked to ‘thermal’.

Moreover, Figure 7.2 successfully demonstrates the progress of a particular technology's advancement by effectively retrieving pertinent published literature. This study focuses on the evolution of the water-energy-food (WEF) nexus research, highlighting the development of the concept and its associated key terminology [174]. Additionally, the research establishes linkages among various WEF nexus topics, providing a comprehensive overview of the phenomenon's interconnections. On the basis of this idea, it would be beneficial to examine past trends in material science technology and predict how they will evolve in the years to come. Such an investigation would facilitate researchers in gaining a deeper understanding of how technology is likely to advance, ultimately proving advantageous to their work.

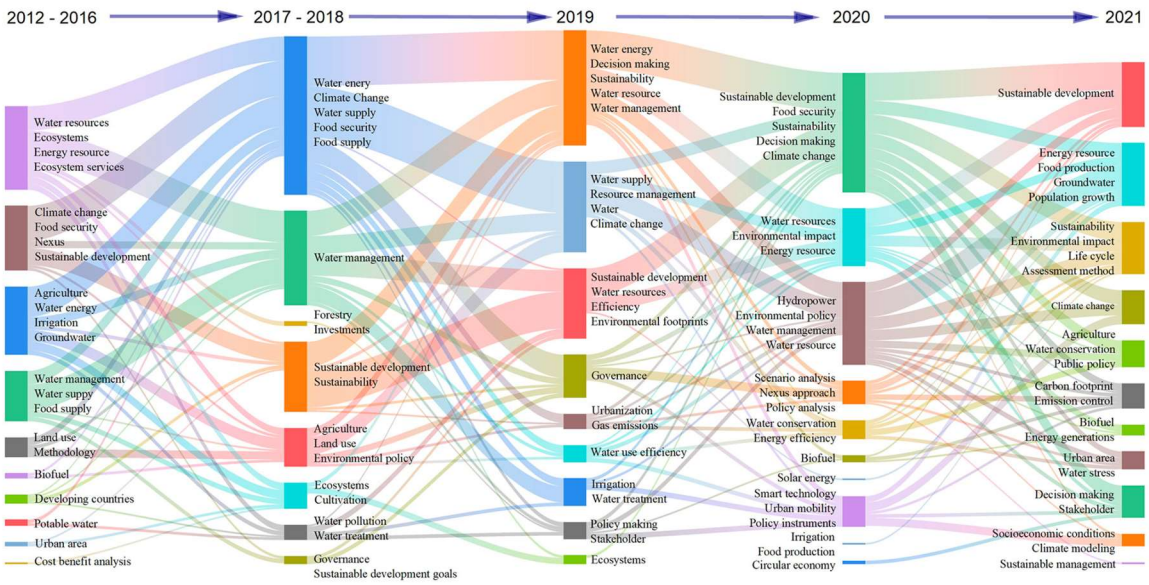


Figure 7.2 Sankey diagram of the thematic evolution on the WEF nexus research (2012–2021). [174]

7.3.2 Automate Database Building from Various Data Format

As evidenced in chapter 6, the current research has successfully demonstrated the capability to extract information solely from textual sources provided by CDE. However, there is a pressing need to expand the scope of data extraction to include not only text but also tables and images. Although state-of-the-art algorithms for table parsing have been developed, data extraction methods from tables and figures are still in their nascent stages and require further improvement [46]. Additionally, efforts will be directed towards developing and optimising algorithms, such as TableDataExtractor [47] and ImageDataExtractor [175]. This is because valuable material related to the target materials exists not only in text but also in other sources, and a more comprehensive approach to data extraction is necessary to fully capture the richness of the available information.

Reference

[1] D. Khurana, A. Koli, K. Khatter, and S. Singh, "Natural language processing: state of the art, current trends and challenges," *Multimed Tools Appl*, vol. 82, no. 3, pp. 3713–3744, 2023, doi: 10.1007/s11042-022-13428-4.

[2] R. M. Samant, M. R. Bachute, S. Gite, and K. Kotecha, "Framework for Deep Learning-Based Language Models Using Multi-Task Learning in Natural Language Understanding: A Systematic Literature Review and Future Directions," *IEEE Access*, vol. 10, pp. 17078–17097, 2022, doi: 10.1109/ACCESS.2022.3149798.

[3] X. Chen, H. Xie, and X. Tao, "Vision, status, and research topics of Natural Language Processing," *Natural Language Processing Journal*, vol. 1, p. 100001, 2022, doi: https://doi.org/10.1016/j.nlp.2022.100001.

[4] Z. Hong, L. Ward, K. Chard, B. Blaiszik, and I. Foster, "Challenges and advances in information extraction from scientific literature: a review," *JOM*, vol. 73, no. 11, pp. 3383–3400, 2021.

[5] K. Adnan, R. Akbar, and K. S. Wang, "Information extraction from multifaceted unstructured big data," *International Journal of Recent Technology and Engineering (IJRTE)*, vol. 8, pp. 1398–1404, 2019.

[6] P. Johri, S. K. Khatri, A. Al-Taani, M. Sabharwal, S. Suvanov, and A. Chauhan, "Natural Language Processing: History, Evolution, Application, and Future Work," 2021, pp. 365–375. doi: 10.1007/978-981-15-9712-1_31.

[7] W. Hu, J. Zhang, and N. Zheng, "Different Contexts Lead to Different Word Embeddings," in *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, Osaka, Japan: The COLING 2016 Organizing Committee, Dec. 2016, pp. 762–771. [Online]. Available: https://aclanthology.org/C16-1073

[8] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *Journal of the American society for information science*, vol. 41, no. 6, pp. 391–407, 1990.

[9] E. Cambria and B. White, “Jumping NLP curves: A review of natural language processing research,” *IEEE Computational Intelligence Magazine*, vol. 9, no. 2. Institute of Electrical and Electronics Engineers Inc., pp. 48–57, 2014. doi: 10.1109/MCI.2014.2307227.

[10] J. Hirschberg and C. D. Manning, “Advances in natural language processing.” [Online]. Available: http://science.sciencemag.org/

[11] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, “Distributed Representations of Words and Phrases and their Compositionality,” Oct. 2013, [Online]. Available: http://arxiv.org/abs/1310.4546

[12] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient Estimation of Word Representations in Vector Space,” Jan. 2013, [Online]. Available: http://arxiv.org/abs/1301.3781

[13] S. Sivakumar, L. S. Videla, T. R. Kumar, J. Nagaraj, S. Itnal, and D. Haritha, “Review on Word2Vec Word Embedding Neural Net,” in *2020 International Conference on Smart Electronics and Communication (ICOSEC)*, 2020, pp. 282–290. doi: 10.1109/ICOSEC49089.2020.9215319.

[14] M. Gutmann and A. Hyvärinen, “Noise-contrastive estimation: A new estimation principle for unnormalized statistical models,” in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, JMLR Workshop and Conference Proceedings, 2010, pp. 297–304.

[15] J. Pennington, R. Socher, and C. D. Manning, “GloVe: Global Vectors for Word Representation.” [Online]. Available: http://nlp.

[16] W. Zaremba, I. Sutskever, and O. Vinyals, “Recurrent Neural Network Regularization,” Sep. 2014, [Online]. Available: http://arxiv.org/abs/1409.2329

[17] I. Sutskever, O. Vinyals, and Q. v. Le, “Sequence to Sequence Learning with Neural Networks,” Sep. 2014, [Online]. Available: http://arxiv.org/abs/1409.3215

[18] C. Olah, “Understanding lstm networks,” 2015.

[19] K. Cho, B. van Merrienboer, D. Bahdanau, and Y. Bengio, “On the Properties of Neural Machine Translation: Encoder-Decoder Approaches,” Sep. 2014, [Online]. Available: http://arxiv.org/abs/1409.1259

[20] A. Vaswani *et al.*, “Attention Is All You Need.”

[21] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” Oct. 2018, [Online]. Available: http://arxiv.org/abs/1810.04805

[22] A. R. Openai, K. N. Openai, T. S. Openai, and I. S. Openai, “Improving Language Understanding by Generative Pre-Training.” [Online]. Available: https://gluebenchmark.com/leaderboard

[23] T. B. Brown *et al.*, “Language Models are Few-Shot Learners,” May 2020, [Online]. Available: http://arxiv.org/abs/2005.14165

[24] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, “DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter,” Oct. 2019, [Online]. Available: http://arxiv.org/abs/1910.01108

[25] P. F. Brown *et al.*, “A STATISTICAL APPROACH TO MACHINE TRANSLATION.”

[26] P. Williams, R. Sennrich, M. Post, and P. Koehn, “Syntax-based Statistical Machine Translation,” *Synthesis Lectures on Human Language Technologies*, vol. 9, no. 4, pp. 1–211, 2016, doi: 10.2200/S00716ED1V04Y201604HLT033.

[27] D. Bahdanau, K. Cho, and Y. Bengio, “Neural Machine Translation by Jointly Learning to Align and Translate,” Sep. 2014, [Online]. Available: http://arxiv.org/abs/1409.0473

[28] K. Cho *et al.*, “Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation,” Jun. 2014, [Online]. Available: http://arxiv.org/abs/1406.1078

[29] B. Klimova, M. Pikhart, A. D. Benites, C. Lehr, and C. Sanchez-Stockhammer, “Neural machine translation in foreign language teaching and learning: a systematic review,” *Educ Inf Technol (Dordr)*, 2022, doi: 10.1007/s10639-022-11194-2.

[30] N. Jain, M. Popovic, D. Groves, and E. Vanmassenhove, “Generating Gender Augmented Data for NLP,” Jul. 2021, [Online]. Available: http://arxiv.org/abs/2107.05987

[31] Y. Peng, S. Yan, and Z. Lu, “Transfer learning in biomedical natural language processing: an evaluation of BERT and ELMo on ten benchmarking datasets,” *arXiv preprint arXiv:1906.05474*, 2019.

[32] J. Zhu *et al.*, “Incorporating BERT into Neural Machine Translation,” Feb. 2020, [Online]. Available: http://arxiv.org/abs/2002.06823

[33] H. P. Luhn, “The Automatic Creation of Literature Abstracts,” *IBM J Res Dev*, vol. 2, no. 2, pp. 159–165, 1958, doi: 10.1147/rd.22.0159.

[34] H. P. Edmundson, “New Methods in Automatic Extracting,” *J. ACM*, vol. 16, no. 2, pp. 264–285, Apr. 1969, doi: 10.1145/321510.321519.

[35] R. Mihalcea and P. Tarau, “Textrank: Bringing order into text,” in *Proceedings of the 2004 conference on empirical methods in natural language processing*, 2004, pp. 404–411.

[36] S. Brin and L. Page, “The anatomy of a large-scale hypertextual web search engine,” *Computer networks and ISDN systems*, vol. 30, no. 1–7, pp. 107–117, 1998.

[37] D. Miller, “Leveraging BERT for Extractive Text Summarization on Lectures.” [Online]. Available: https://github.com/dmmiller612/lecture-summarizer.

[38] J. Zhang, Y. Zhao, M. Saleh, and P. J. Liu, “PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization,” Dec. 2019, [Online]. Available: http://arxiv.org/abs/1912.08777

[39] O. Kononova, T. He, H. Huo, A. Trewartha, E. A. Olivetti, and G. Ceder, “iScience Opportunities and challenges of text mining in materials research,” 2021, doi: 10.1016/j.isci.

[40] V. Tshitoyan *et al.*, “Unsupervised word embeddings capture latent knowledge from materials science literature,” *Nature*, vol. 571, no. 7763, pp. 95–98, Jul. 2019, doi: 10.1038/s41586-019-1335-8.

[41] P. Cerda, G. Varoquaux, and B. Kégl, “Similarity encoding for learning with dirty categorical variables,” Jun. 2018, [Online]. Available: http://arxiv.org/abs/1806.00979

[42] M. C. Swain and J. M. Cole, “ChemDataExtractor: A Toolkit for Automated Extraction of Chemical Information from the Scientific Literature,” *J Chem Inf Model*, vol. 56, no. 10, pp. 1894–1904, Oct. 2016, doi: 10.1021/acs.jcim.6b00207.

