Prediction of future research trends in Optics using Semantic Analysis and Artificial Neural Networks

Master's Thesis

Submitted By

Juilee Prasad Kulkarni

Written At

The Artificial Scientist Lab

Max Planck Institute for the Science of Light

Under the Supervison of

Dr. Mario Krenn

Group Leader, The Artificial Scientist Lab

Friedrich-Alexander-Universität Erlangen-Nürnberg

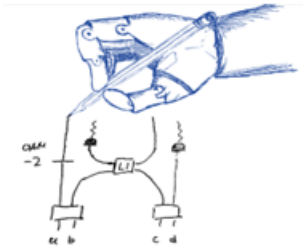
November 2022

The Participating Institutions

MAX PLANCK INSTITUTE
FOR THE SCIENCE OF LIGHT

Max Planck Institute for the Science of Light, Erlangen

MAOT
master
programme in
advanced optical technologies



Master Programme in Advanced Optical Technologies, Friedrich-Alexander-Universität Erlangen-Nürnberg

The Artificial Scientist Lab, Max Planck Institute for the Science of Light, Erlangen

Ich versichere, dass ich die Arbeit ohne fremde Hilfe und ohne Benutzung anderer als der angegebenen Quellen angefertigt habe und dass die Arbeit in gleicher oder ähnlicher Form noch keiner anderen Prüfungsbehörde vorgelegen hat und von dieser als Teil einer Prüfungsleistung angenommen wurde. Alle Ausführungen, die wörtlich oder sinngemäß übernommen wurden, sind als solche gekennzeichnet.

Die Richtlinien des Lehrstuhls für Studien- und Diplomarbeiten habe ich gelesen und anerkannt, insbesondere die Regelung des Nutzungsrechts.

Erlangen, den 29. November 2022

Acknowledgement

I would like to express my deepest gratitude to the people without whom this thesis won't have been a smooth ride. First of all, I would like to thank Prof. Dr. Florian Marquardt for letting me be a part of the amazing Theory Division at the Max Planck Institute for the Science of Light, Erlangen. I would like to thank Dr. Mario Krenn who leads the Artificial Scientist Lab at the Institute, for giving me an opportunity to work on such an interesting topic that has upgraded my skill set by exposing me to a plethora of domains like Artificial Intelligence, Natural Language Processing, Graph and Semantic Analysis. This thesis could be successful only because of Mario's valuable suggestions, scholarly guidance and constant encouragement at every step of the project. I would also like to thank Mario for organizing interesting think tanks that gave me a chance to know new things every week. I would also like to thank the other members at the Artificial Scientist Lab - Carla, Carlos, Jan, Sören, Tareq and Xuemei for helpful discussions and for providing a friendly and enjoyable atmosphere during the course of the project. Last but not the least, I would also like to thank the other friendly staff at the Max Planck Institute for the Science of Light, Erlangen for helping me out whenever required during my tenure at the institute.

Abstract

The rapid growth in the number of publications in Optics over the years calls for the development of a robust methodology to store the necessary data and utilize it in the best possible way. Semantic graphs can store huge amounts of data and the hidden relationships in it in the most efficient and easy-to-access format, thus making information management easier. The field of graph-based data analysis and tools powered by Natural Language Processing software in the back end have emerged intensely in recent times. This thesis aims at utilizing such methodologies to pick topics in the field of Optics that highlight past and ongoing research. The aim here is to interlink the research fields in Optics in an efficient manner to unleash some new, surprising, and interesting research directions in this field that might not be visible otherwise. To tackle this idea, a semantic network is developed based on 35,717 papers published on arXiv under the category, *physics.optics* by analyzing historic patterns in the titles and abstracts of these papers. The extracted patterns are then fed to an artificial neural network to predict the chances of a pair of research topics being investigated together in the future, which was not tackled together in the past. After building and testing this model, it is deployed to predict personalized research topic combinations based on the interest of one specific chosen Scientist. As a test case, this idea has been implemented for one Scientist at the Max Planck Institute for the Science of Light, Erlangen.

Contents

1 Introduction **1**

1.1 Accelerated growth in scientific publications 1

1.2 The idea of recommender systems 2

1.3 Generative models based on text data 4

1.4 Motivation and goal of the thesis 5

2 Background **7**

2.1 Related work 7

2.2 Network theory 9

2.2.1 Semantic networks 9

2.2.2 Mathematics of networks 10

2.3 Artificial neural networks 15

2.3.1 What is an artificial neural network? 15

2.3.2 Feedforward neural networks 16

2.3.3 Training a feedforward neural network 17

2.4 Computation tools 21

3 Creation of the semantic network **25**

3.1 The dataset 25

3.1.1 Data source (arXiv) 25

3.1.2 Source data structure 26

3.1.3 Data preprocessing 27

3.1.4 Keyword extraction 28

3.1.5 Final concept list 31

3.1.6 Further analysis 32

3.2 Creation of the semantic network 32

3.2.1 Nodes of the semantic network of Optics . . . 33
3.2.2 Edges of the semantic network of Optics . . . 33
3.2.3 An attempt at synonym detection . . . 34
3.2.4 Visualizing the semantic network . . . 36

4 The evolution of concepts in Optics 37
4.1 Data analysis of popular concepts . . . 37
4.1.1 Concepts evolved over 5 years . . . 37
4.1.2 Concepts evolved over 3 years . . . 44
4.1.3 Popular concept over 1 year . . . 44

5