[43] C. J. Court and J. M. Cole, “Auto-generated aterials database of Curie and Neél temperatures via semisupervised relationship extraction,” *Sci Data*, vol. 5, Jun. 2018, doi: 10.1038/sdata.2018.111.

[44] J. M. Cole, “A Design-to-Device Pipeline for Data-Driven Materials Discovery,” *Acc Chem Res*, vol. 53, no. 3, pp. 599–610, Mar. 2020, doi: 10.1021/acs.accounts.9b00470.

[45] S. Huang and J. M. Cole, “A database of battery materials auto-generated using ChemDataExtractor,” *Sci Data*, vol. 7, no. 1, Dec. 2020, doi: 10.1038/s41597-020-00602-2.

[46] J. Zhao and J. M. Cole, “A database of refractive indices and dielectric constants auto-generated using ChemDataExtractor,” *Sci Data*, vol. 9, no. 1, Dec. 2022, doi: 10.1038/s41597-022-01295-5.

[47] J. Mavračić, C. J. Court, T. Isazawa, S. R. Elliott, and J. M. Cole, "ChemDataExtractor 2.0: Autopopulated Ontologies for Materials Science," *J Chem Inf Model*, vol. 61, no. 9, pp. 4280–4289, Sep. 2021, doi: 10.1021/acs.jcim.1c00446.

[48] E. Loper and S. Bird, "Nltk: The natural language toolkit," *arXiv preprint cs/0205028*, 2002.

[49] E. Zdravevski *et al.*, "Automation in systematic, scoping and rapid reviews by an NLP toolkit: a case study in enhanced living environments," *Enhanced living environments*, pp. 1–18, 2019.

[50] L. A. Kahale *et al.*, "PRISMA flow diagrams for living systematic reviews: a methodological survey and a proposal [version 1; peer review," 2021.

[51] M. Krenn and A. Zeilinger, "Predicting research trends with semantic and neural networks with an application in quantum physics", doi: 10.1073/pnas.1914370116/-/DCSupplemental.y.

[52] S. Rose, D. Engel, N. Cramer, and W. Cowley, "Automatic Keyword Extraction from Individual Documents," in *Text Mining: Applications and Theory*, John Wiley and Sons, 2010, pp. 1–20. doi: 10.1002/9780470689646.ch1.

[53] F. Kuniyoshi, J. Ozawa, and M. Miwa, "Analyzing Research Trends in Inorganic Materials Literature Using NLP," Jun. 2021, [Online]. Available: http://arxiv.org/abs/2106.14157

[54] D. Nadeau and S. Sekine, "A survey of named entity recognition and classification," *Lingvisticae Investigationes*, vol. 30, no. 1, pp. 3–26, 2007.

[55] L. Weston *et al.*, "Named Entity Recognition and Normalization Applied to Large-Scale Information Extraction from the Materials Science Literature," *J Chem Inf Model*, vol. 59, no. 9, pp. 3692–3702, Sep. 2019, doi: 10.1021/acs.jcim.9b00470.

[56] A. Trewartha *et al.*, "Quantifying the advantage of domain-specific pre-training on named entity recognition tasks in materials science," *Patterns*, vol. 3, no. 4, Apr. 2022, doi: 10.1016/j.patter.2022.100488.

[57] J. Lee *et al.*, "BioBERT: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020.

[58] I. Beltagy, K. Lo, and A. Cohan, "SciBERT: A pretrained language model for scientific text," *arXiv preprint arXiv:1903.10676*, 2019.

[59] K. Huang, J. Altosaar, and R. Ranganath, "Clinicalbert: Modeling clinical notes and predicting hospital readmission," *arXiv preprint arXiv:1904.05342*, 2019.

[60] J. Libovický, R. Rosa, and A. Fraser, "On the language neutrality of pre-trained multilingual representations," *arXiv preprint arXiv:2004.05160*, 2020.

[61] D. Araci, "Finbert: Financial sentiment analysis with pre-trained language models," *arXiv preprint arXiv:1908.10063*, 2019.

[62] T. Gupta, M. Zaki, N. M. A. Krishnan, and Mausam, "MatSciBERT: A materials domain language model for text mining and information extraction," *NPJ Comput Mater*, vol. 8, no. 1, Dec. 2022, doi: 10.1038/s41524-022-00784-w.

[63] V. Venugopal, S. Sahoo, M. Zaki, M. Agarwal, N. N. Gosvami, and N. M. A. Krishnan, "Looking through glass: Knowledge discovery from materials science literature using natural language processing," *Patterns*, vol. 2, no. 7, p. 100290, 2021.

[64] Z. Wang *et al.*, "Dataset of Solution-based Inorganic Materials Synthesis Recipes Extracted from the Scientific Literature," Nov. 2021, [Online]. Available: http://arxiv.org/abs/2111.10874

[65] O. Kononova *et al.*, "Text-mined dataset of inorganic materials synthesis recipes," *Sci Data*, vol. 6, no. 1, Dec. 2019, doi: 10.1038/s41597-019-0224-1.

[66] D. C. Elton *et al.*, "Using natural language processing techniques to extract information on the properties and functionalities of energetic materials from large text corpora," Mar. 2019, [Online]. Available: http://arxiv.org/abs/1903.00415

[67] A. C. Vaucher, F. Zipoli, J. Geluykens, V. H. Nair, P. Schwaller, and T. Laino, "Automated extraction of chemical synthesis actions from experimental procedures," *Nat Commun*, vol. 11, no. 1, Dec. 2020, doi: 10.1038/s41467-020-17266-6.

[68] K. Cruse *et al.*, "Text-mined dataset of gold nanoparticle synthesis procedures, morphologies, and size entities," Apr. 2022, [Online]. Available: http://arxiv.org/abs/2204.10379

[69] J. Reiss and J. Sprenger, "Scientific Objectivity," 2013. [Online]. Available: http://www.laeuferpaar.de.

[70] Y.-B. Zhou, L. Lü, and M. Li, "Quantifying the influence of scientists and their publications: distinguishing between prestige and popularity," *New J Phys*, vol. 14, no. 3, p. 033033, 2012.

[71] N. I. Zheludev, "The road ahead for metamaterials," *Science (1979)*, vol. 328, no. 5978, pp. 582–583, 2010.

[72] K. Z. Rajab, Y. Hao, D. Bao, C. G. Parini, J. Vazquez, and M. Philippakis, "Stability of active magnetoinductive metamaterials," *J Appl Phys*, vol. 108, no. 5, p. 054904, 2010.

[73] E. Jan and N. A. Kotov, “Successful differentiation of mouse neural stem cells on layer-by-layer assembled single-walled carbon nanotube composite,” *Nano Lett*, vol. 7, no. 5, pp. 1123–1128, 2007.

[74] S. Walia *et al.*, “Flexible metasurfaces and metamaterials: A review of materials and fabrication processes at micro-and nano-scales,” *Appl Phys Rev*, vol. 2, no. 1, p. 011303, 2015.

[75] Z. Wang, F. Cheng, T. Winsor, and Y. Liu, “Optical chiral metamaterials: a review of the fundamentals, fabrication methods and applications,” *Nanotechnology*, vol. 27, no. 41, p. 412001, 2016.

[76] “Content Coverage Guide.”

[77] W. L. Belcher, *Writing your journal article in twelve weeks: A guide to academic publishing success*. University of Chicago Press, 2019.

[78] M. F. Porter, “An algorithm for suffix stripping,” *Program*, 1980.

[79] R. Bellman, “On the theory of dynamic programming,” *Proceedings of the national Academy of Sciences*, vol. 38, no. 8, pp. 716–719, 1952.

[80] C.-W. Lee, H. J. Choi, and H. Jeong, “Tunable metasurfaces for visible and SWIR applications,” *Nano Converg*, vol. 7, no. 1, pp. 1–11, 2020.

[81] T. Mikolov, W. Yih, and G. Zweig, “Linguistic regularities in continuous space word representations,” in *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies*, 2013, pp. 746–751.

[82] T. Bolukbasi, K.-W. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai, “Man is to computer programmer as woman is to homemaker? debiasing word embeddings,” *Adv Neural Inf Process Syst*, vol. 29, 2016.

[83] A. E. Ezugwu, A. K. Shukla, M. B. Agbaje, O. N. Oyelade, A. José-García, and J. O. Agushaka, “Automatic clustering algorithms: a systematic review and bibliometric analysis of relevant literature,” *Neural Comput Appl*, vol. 33, pp. 6247–6306, 2021.

[84] A. Likas, N. Vlassis, and J. J. Verbeek, “The global k-means clustering algorithm,” *Pattern Recognit*, vol. 36, no. 2, pp. 451–461, 2003.

[85] A. M. Fahim, A. M. Salem, F. A. Torkey, and M. Ramadan, “An efficient enhanced k-means clustering algorithm,” *Journal of Zhejiang University-Science A*, vol. 7, pp. 1626–1633, 2006.

[86] C. Shi, B. Wei, S. Wei, W. Wang, H. Liu, and J. Liu, “A quantitative discriminant method of elbow point for the optimal number of clusters in clustering algorithm,” *EURASIP J Wirel Commun Netw*, vol. 2021, no. 1, pp. 1–16, 2021.

[87] R. Nainggolan, R. Perangin-angin, E. Simarmata, and A. F. Tarigan, “Improved the performance of the k-means cluster using the sum of squared error (SSE) optimized by using the Elbow method,” in *Journal of Physics: Conference Series*, IOP Publishing, 2019, p. 012015.

[88] L. Cohen, *Time-frequency analysis*, vol. 778. Prentice hall New Jersey, 1995.

[89] G. E. P. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, *Time series analysis: forecasting and control*. John Wiley & Sons, 2015.

[90] M. Roondiwala, H. Patel, and S. Varma, “Predicting stock prices using LSTM,” *International Journal of Science and Research (IJSR)*, vol. 6, no. 4, pp. 1754–1756, 2017.

[91] K. Park, Y. Choi, W. J. Choi, H.-Y. Ryu, and H. Kim, “LSTM-based battery remaining useful life prediction with multi-channel charging profiles,” *Ieee Access*, vol. 8, pp. 20786–20798, 2020.

[92] R. Adhikari and R. K. Agrawal, “An introductory study on time series modeling and forecasting,” *arXiv preprint arXiv:1302.6613*, 2013.

[93] G. P. Zhang, “Time series forecasting using a hybrid ARIMA and neural network model,” *Neurocomputing*, vol. 50, pp. 159–175, 2003.

[94] A. Bejan and S. Lorente, “The constructal law origin of the logistics S curve,” *J Appl Phys*, vol. 110, no. 2, p. 024901, 2011.

[95] M. Campani and R. Vaglio, “A simple interpretation of the growth of scientific/technological research impact leading to hype-type evolution curves,” *Scientometrics*, vol. 103, no. 1, pp. 75–83, 2015.

[96] W. J. Padilla, D. N. Basov, and D. R. Smith, “Negative refractive index metamaterials,” *Materials today*, vol. 9, no. 7–8, pp. 28–35, 2006.

[97] V. J. Einck *et al.*, “Scalable nanoimprint lithography process for manufacturing visible metasurfaces composed of high aspect ratio TiO_2 meta-atoms,” *ACS Photonics*, vol. 8, no. 8, pp. 2400–2409, 2021.

[98] R. C. Mitchell-Thomas, T. M. McManus, O. Quevedo-Teruel, S. A. R. Horsley, and Y. Hao, “Perfect surface wave cloaks,” *Phys Rev Lett*, vol. 111, no. 21, p. 213901, 2013.

[99] M. McCall, “Transformation optics and cloaking,” *Contemp Phys*, vol. 54, no. 6, pp. 273–286, 2013.

[100] S. Xiao, T. Wang, T. Liu, C. Zhou, X. Jiang, and J. Zhang, “Active metamaterials and metadevices: a review,” *J Phys D Appl Phys*, vol. 53, no. 50, p. 503002, 2020.

[101] H. Chu *et al.*, “Invisible surfaces enabled by the coalescence of anti-reflection and wavefront controllability in ultrathin metasurfaces,” *Nat Commun*, vol. 12, no. 1, p. 4523, 2021.

[102] B. Wu *et al.*, “Experimental demonstration of a transparent graphene millimetre wave absorber with 28% fractional bandwidth at 140 GHz,” *Sci Rep*, vol. 4, no. 1, pp. 1–7, 2014.

[103] A. Ihalage and Y. Hao, “Analogical discovery of disordered perovskite oxides by crystal structure information hidden in unsupervised material fingerprints,” *NPJ Comput Mater*, vol. 7, no. 1, p. 75, 2021.

[104] N. Scarpato, A. Pieroni, L. Di Nunzio, and F. Fallucchi, “E-health-IoT universe: A review,” *management*, vol. 21, no. 44, p. 46, 2017.

[105] D. S. W. Ting, L. Carin, V. Dzau, and T. Y. Wong, “Digital technology and COVID-19,” *Nat Med*, vol. 26, no. 4, pp. 459–461, 2020.

[106] R. W. Palmatier, M. B. Houston, and J. Hulland, “Review articles: Purpose, process, and structure,” *Journal of the Academy of Marketing Science*, vol. 46. Springer, pp. 1–5, 2018.

[107] R. Bos, S. De Waele, and P. M. T. Broersen, “Autoregressive spectral estimation by application of the Burg algorithm to irregularly sampled data,” *IEEE Trans Instrum Meas*, vol. 51, no. 6, pp. 1289–1294, Dec. 2002, doi: 10.1109/TIM.2002.808031.

[108] J.-F. Chen, W.-M. Wang, and C.-M. Huang, “Analysis of an adaptive time-series autoregressive moving-average (ARMA) model for short-term load forecasting,” *Electric Power Systems Research*, vol. 34, no. 3, pp. 187–196, 1995.

[109] E. S. Gardner Jr, “Exponential smoothing: The state of the art,” *J Forecast*, vol. 4, no. 1, pp. 1–28, 1985.

[110] S. Lee and D. B. Fambro, “Application of subset autoregressive integrated moving average model for short-term freeway traffic volume forecasting,” *Transp Res Rec*, vol. 1678, no. 1, pp. 179–188, 1999.

[111] M. Pirani, P. Thakkar, P. Jivrani, M. H. Bohara, and D. Garg, “A Comparative Analysis of ARIMA, GRU, LSTM and BiLSTM on Financial Time Series Forecasting,” in *2022 IEEE International Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE)*, 2022, pp. 1–6. doi: 10.1109/ICDCECE53908.2022.9793213.

[112] S. Siami-Namini, N. Tavakoli, and A. S. Namin, “A Comparison of ARIMA and LSTM in Forecasting Time Series,” in *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 2018, pp. 1394–1401. doi: 10.1109/ICMLA.2018.00227.

[113] B. Srinivasa-Desikan, *Natural Language Processing and Computational Linguistics: A practical guide to text analysis with Python, Gensim, spaCy, and Keras*. Packt Publishing Ltd, 2018.

[114] A. N. Khan, Y.-O. Cha, H. Giddens, and Y. Hao, “Recent Advances in Organ Specific Wireless Bioelectronic Devices: Perspective on Biotelemetry and Power Transfer Using Antenna Systems,” *Engineering*, vol. 11, pp. 27–41, 2022, doi: https://doi.org/10.1016/j.eng.2021.10.019.

[115] K. Sharma, A. Gaikwad, S. Patil, P. Kumar, and D. P. Salapurkar, “Automated Document Summarization and Classification Using Deep Learning,” *International Research Journal of Engineering and Technology*, vol. 5, no. 06, 2018.

[116] M. E. Peters *et al.*, “Deep contextualized word representations,” Feb. 2018, [Online]. Available: http://arxiv.org/abs/1802.05365

[117] C.-Y. Lin, “Rouge: A package for automatic evaluation of summaries,” in *Text summarization branches out*, 2004, pp. 74–81.

[118] P. Wilkinson, P. S. Cannon, and W. R. Stone, “100 Years of the International Union of Radio Science”.

[119] C. A. Balanis, “Antenna theory: A review,” *Proceedings of the IEEE*, vol. 80, no. 1, pp. 7–23, 1992.

[120] M. A. Jensen and J. W. Wallace, “A review of antennas and propagation for MIMO wireless communications,” *IEEE Trans Antennas Propag*, vol. 52, no. 11, pp. 2810–2824, 2004.

[121] M. A. Ullah, R. Keshavarz, M. Abolhasan, J. Lipman, K. P. Esselle, and N. Shariati, “A review on antenna technologies for ambient rf energy harvesting and wireless power transfer: Designs, challenges and applications,” *IEEE Access*, 2022.

[122] H. M. El Misilmani, T. Naous, and S. K. Al Khatib, “A review on the design and optimization of antennas using machine learning algorithms and techniques,” *International Journal of RF and Microwave Computer-Aided Engineering*, vol. 30, no. 10, p. e22356, 2020.

[123] M. O. Akinsolu, K. K. Mistry, B. Liu, P. I. Lazaridis, and P. Excell, “Machine learning-assisted antenna design optimization: A review and the state-of-the-art,” in

2020 14th European conference on antennas and propagation (EuCAP), IEEE, 2020, pp. 1–5.

[124] J. Anguera, A. Andújar, M.-C. Huynh, C. Orlenius, C. Picher, and C. Puente, "Advances in antenna technology for wireless handheld devices," *Int J Antennas Propag*, vol. 2013, 2013.

[125] H. Huang, "Flexible wireless antenna sensor: A review," *IEEE Sens J*, vol. 13, no. 10, pp. 3865–3872, 2013.

[126] K. N. Paracha, S. K. A. Rahim, P. J. Soh, and M. Khalily, "Wearable antennas: A review of materials, structures, and innovative features for autonomous communication and sensing," *IEEE Access*, vol. 7, pp. 56694–56712, 2019.

[127] Z. Wang, Z. Yang, and T. Dong, "A review of wearable technologies for elderly care that can accurately track indoor position, recognize physical activities and monitor vital signs in real time," *Sensors*, vol. 17, no. 2, p. 341, 2017.

[128] J. Hirschberg and C. D. Manning, "Advances in natural language processing," *Science (1979)*, vol. 349, no. 6245, pp. 261–266, 2015.

[129] E. Kim *et al.*, "Machine-learned and codified synthesis parameters of oxide materials," *Sci Data*, vol. 4, p. 170127, Sep. 2017, doi: 10.1038/sdata.2017.127.

[130] Y. Liu, T. Zhao, W. Ju, and S. Shi, "Materials discovery and design using machine learning," *Journal of Materiomics*, vol. 3, no. 3. Chinese Ceramic Society, pp. 159–177, Sep. 01, 2017. doi: 10.1016/j.jmat.2017.08.002.

[131] M.-T. Luong, H. Pham, and C. D. Manning, "Effective Approaches to Attention-based Neural Machine Translation," Aug. 2015, [Online]. Available: http://arxiv.org/abs/1508.04025

[132] S. Rose, D. Engel, N. Cramer, and W. Cowley, "Automatic keyword extraction from individual documents," *Text mining: applications and theory*, pp. 1–20, 2010.

[133] B. González-Pereira, V. P. Guerrero-Bote, and F. Moya-Anegón, "A new approach to the metric of journals' scientific prestige: The SJR indicator," *J Informetr*, vol. 4, no. 3, pp. 379–391, 2010.

[134] Z. Gyöngyi, P. Berkhin, H. Garcia-Molina, and J. O. Pedersen, "Link spam detection based on mass estimation," in *VLDB*, Citeseer, 2006, pp. 439–450.

[135] S. Brin and L. Page, "The anatomy of a large-scale hypertextual web search engine," *Computer networks and ISDN systems*, vol. 30, no. 1–7, pp. 107–117, 1998.

[136] E. Garfield, "The history and meaning of the journal impact factor," *JAMA*, vol. 295, no. 1, pp. 90–93, 2006.

[137] M. E. Falagas, V. D. Kouranos, R. Arencibia-Jorge, and D. E. Karageorgopoulos, "Comparison of SCImago journal rank indicator with journal impact factor," *The FASEB journal*, vol. 22, no. 8, pp. 2623–2628, 2008.

[138] R. Cagan, "San Francisco declaration on research assessment," *Disease models & mechanisms*. The Company of Biologists Ltd, p. dmm-012955, 2013.

[139] V. K. Singh, P. Singh, M. Karmakar, J. Leta, and P. Mayr, "The journal coverage of Web of Science, Scopus and Dimensions: A comparative analysis," *Scientometrics*, vol. 126, pp. 5113–5142, 2021.

[140] M. E. Rose and J. R. Kitchin, "pybliometrics: Scriptable bibliometrics using a Python interface to Scopus," *SoftwareX*, vol. 10, p. 100263, 2019.

[141] J. G. Andrews *et al.*, "What will 5G be?," *IEEE Journal on selected areas in communications*, vol. 32, no. 6, pp. 1065–1082, 2014.

[142] N. Yu *et al.*, "Light propagation with phase discontinuities: generalized laws of reflection and refraction," *Science (1979)*, vol. 334, no. 6054, pp. 333–337, 2011.

[143] J. N. Laneman, D. N. C. Tse, and G. W. Wornell, "Cooperative diversity in wireless networks: Efficient protocols and outage behavior," *IEEE Trans Inf Theory*, vol. 50, no. 12, pp. 3062–3080, 2004.

[144] E. G. Larsson, O. Edfors, F. Tufvesson, and T. L. Marzetta, "Massive MIMO for next generation wireless systems," *IEEE communications magazine*, vol. 52, no. 2, pp. 186–195, 2014.

[145] S. M. Alamouti, "A simple transmit diversity technique for wireless communications," *IEEE Journal on selected areas in communications*, vol. 16, no. 8, pp. 1451–1458, 1998.

[146] T. S. Rappaport *et al.*, "Millimeter wave mobile communications for 5G cellular: It will work!," *IEEE access*, vol. 1, pp. 335–349, 2013.

[147] F. Rusek *et al.*, "Scaling up MIMO: Opportunities and challenges with very large arrays," *IEEE Signal Process Mag*, vol. 30, no. 1, pp. 40–60, 2012.

[148] M. W. Berry and J. Kogan, *Text mining: applications and theory*. John Wiley & Sons, 2010.

[149] Y. Cha and Y. Hao, "The Dawn of Metamaterial Engineering Predicted via Hyperdimensional Keyword Pool and Memory Learning," *Adv Opt Mater*, vol. 10, no. 8, p. 2102444, 2022.

[150] S. Venugopalan, M. Rohrbach, J. Donahue, R. Mooney, T. Darrell, and K. Saenko, "Sequence to sequence-video to text," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 4534–4542.

[151] X. Xiao, L. Wang, K. Ding, S. Xiang, and C. Pan, "Deep Hierarchical Encoder-Decoder Network for Image Captioning," *IEEE Trans Multimedia*, vol. 21, no. 11, pp. 2942–2956, Nov. 2019, doi: 10.1109/TMM.2019.2915033.

[152] T. Wang, P. Chen, K. Amaral, and J. Qiang, "An Experimental Study of LSTM Encoder-Decoder Model for Text Simplification," Sep. 2016, [Online]. Available: http://arxiv.org/abs/1609.03663

[153] D. Bahdanau, K. Cho, and Y. Bengio, "Neural Machine Translation by Jointly Learning to Align and Translate," Sep. 2014, [Online]. Available: http://arxiv.org/abs/1409.0473

[154] M. Steinert and L. Leifer, *Scrutinizing Gartner's hype cycle approach*. 2010.

[155] G. Palikaras and E. Kallos, "The Gartner Hype Cycle for metamaterials," in *2014 8th International Congress on Advanced Electromagnetic Materials in Microwaves and Optics, METAMATERIALS 2014*, Institute of Electrical and Electronics Engineers Inc., Nov. 2014, pp. 397–399. doi: 10.1109/MetaMaterials.2014.6948573.

[156] W. H. Chin, Z. Fan, and R. Haines, "Emerging technologies and research challenges for 5G wireless networks," *IEEE Wirel Commun*, vol. 21, no. 2, pp. 106–112, 2014, doi: 10.1109/MWC.2014.6812298.

[157] W. A. Awan, A. Zaidi, N. Hussain, S. Khalid, and A. Baghdad, "Frequency Reconfigurable patch antenna for millimeter wave applications," in *2019 2nd International Conference on Computing, Mathematics and Engineering Technologies (iCoMET)*, IEEE, 2019, pp. 1–5.

[158] J. L. Valdes, L. Huitema, E. Arnaud, D. Passerieux, and A. Crunteanu, "A Polarization Reconfigurable Patch Antenna in the Millimeter-Waves Domain Using Optical Control of Phase Change Materials," *IEEE Open Journal of Antennas and Propagation*, vol. 1, pp. 224–232, 2020, doi: 10.1109/OJAP.2020.2996767.

[159] A.-K. U. Michel, P. Zalden, D. N. Chigrin, M. Wuttig, A. M. Lindenberg, and T. Taubner, "Reversible Optical Switching of Infrared Antenna Resonances with Ultrathin Phase-Change Layers Using Femtosecond Laser Pulses," *ACS Photonics*, vol. 1, no. 9, pp. 833–839, Sep. 2014, doi: 10.1021/ph500121d.

[160] M. Wuttig, H. Bhaskaran, and T. Taubner, "Phase-change materials for non-volatile photonic applications," *Nat Photonics*, vol. 11, no. 8, pp. 465–476, 2017.

[161] W. Zhang, R. Mazzarello, M. Wuttig, and E. Ma, “Designing crystallization in phase-change materials for universal memory and neuro-inspired computing,” *Nat Rev Mater*, vol. 4, no. 3, pp. 150–168, 2019.

[162] X. Yin *et al.*, “Beam switching and bifocal zoom lensing using active plasmonic metasurfaces,” *Light Sci Appl*, vol. 6, no. 7, pp. e17016–e17016, 2017.

[163] P. Xu, J. Zheng, J. K. Doylend, and A. Majumdar, “Low-loss and broadband nonvolatile phase-change directional coupler switches,” *ACS Photonics*, vol. 6, no. 2, pp. 553–557, 2019.

[164] X. Wang *et al.*, “Advances in photonic devices based on optical phase-change materials,” *Molecules*, vol. 26, no. 9, p. 2813, 2021.

[165] X. Li *et al.*, “Fast and reliable storage using a 5 bit, nonvolatile photonic memory cell,” *Optica*, vol. 6, no. 1, pp. 1–6, 2019.

[166] Q. Zhang, Y. Zhang, J. Li, R. Soref, T. Gu, and J. Hu, “Broadband nonvolatile photonic switching based on optical phase change materials: beyond the classical figure-of-merit,” *Opt Lett*, vol. 43, no. 1, pp. 94–97, 2018.

[167] Z. Sun, J. Zhou, and R. Ahuja, “Structure of phase change materials for data storage,” *Phys Rev Lett*, vol. 96, no. 5, p. 055507, 2006.

[168] E. Haddad *et al.*, “Review of the VO2 smart material applications with emphasis on its use for spacecraft thermal control,” *Front Mater*, vol. 9, 2022, doi: 10.3389/fmats.2022.1013848.

[169] M. K. Dietrich, F. Kuhl, A. Polity, and P. J. Klar, “Optimizing thermochromic VO2 by co-doping with W and Sr for smart window applications,” *Appl Phys Lett*, vol. 110, no. 14, p. 141907, 2017.

[170] A. Krammer, O. Bouvard, and A. Schüler, “Study of Si doped VO2 thin films for solar thermal applications,” *Energy Procedia*, vol. 122, pp. 745–750, 2017.

[171] J.-L. Victor, M. Gaudon, G. Salvatori, O. Toulemonde, N. Penin, and A. Rougier, “Doubling of the Phase Transition Temperature of VO2 by Fe Doping,” *J Phys Chem Lett*, vol. 12, no. 32, pp. 7792–7796, Aug. 2021, doi: 10.1021/acs.jpclett.1c02179.

[172] H. Ji, D. Liu, and H. Cheng, “Infrared optical modulation characteristics of W-doped VO2(M) nanoparticles in the MWIR and LWIR regions,” *Mater Sci Semicond Process*, vol. 119, p. 105141, 2020, doi: https://doi.org/10.1016/j.mssp.2020.105141.

[173] I. Robinson, J. Webber, and E. Eifrem, *Graph databases: new opportunities for connected data*. “ O’Reilly Media, Inc.,” 2015.

[174] L. L. B. Lazaro, R. A. Bellezoni, J. A. Puppim de Oliveira, P. R. Jacobi, and L. L. Giatti, “Ten years of research on the water-energy-food nexus: An analysis of topics evolution,” *Frontiers in Water*, vol. 4, p. 859891, 2022.

[175] K. T. Mukaddem, E. J. Beard, B. Yildirim, and J. M. Cole, “ImageDataExtractor: A Tool to Extract and Quantify Data from Microscopy Images,” *J Chem Inf Model*, vol. 60, no. 5, pp. 2492–2509, May 2020, doi: 10.1021/acs.jcim.9b00734.

Appendix A

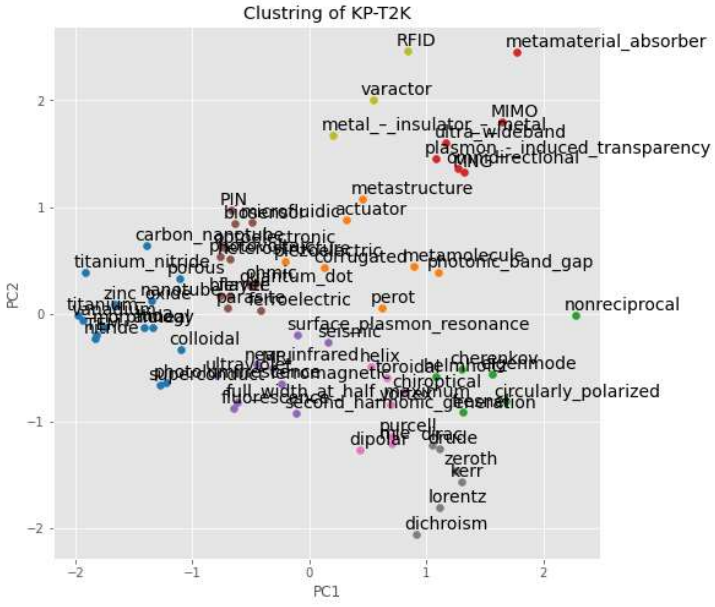
Metamaterial Research Keywords Data

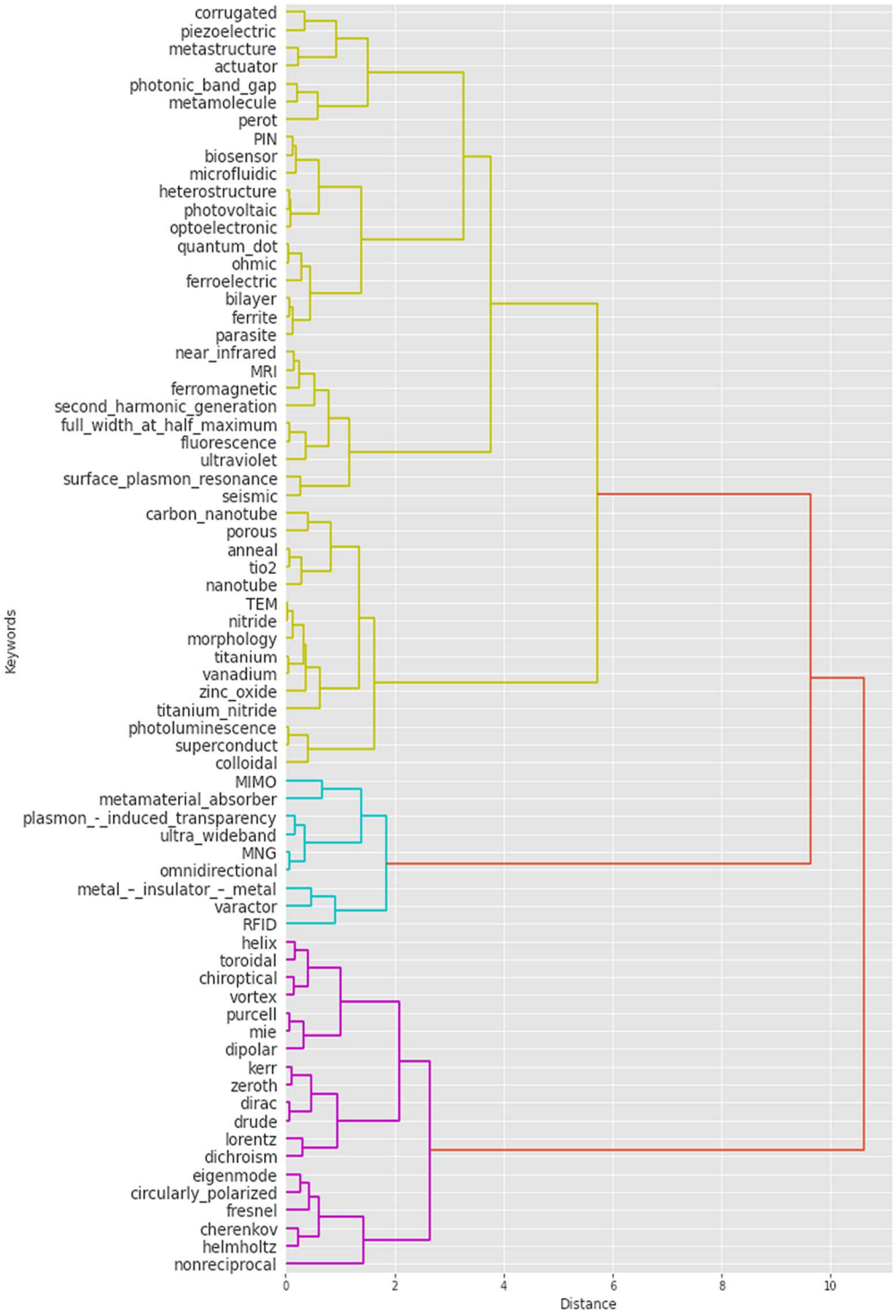
A.1 All Keywords by Frequency Scale.

Frequency Scale	Keywords
0.1–0.7	optic, antenna, magnet, electromagnetic, SRR, plasmonic, dielectric, propagation, transmissive, terahertz, refractive, electric, waveguide, polarization, permittivity, radiation, gigahertz, permeability, absorption, nonlinearity, absorber, transmission_line, bandwidth, wavelength, acoustic, cloaking, tunable, photonic, anisotropy, electron
0.01–0.09	photonic_crystal, lens, spatial, chiral, electromagnetic_wave, infrared, crystal, nanoparticle, metasurface, spectral, negative_refraction, nanostructure, cavity, thermal, plasma, isotropy, graphene, perfect, quantum, bandgap, finite_difference_time_domain, silicon, microplasma, elastic, transverse_electric, semiconductor, spectroscopy, oscillation, hyperbolic_metamaterial, atom, microstructure, genetic_algorithm, negative_refractive_index, radio_frequency, nanoantenna, negative_index_material, epsilon_near_zero, spheric, molecule, LC_circuit, electromagnetically_induced_transparency, phononic, fishnet, plasmonic_metamaterial, oxide, nanophotonic, Ag, artificial_magnetic_conductor, copper, Au, ferrite, bianisotropy, varactor, coding, colloidal, MIE, omnidirectional, eigenmode, piezoelectric, bilayer, microstrip_patch_antenna, dichroism, superconduct, corrugated, mu-negative, ferromagnetic, digital, Drude, Kerr, optoelectronic, actuator,

	parasite, toroidal, nanotube, heterostructure, ultra_wideband, dipolar, perfect_absorber, Lorent
0.0009–0.009	zeroth_order_resonator, specific_absorption_rate, circularly_polarized, morphology, mushroom, fluorescence, nonreciprocal, porous, MIMO, biosensor, metamaterial_absorber, RFID, leaky_wave_antenna, perfect_electric_conductor, ferroelectric, metamolecule, grid_antenna_array, acoustic_metamaterial, second_harmonic_generation, Cherenkov, Helmholtz, UV, photovoltaic, Dirac, photonic_band_gap, phased_array, nitride, PIN-diode, microfluidic, plasmon_induced_transparency, Fresnel, zinc_oxide, helix, seismic, zero_index_metamaterial, TEM, vortex, annealing, ohmic, surface_plasmon_resonance, MRI, titanium_nitride, metastructure, substrate_integrated_waveguide, Purcell, full_width_at_half_maximum, metal_insulator_metal, quantum_dot, near_infrared, vanadium, chiroptical, photoluminescence, parity_time, carbon_nanotubes, topological_insulator, perfect_metamaterial_absorber, negative_stiffness, titanium, vanadium_oxide, refractive_index_unit , TiO_2, Fabry_Perot

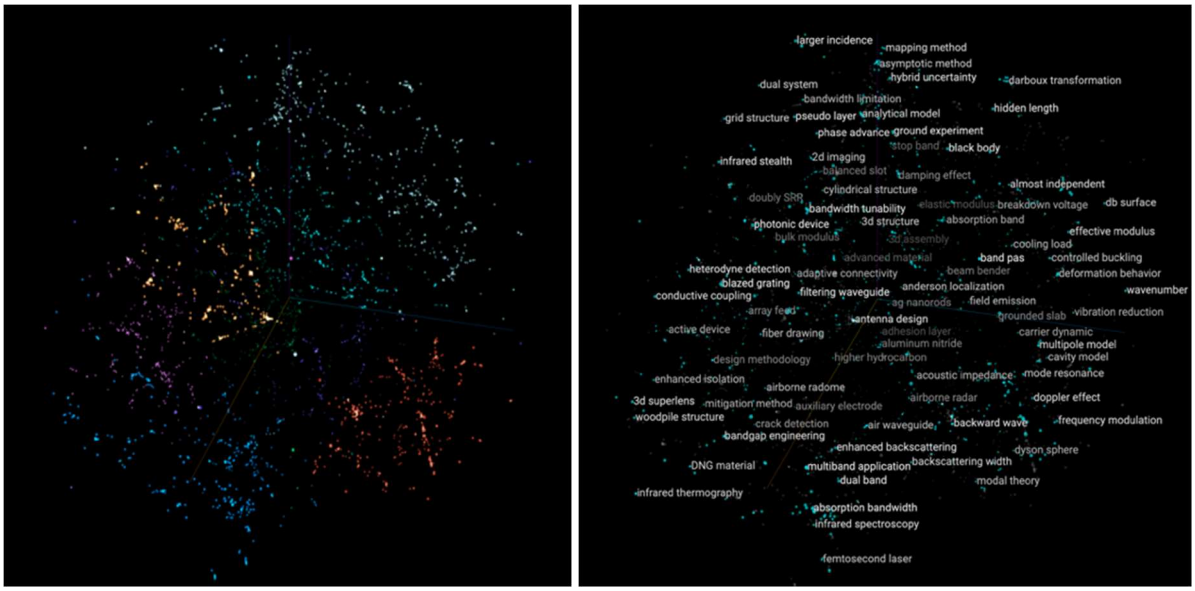
A.2 The Result of Clustering and Dendrogram of KP-T2K.




corrugated
piezoelectric
metastructure
actuator
photonic_band_gap
metamolecule
perot
PIN
biosensor
microfluidic
heterostructure
photovoltaic
optoelectronic
quantum_dot
ohmic
ferroelectric
bilayer
ferrite
parasite
near_infrared
MRI
ferromagnetic
second_harmonic_generation
full_width_at_half_maximum
fluorescence
ultraviolet
surface_plasmon_resonance
seismic
carbon_nanotube
porous
anneal
tio2
nanotube
TEM
nitride
morphology
titanium
vanadium
zinc_oxide
titanium_nitride
photoluminescence
superconduct
colloidal
MIMO
metamaterial_absorber
plasmon_-_induced_transparency
ultra_wideband
MNG
omnidirectional
metal_-_insulator_-_metal
varactor
RFID
helix
toroidal
chiroptical
vortex
purcell
mie
dipolar
kerr
zeroth
dirac
drude
lorentz
dichroism
eigenmode
circularly_polarized
fresnel
cherenkov
helmholtz
nonreciprocal
Keywords
0
2
4
6
8
10
Distance

A.3 Visualising the Embedding of 3k Keyword Pool via TensorBoard Embedding Projector.

(a)



(b)

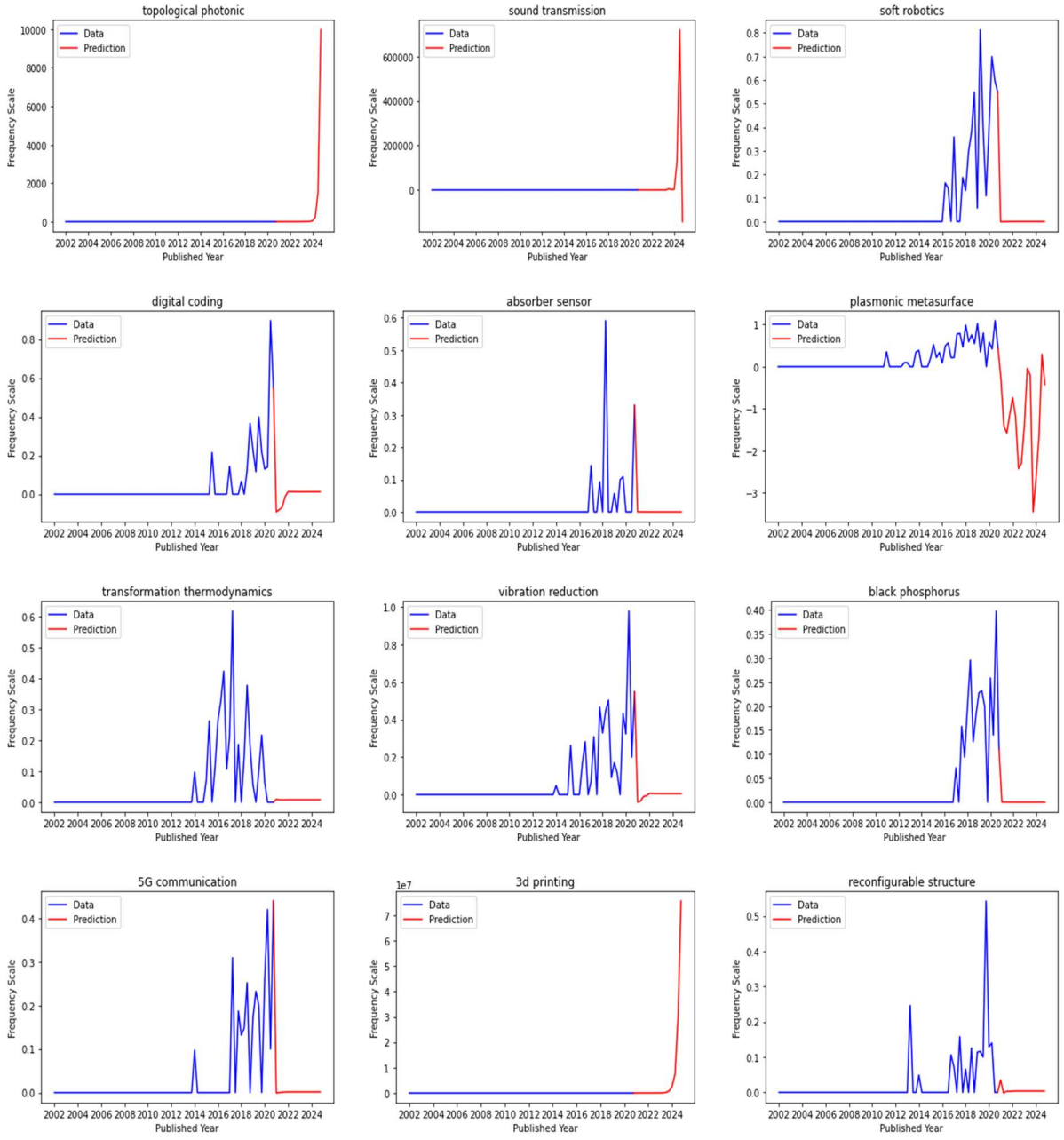
Nearest points in the original space:

frequency coding	0.236
complex coding	0.245
coding metasurface	0.337
digital modulation	0.364
digital camouflage	0.422
digital alloy	0.476
homomorphic encryption	0.478
digital twin	0.489
programmable hologram	0.501
keypoints detection	0.527
optical encryption	0.533
supervised learning	0.546
genetic algorithm	0.569
5G communication	0.579
spatial beamforming	0.582
wireless communication	0.590
interferometric autocorrelation	0.591
recursive algorithm	0.601
MIMO system	0.603
machine learning	0.604

A.3(a) shows the visualisation of 3,187 keywords embedding onto 3-dimensional spaces using PCA via TensorBoard Embedding Projector. All keywords are presented in the vector space. The nearest keywords of ‘digital coding’ and the related words to the chosen keyword ‘digital coding’ are visualised in A.3(b). Nearer to the chosen word would mean the higher

probability of co-occurrence. The keywords, ‘frequency coding’ and ‘digital coding’ are closely located meaning that their co-occurrence is high.

A.4 Examples of Prediction Outcomes from Emerging Keywords of Metamaterial Research.



Appendix B

Keywords Extraction & NLP-Summarisation Outcomes on Body Sensing Technologies

B.1 The Most Frequent Keyword by RAKE from 628 Publications of Flexible Electronic Sensors.

2003	previous projects like	modified navy f	incorporates connections wires	hands free communication	first prototype version
2004	voice recognition mainly	uneven contours resulting	specific diagnostic questions	skeletal muscles accessible	silicon wafer carrier
2005	standard spice simulator	standard goniometer used	sensitivity template matched	preliminary experiments involving	postoperative pacing thresholds
2006	voice recognition mainly	sheet braille displays	pressure recognition compared	potentially ultra-low	plastic actuator arrays
2007	success devices based	selective medium chip	polymer semiconductors allow	physiological parameters monitoring	large area feature
2008	wind resistant ability	steady state accuracy	signal processing circuits	self-organized network	protection safety emergency
2009	two major parts	thin film metals	robotics biomedicine aerodynamics	review recent progress	random loading conditions
2010	using different microstructures	standard manufacturing processes	semiconductor nanowires	RHID tag module	predict heat stroke

2011	tiny force exerted	system level solution	sensitivity ali javey	relevant physical issues	present benchmarking results
2012	ultraviolet ozone uv	transparent conducting electrode	transepithelial osmotic gradient	reversible directional manner	oxygen plasma produces
2013	versatile component including	tunnelling effect theory	thin film transistor	taking full advantage	reconstruction algorithm running
2014	voice vibrations highlighting	vibration the proposed theoretical model	tree planes conclusion	reference values achieving	presented whole package
2015	validated experimentally unlike	vacuum assisted infusion	thus integrate intimately	spontaneously grown upon	spontaneous buckle formation
2016	wheat stone bridge configuration	serially stacked springs	relatively longer pinfin	problematic issues faced	probable operational situations
2017	wind turbine blades	subtle blood pulses	stainless steel threads	polyvinylidene fluoride pvdf	polystyrene PS microspheres
2018	œ phase content	roadmap becomes clearer	outlet passive pumps	newly independent scientists	muscles upon swallowing
2019	waving badminton racket	tremendous efforts dedicated	tpu fibrous mats	task ideally suited	sub-zero temperatures resulting
2020	uv äìvis wavelength	true äúskin äù	triboelectric nanogenerator specially	triboelectric nanogenerator contacting	tf teg allowing

B.2 Summary Results using NLP Algorithm.

Section 1. Gastrointestinal tract monitoring. (5%)

Wireless Capsule Endoscopy (WCE) is a patient-friendly approach for digestive tract monitoring to support medical experts towards identifying any anomaly inside human's Gastrointestinal (GI) tract. The navigation techniques suggested for Wireless Capsule Endoscopy are image-based that are required to transfer and process a significant amount of data in real-Time operation. A novel navigation system for Wireless Capsule Endoscopy/ordinary endoscopy that does not depend on any external source for operation and can handle the uncertainties of the path even in a dark or liquid environment (*i.e.* mucosa) of the human body is presented in this paper. Working algorithms for flexible and rigid environments are described in this paper. Ingestible electronic systems that are capable of embedded sensing, particularly within the gastrointestinal (GI) tract and its accessory organs, have the potential to screen for diseases that are difficult if not impossible to detect at an early stage using other means. Electronic drug delivery systems such as capsules, on the other side, can be used not only to deliver drugs to a specific site in the gastrointestinal tract but can also record data and report the state of patients' gastrointestinal tract, and after excretion, this information can be studied and used to present them graphically. Because of the weak correlation that exists between symptoms and endoscopic disease activity, the treat-to-target paradigm has been developed, and the associated treatment goal is to achieve and maintain deep remission, encompassing both clinical and endoscopic remission. This review also demonstrated that small bowel capsule endoscopy (SBCE) can detect post-operative recurrence to a similar extent as ileocolonoscopic, and proximal SB lesions that are beyond the reach of the colonoscope in over half of the patients. The dominant T wave cluster and one-class SVM based analysis of multilead ECG for classification of myocardial infarction, and dysmenorrhea on blood pressure and radial pulse spectrum in women. This paper thus addresses the design requirements for an implant to treat GI dysmotility and presents a miniaturized wireless implant capable of modulating and recording GI motility. Everybody has experienced an embarrassing stomach growl during a lull in conversation or the satisfying and uncomfortable fullness after dessert. Gaining even more information about what is happening in the stomach and gastrointestinal (GI) tract has been an important clinical goal. Although previous biocompatible power-harvesting systems for in vivo use have demonstrated short (minute-long) bursts of power from the stomach, little is known about the potential for powering electronics in the longer term and throughout the gastrointestinal tract. Body Area Networks (BANs) has great potential to provide real-time health monitoring of a patient and diagnose many life-threatening diseases. The wireless capsule endoscopy (WCE) is one of the promising Body Area Networks (BANs) applications that provides a non-invasive way to inspect the entire Gastrointestinal (GI) tract. Combining ASIC and multiple microsensors low-power wireless electronic capsule was developed for the long-term monitoring of the entire human gastro-

intestinal (GI) tract. This paper surveys the different propagation models for implant communication that have been presented in the literature for narrowband and ultra-wideband signals. We developed an ingestible electronic drug delivery and monitoring system. This electronic drug delivery and monitoring system may be a promising tool for targeted delivery of substances to well-defined areas of the GI tract. The results of this study showed that the localization root mean square error (RMSE) of our Bayesian-based method when a sensor node was covered by four anchors was 1.0 mm which is smaller than that of other existing localization approaches under the same conditions such as classical MDS (43.1 mm), dwMDS (24.7 mm), MLE (21.8 mm) and POCS (1.7 mm). The WAC used in this study provided a non-invasive technique that produced novel information about the pony gastrointestinal tract, but owing to the substantial variability in GET values and long transit time it may not be a reliable clinical tool at this time. Tolerance is good with sometimes some transient chest pain. 48-h pH data and endoscopic findings were recorded. P=0.94) in patients on PPI therapy for each point increase on the GerdQ. The odds of an abnormal SAP 95% in patients studied off PPI therapy was 1.18 (95% CI, 1.09-1.28; Seven healthy, active participants (3 men, 4 women; Participants completed a 45-minute exercise trial at approximately 70% Vo(2peak). The transmission power levels needed to establish a reliable connection from the different gastrointestinal districts are reported and compared with safety levels from international guidelines. The increasing risk of small bowel carcinoma and prevention of obstruction and intussusception have been making frequent and acute surgical interventions in avoidably led to the necessity of screening and surveillance the patients. Gastric transit time and the rate of complete small bowel examination were compared. Serial radiographs were performed weekly until capsule release. Design/methodology/approach - The circuit requirements and methods of data transfer are examined.

Section 2. Retinal prosthesis. (3%)

We sought to describe the surgical techniques required in the ab-interno method to implant subretinal prostheses in mini-pigs and suggest tips to facilitate optimal outcomes. In comparison to implants which utilize inductively coupled coils, laser power delivery enables a high degree of miniaturization and lower surgical complexity. 175 implants were tested for up to 33 months. Here we show that within a visual angle of 46.3 degrees, POLYRETINA embeds 2215 stimulating pixels, of which 967 are in the central area of 5 mm, it is foldable to allow implantation through a small scleral incision, and it has a hemispherical shape to match the curvature of the eye. This paper thus addresses the design requirements for an implant to treat GI dysmotility and presents a miniaturized wireless implant capable of modulating and recording GI motility. In this paper, features of STS-based retinal prosthesis will be described in detail especially from an electrical circuit perspective. Retinal prostheses have the potential to restore some level of visual function to the patients suffering from retinal degeneration. Improved mobility and object detection are some of the more notable findings from the clinical trials. In-vitro experiment conducted in artificial vitreous humour is designed and set-up to investigate stimulation waveforms for better visual resolution. This simple architecture provides a hybrid design that involves a high-efficiency antenna which is used to send the signal wirelessly by means of MedRadio band. Finally, we show that the coil can be deformed into spherical shape to fit curvature of the eye for retinal prosthesis application without degradation of electrical properties and link performance. The complete package is demonstrated with a mechanical model with a parylene-C flexible circuit board, *i.e*. parylene flex, to show the placement of the IC chips, discrete components, and coils. The turning of the eye model is controlled with a microcontroller board. Measurement results show an enormous reduction of the transmitted power for large eye movements and therewith the need for power control in biomedical implants. Delivering power to an implanted device located deep inside the body is not trivial. Threshold measurements showed safe and stable current levels. Our laboratory, as described in Ray et al, is investigating the interface between stimulating microelectrodes and the retina, to inform the design of a high-resolution retinal prosthesis. Instead, the image data are represented by the timing of pulses or pulse edges. The required stimulation thresholds were found to be very low. This array is implanted in the subretinal space using a specially designed ab externo surgical technique that uses the retina to hold the array in place while leaving the bulk of the prosthesis outside the eye. Operation of the retinal implant has been verified in vivo in two pigs for up to five and a half months by measuring stimulus artifact on the eye surface using a contact lens electrode. Also, the wireless module is expected to operate in the reactive near-field region due to small separation between the transmit and receive antennas compared to their size and corresponding operating wavelength. The prosthesis conforms to the eye and drives a microfabricated polyimide stimulating electrode array with sputtered iridium oxide electrodes. The implanted device includes a hermetic titanium case containing a 15-channel stimulator chip and discrete power supply components. The goal of this design is to achieve high programmability with limited wireless transmission bandwidth. This digital controller can be used in other implantable multi-channel stimulation systems. Wireless operation of the retinal implant has been verified both in vitro and in vivo in three pigs for more than seven months, the latter by measuring stimulus artifacts on the eye surface using contact lens electrodes. Maintaining close proximity between the electrode array and the retinal surface is critical in developing a successful retinal implant. A temperature increases of 3.2degrees C in the living rabbit eye is to be expected when powering a subretinal implant with 15 mW (4.8 mW/mm2) IR power, the wattage used in an external power supply for an active implant with 1,500 electrodes. The stimulator was connected to the array by a multiwire cable and was controlled by a computer based external system that allowed precise control over each electrode.

Section 3. Cochlear implants for auditory nerve stimulation (3%)

Especially, medical implant devices such as a cochlear implant are one of the typical applications of the WPT technology. Purpose To evaluate the effect of the digital remote wireless microphone system, RogerTM, on speech recognition at different levels of multisource noise in SSD CI recipients using MED-EL CI sound processor OPUS 2. Consequently, such infections can lead to explantation and, in severe cases, amputation or even death. Pediatric listeners were tested in quiet and

in level 1 noise in A-only and AV environments. Kim questionnaire results showed statistically significant differences (P <.001) in the subjective satisfaction of the Bluetooth-implemented CI compared to the conventional mode for sound quality, noise interference, and sound accuracy. The aim of the study was to determine if contralateral routing of signal (CROS) technology results in improved hearing outcomes in unilateral cochlear implant (CI) patients and provides similar gains in speech perception in noise to traditional monaural listeners (MLs). The speech recognition was significantly poorer for children with CI than children with NH in quiet and in noise when using the TM alone. This structure encounter drawbacks, like maintenance of cable, expensive as any small damage requires huge money to be spent on cable. Finally, we discuss and anticipate future developments that will enhance the capabilities of current-day wirelessly powered implants and make them more efficient and integrable with other electronic components in IMDs. A three-by-two-way repeated measures design was used to evaluate mobile telephone sentence recognition performance differences obtained in quiet and in noise with and without the wireless HAT accessory coupled to the hearing aid alone, CI sound processor alone, and in the bimodal condition. Eleven Advanced Bionics (AB) cochlear implant recipients, ages 11 to 68 yr Data Collection and Analysis: In this paper we design PMPH that is able to harvest environmental vibration sounds and convert it to usable electrical power for artificial cochlea. This project deals with the design of antenna for teeth implantable device which performs the measurement and monitor of temperature by sensing the basal body temperature (BBT). robustness analysis of 10-dimensional cell cycle systems based on periodic sensitivity; proof-of-concept microwave sensor on flexible substrate for real-time water composition analysis; Modern implantable microelectronic devices (IMDs) require higher performance and power efficiency to enable more efficacious therapies, particularly in neuro-prostheses such as retinal and cochlear implants [1]. And the whole CDR circuit could recovery correct clock and data within a broad duty cycle range and consumes only 29.52ŒºW. The main focus of this review is to provide a holistic amalgamated overview of the most recent human in vivo techniques for implementing brain-computer interfaces (BCIs), bidirectional interfaces, and neuroprosthetics. a study on the relation between stability of EEG and respiration; a confidence measure for real-time eye movement detection in video-oculography; development of the tongue diagnosis system by using surface coating mirror; using saliency features for graphcut segmentation of perfusion kidney images; multi scale assessment of bone architecture and quality from CT images; an empirical approach for objective pain measurement using dermal and cardiac parameters; The imager is intended for use in a brain-machine visual prosthesis for the blind where energy efficiency and power are of paramount importance. In the circuit design for CIs, power is always a key issue which incorporates efficient power supply and low power consumption. Multiple sites can be driven in parallel to provide higher current levels. Backing structures and articulated insertion tools are being developed for dynamic closed-loop insertion control. This paper discusses some of the advances and limitations of the field of functional electrical stimulation (FES) and why the field has difficulty in achieving its promise whenever many wires are used throughout the body. Researchers are exploring pressure sensors that run at nanowatt levels and are small enough to fit inside the eye. To address these problems that hearing-impaired people experience with telephones, this paper proposes a wireless phone adapter that can be used to route the audio signal directly to the hearing aid or cochlear implant processor. This adapter is based on Bluetooth technology. The favourable features of this new wireless technology make the adapter superior to traditional assistive listening devices. A new telemetry system has been developed in order to fit with the common characteristics of future high-performance radio-powered implantable stimulators which are requiring as high as possible data bit rate to control higher number of electrodes and/or accurate definition in the stimulus waveforms with enough power transfer. Because of this high data bit rate and to maintain a low bit error rate (BER), the chip works properly with an inductive coupling link quite similar to current cochlear implants.

Section 4. Hyperthermia treatment. (8%)

In-stent restenosis concerning the coronary artery refers to the blood clotting-caused re-narrowing of the blocked section of the artery, which is opened using a stent. The problems and challenges associated with current stent monitoring technology were illustrated, along with its typical applications. Lastly, the challenges and concerns associated with nurturing a healthcare system were deliberated with meaningful evaluations. This paper presents a low-voltage low-power implantable telemonitoring system in the context of a smart stent that uses wireless endo-hyperthermia for the treatment of in-stent restenosis. The device can detect and response to the temperature variations in the range of 30 to 50 °C. The remote power link is established when the power received by the implantable device is about -8 dBm. Many of these applications can benefit from replacing external temperature probes with injectable wireless devices. This article presents an electromagnetically powered stent designed for hyperthermia treatment of in-stent restenosis. The study will pave the path to further design optimization and performance improvement for resonant stent technology towards its application to wireless thermal treatment of in-stent restenosis. We have developed a non-invasive wireless temperature measurement method by utilizing the magnetic permeability property of a ferromagnetic implant with low Curie temperature (FILCT) that varies with the temperature. The experimental results agree well with the simulation. It was very difficult, however, to detect the FILCT temperature if the applied magnetic field was in an unstable state. Also, their interaction with the human body should be well considered. The fabricated stent device with the initial diameter of 2 mm is expanded up to 6 mm in diameter to simultaneously deploy the resonant heater circuit and the mechanical scaffolding structure inside an artificial artery using a balloon catheter. Wireless heating tests of the prototypes deployed into artificial artery using commercial balloon catheters demonstrate the designed function of the circuit breaker, regulating the stent temperature within 50 °C - 66 °C when excited in air at an output RF power of 320 mW, which heats the device to 78 °C without the breaker. However, it cannot determine the temperature of the material during the power-off period. Cavity method is used to measure the bio-material parameters for design. An in vitro experiment has been carried out to verify the radiation efficiency of antenna and function properness of the whole system. To improve the accuracy of low-invasive temperature measurement, we formulated a method that reduces the drift in the pickup coil voltage caused by instability from the applied magnetic field. and low-cost, thermistor-based respiration monitor. To increase this low figure, we have developed microwave technology aiming to differentiate hemorrhagic from ischemic stroke patients. Wireless communication has played a significant role in modern healthcare

systems. Limitations in existing in vivo thermal-aware routing algorithms motivated us to use the in vivo "lightweight rendezvous routing" approach. Study design Prospective, randomized, crossover study. Following recovery from anesthesia, temperature in treatments IH and IHK was different from baseline ($p < 0.002$). When considering the efficiency of electromagnetic (EM) propagation inside the human body for BAN and hyperthermia treatment using RF, it is important to determine the mechanism of EM dissipation in the human body. In case of deep-seated brain tumours or cancerous cells, focused ultrasound is found to be an effective treatment tool. A Wireless Sensor Network (WSN) is developed to transmit the tumor thermal information to an expert clinician to monitor the health condition and for therapeutic control of ultrasound dosage level, particularly in case of elderly patients living in remote areas. Promising results from both methods are obtained.

Section 5. Cardiovascular healthcare (3%)

In recent years, flexible and sensitive pressure sensors are of extensive interest in healthcare monitoring, artificial intelligence, and national security. Cost efficiency comes from efficient signal processing and replacing manual analysis with AI based machine classification. IoT (Internet of Things) plays a vital role in modern medical field. Moreover, an effective system called – EmoStrokeSys - has been proposed for health monitoring that combines three wearable sensors: Interestingly, we found that perceived usefulness had an indirect effect on behavioural intention through attitude. Though the number of aged people is increasing, it is irrefutable that the need of a disperse medical care system providing remote monitoring intending to reduce the escalating healthcare expenditure is very urgent. The test results showed that this system could measure the patient's physiological data with high accuracy. Such devices can be deployed in any public spaces provide detailed information about the behaviour of individuals such as personalization, behaviour change and personal health monitoring. IoT technology which is being deployed is specially designed to make it invisible, such that the technology does not manifest its presence to the users it is monitoring. The ability to monitor subtle changes in vital and arterial signals using flexible devices attached to the human skin can be valuable for the detection of various health conditions such as cardiovascular disease. In this paper polynomial-based curve is generated and steganography technique has been used for secure health monitoring which provides data confidentiality and authentication to maintain the privacy of a patient. sleep, neurology, movement disorders and mental health; The objective was to design the Cardio-Sensor Aggregation Platform (CSAP) in order to capture the data (ECG, Pulse Rate and Heart Sound) in real-time and transmit the same from the CSAP to any mobile device running the application, by using wireless communication technology. In addition, the results provide insights into possible self-efficacy failings in traditional training and the benefits of embedding self-efficacy theory into the technology design process. A specific application of telemedicine for hypertension management is blood pressure telemonitoring (BPT), which allows remote data transmission of BP and additional information on patients' health status from their living site or from a community setting to the doctor's office or the hospital. Internet of Things (IoT)-devices are now expanding inter-connecting networking technologies to invent healthcare monitoring system especially for assessing physiological conditions of the chronically ill patients those with cardiovascular diseases. This study aims to develop a community-based electrocardiogram (ECG) monitoring system for cardiac outpatients to wirelessly detect heart rate, provide personalized healthcare, and enhance interactive social contact because of the prevalence of deaths from cardiovascular disease and the growing problem of aging in the world. This work gives a detailed insight into a novel wireless body sensor network and addresses critical aspects such as signal quality, synchronicity among multiple devices as well as the system's overall capabilities and limitations in cardiovascular monitoring. However, few fully automatic myocardial infarctions (MI) disease detection algorithms have well been developed. However, it is very expensive to have a professional expert assisting the patient throughout the treatment. The result obtained from DFA is used to display the patient's health condition on a smartphone anytime and anywhere. Advanced wireless technology, high speed internet facility and availability of other communication systems can be used to provide the accessibility of state-of-the-art healthcare facilities to the patients in remote and rural areas for monitoring and diagnosis of cardiovascular diseases, one of the prime causes of human mortality today. Wanda-CVD is a smartphone-based RHM system designed to assist participants to reduce identified CVD risk factors by motivating participants through wireless coaching using feedback and prompts as social support. In this study set in the Indian healthcare environment, an auto-triggered, wireless patch-type ELR was used with 125 patients (62.5 ± 16.7 years, 76 males) presenting a broad range of symptoms. We developed an adaptive home-based platform that integrates multifarious sensors and telecommunication services to cover the needs of both users and healthcare professionals. For home use, sensors and measuring apparatus are embedded in a chair, which is named as Sensing Chair. The possible obtained signs are electrocardiogram, body weight and pulse wave etc. The complete system also includes a wireless gateway for signal collection/transmission and a server that integrates database, web interface and decision support system. The systems are intended for training of personnel in the Health Care industry. However, if the value of remote monitoring is demonstrated to extend beyond the previous boundaries of in-person interrogations, a rational request can be made to reconsider the relative value of remote monitoring. This proposed wireless type biosignal alerting system aims at designing and developing a module that detects the abnormal interpretations in the PQRST complex (electrocardiography) and heart rate of a patient in advance, gives a self-warning ring to the patient, and also sends a short message service warning to the doctor's mobile phone through the Global System for Mobile Communication. Cardiovascular disease (CVD) is the leading cause of mortality in Australia, and places large burdens on the healthcare system. Both HR and RR are the most important vital signs during exercise but only used one physiological signal recorder in this system. Ubiquitous sensor network is technology for the domain of unobtrusive medical monitoring system in healthcare applications. The uCare device has been tested in a trial in Beijing Hospital. The system was developed in conjunction with the UCLA School of Nursing and the UCLA Wireless Health Institute to enable early detection of key clinical symptoms indicative of CHF-related decompensation in a real-time automated fashion and allows health professionals to offer surveillance, advice, and continuity of care and triggers early implementation of strategies to enhance adherence behaviours. Each sensor module is

combined with a tri-axis accelerometer for patient's posture and activity measurement. The platform hierarchy comprises three layers for sensing, communication, and management.

Section 6. Drug delivery devices (1%)

In the future, overcoming the problems will make TENG as an alternative power source for the biomedical and healthcare applications. Inspired by the drop-counting principle implemented in a clinical gravity drip, we propose a novel microfluidic flowmetry technology for polydimethylsiloxane (PDMS)-based conventional microfluidic devices, known as a microfluidic digital meter-on-chip (DMC), to achieve on-chip and localized microflow measurements with ultrahigh precision and a wide tunable range. The microrobot's function is validated by an in vitro experiment that mimics the anatomy of the stomach filled with fluid at a centimeter range underneath the skin. In this work, a smartphone controlled interactive theranostic device has been developed to perform in vitro photodynamic therapy (PDT) and diagnostic assays for treatment assessment on a single platform. Implantation of biodegradable wafers near the brain surgery site to deliver anti-cancer agents which target residual tumor cells by bypassing the blood-brain barrier has been a promising method for brain tumor treatment. This paper presents modeling and finite element analysis of a thermopneumatic micropump with a novel design that does not affect the temperature of the working fluid. The micropump is operated by activating a passive wireless heater using wireless power transfer when the magnetic field is tuned to match the resonant frequency of the heater. The proposed unrestrained motions are demonstrated through feasibility test. Miniaturized versions of electronically actuated (lead-screw and pulley) mechanisms are used for the specific purpose of controlled drug delivery. While webcams, camera phones, and iPads have been explored as potential new methods of real-time information sharing, the non-“hands-free” nature and lack of viewer and observer point-of-view render them unsuitable for the R&D laboratory or manufacturing setting. Herein, this study presents an advanced multifunctional dressing (GelDerm) capable of colorimetric measurement of pH, an indicator of bacterial infection, and release of antibiotic agents at the wound site. The ability to accurately measure the range between a sensor implanted in the human body and an external receiver can make a number of new medical applications such as better wireless capsule endoscopy, next generation microrobotic surgery systems, and targeted drug delivery systems possible. This paper will cover the basic physics and modeling of APT and will review the current state of acoustic (or ultrasonic) power transfer for biomedical implants. Our research focuses on the reliability of one particular wireless nanosensor network (WNSN) that is used for monitoring human lung cells. After applying 35% carbamide peroxide to human teeth topically or with the IDE at 1200Hz, 5 Vpp for 20min, spectrophotometric analysis showed that compared to diffusion, the IDE enhanced whitening in specular optic and specular optic excluded modes by 215% and 194% respectively. The force of the coil acting on the magnetic piston and the drug release profile were modeled and assessed on bench-top with a maximum relative error below 5%. We present ultrathin, soft microfluidic neural probes with wireless drug delivery capability that can be injected precisely in the deep brain tissue. In this work, we present the prototype design and the results of bench trials that demonstrated the device ability to trigger the drug deployment by characterizing the magnetic field and resulting force. However, the inevitable mismatch between the nominal and actual force-current maps along with external disturbances affects the positioning accuracy of the motion control system. The control system is implemented on a magnetic system for controlling microparticles of paramagnetic material, which experience magnetic forces that are related to the gradient of the field squared. In this work, the chip extracted a minimum of 1.12 nW from the EP of a guinea pig for up to 5 h, enabling a 2.4 GHz radio to transmit measurement of the EP every 40-360 s. With future optimization of electrode design, we envision using the biologic battery in the inner ear to power chemical and molecular sensors, or drug-delivery actuators for diagnosis and therapy of hearing loss and other disorders. A remote operator was able to establish a wireless link with the microchip to program the schedule of human parathyroid hormone dosing from the device. The progress about the WPT and the active capsule technology is reviewed. Though the idea of an implantable BSN was proposed in parallel with the on-body sensor network, the development in this area is relatively slow due to the complexity of human body, safety concerns, and some technological bottlenecks such as the design of ultralow-power implantable RF transceiver. This paper describes a new wireless implantable BSN that operates in medical implant communication service (MICS) frequency band. This method offers a non-invasive alternative to traditional endoscopy and provides the opportunity for exploring distal areas of the small intestine which are otherwise not accessible. introduction to new methodologies and applications in information retrieval indexing; In order to design effective drugs, biomedical nanotechnology owns a potential prospect of industrialization when it comes to the drugs and the delivery of them in nanometer level. A thermoresponsive hydrogel serves as the actuator that is driven by the passive resonant circuit that effectively generates heat only when the field frequency is tuned to the resonant frequency of the circuit, inducing bulk squeezing of the material for drug release. To protect the vital medical information, efficient security mechanisms must be properly deployed in a resource-constrained wireless BSN, which faces more serious security challenges compared with a wired BSN, *e.g.* one that uses e-textile materials to connect the various sensors. This chapter introduces a variety of security techniques that are applicable to wireless BSNs, with emphasis on a novel biometrics method that utilizes the biological channels (bio-channels) to assist secure information transmission. We have developed a noninvasive wireless pharmaceutical compliance monitoring (PCM) system using an array of magneto-inductive sensors around the patient's neck in the form of a necklace to detect the passage of a pill or capsule embedded with a small permanent magnet as a tracer through the esophagus upon ingestion. A software has been developed in order to minimize power consumption and control the infusion mode. This paper addresses some of the issues surrounding this challenge in the domain of informatics in support of healthcare delivery and management.

Section 7. Brain stimulators (10%)

Wireless optogenetics based on the upconversion technique has recently provided an effective and interference-free alternative for remote brain stimulation and inhibition in behaving animals, which is of great promise for neuroscience research. A resonator-based three-coil inductive link creates a homogeneous magnetic field that continuously delivers

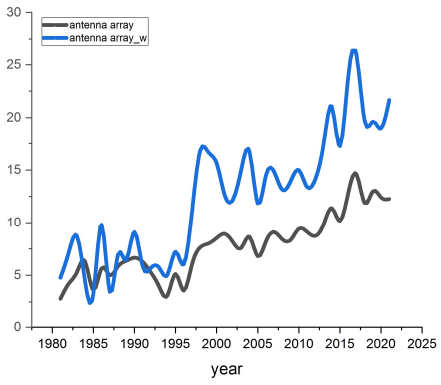
sufficient power (>2.7 mW) at an optimal carrier frequency of 60 MHz to the FF-WIOS in the near field without surpassing the specific absorption rate limit, regardless of the position of the FF-WIOS in a large brain area. LEDs typically require high instantaneous power to emit sufficient light for optical stimulation. The simulation results of these ASICs for addressing of 16 stimulation pixels on the shank are presented. Recently developed microscale light-emitting diodes (micro-LEDs), which can be wirelessly operated, serve as injectable light sources that directly interact with neural systems. The proposed system provides the potential for advanced optical neural interfaces and offers solutions to study complicated animal behaviours in neuroscience research. Both approaches are discussed with respect to size, spatial resolution, opportunity to integrate electrodes for electrical recording and potential interactions with the target tissue. Experimental possibilities include studies in naturalistic, three-dimensional environments, investigations of pairwise or group related social interactions and many other scenarios of interest that cannot be addressed using traditional hardware. Thereby, brain-implantable devices incorporating optical stimulation and low-noise data acquisition means have been designed based on custom integrated circuits (IC) to study the brain of small freely behaving laboratory animals. In this paper, we report an IC for simultaneous multichannel optogenetics and electrophysiological recording addressing both LFP and AP signals all at once. However, there are limitations and challenges with the current technologies. The results show the feasibility of utilizing wireless optogenetic nanonetworking devices (WiOptND) for long-term implants in the brain, and a new direction toward precise stimulation of neurons in the cortical microcolumn of the brain cortex. This feature allows simultaneous implantation of multiple UCNP-optrodes to achieve modulation of brain function to control complex animal behavior. Elsevier Inc. In vivo optogenetics provides unique, powerful capabilities in the dissection of neural circuits implicated in neuropsychiatric disorders. These approaches are prone to infection, vulnerable to damage and restrict the experimental approaches that can be conducted. In parallel to this, the field of optogenetics has emerged where the aim is to stimulate neurons using light, usually by means of optical fibers inserted through the skull. The wireless optogenetic neural dust is equipped with a miniature LED that is able to stimulate the genetically engineered neurons, and at the same time harvest energy from ultrasonic vibrations. With a radiofrequency (RF) power source and controller, this implant produces sufficient light power for optogenetic stimulation with minimal tissue heating (<1 °C). Wireless control and power harvesting systems that operate injectable, cellular-scale optoelectronic components provide important demonstrated capabilities in neuromodulatory techniques such as optogenetics. Here, we report a radio frequency (RF) control/harvesting device that offers dramatically reduced size, decreased weight and improved efficiency compared to previously reported technologies. Existing nontethered optical stimulators either deliver light through a cranial window limiting applications to superficial layers of the brain, are not widely accessible due to highly specialized fabrication techniques, or do not demonstrate robust and flexible control of the optical power emitted. To allow its rapid and widespread adoption, we developed this stimulator using commercially available components. The International Neuromodulation Society (INS) has determined that there is a need to provide an expert consensus that defines the appropriate use of neuromodulation technologies for appropriate patients. Despite these advances, the NACC has identified several additional promising technologies and potential applications for neurostimulation that could move this field forward and expand the applicability of neuromodulation. Light stimulation uses flexible patterns that allow for easy tuning of light intensity and stimulation periods.

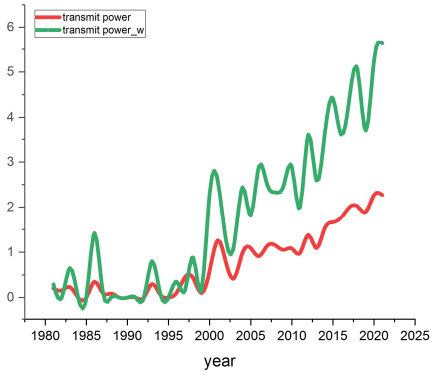
Appendix C

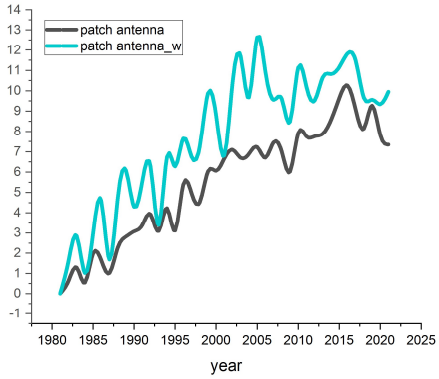
Weighted Keywords Frequency Data and Prediction Results on Antenna & Propagation Research

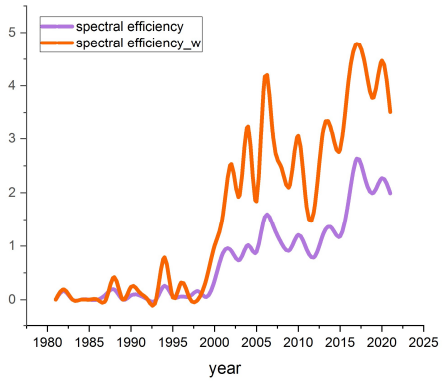
C.1 The Changes of the Keywords Trend by Applying Weights

(a) Increasing keywords

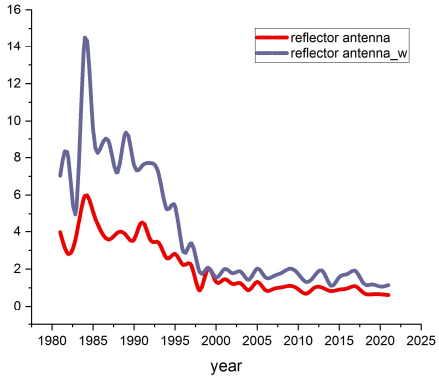


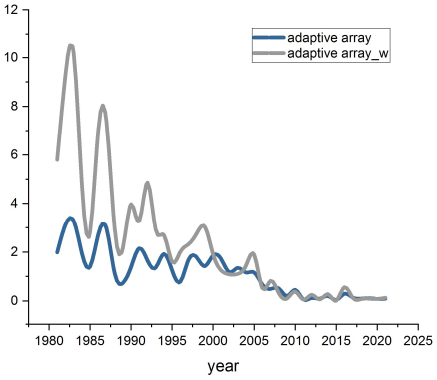


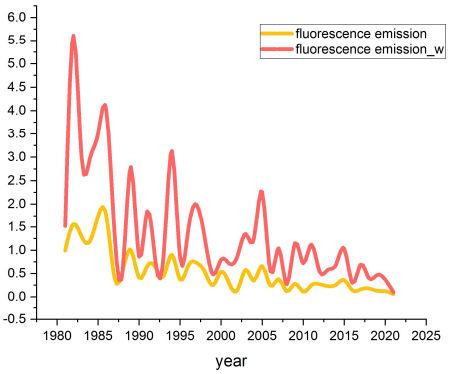


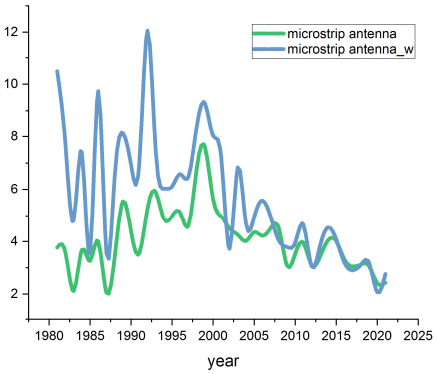


(b) Decreasing keywords

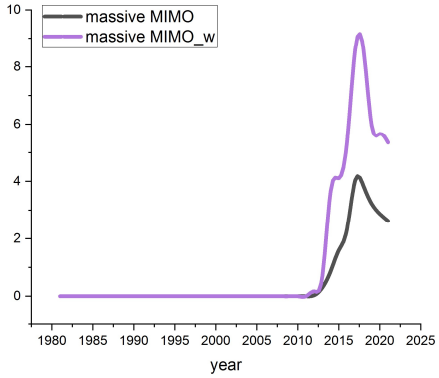


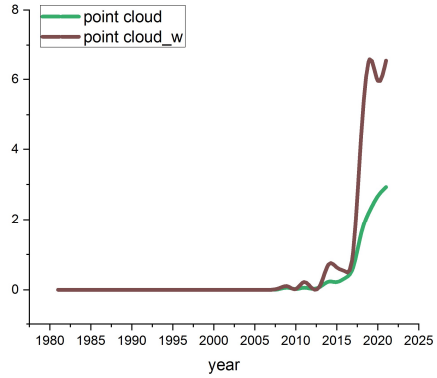


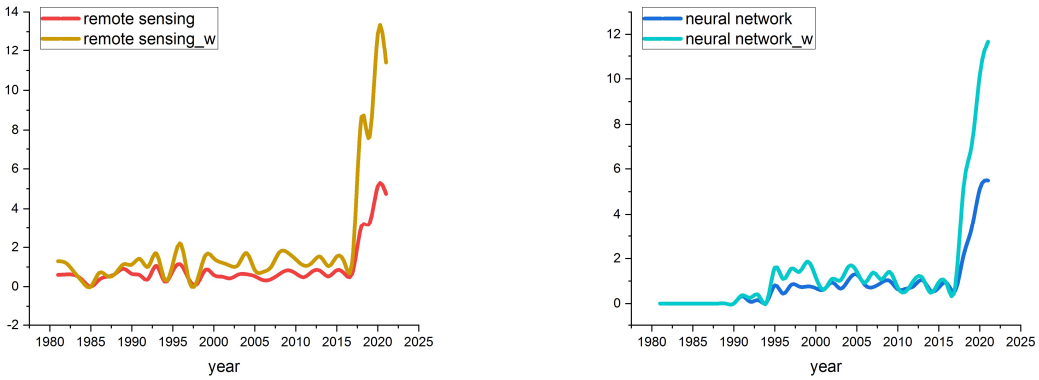




(c) Emerging keywords







C.2 Prediction Results of Some Increasing Keywords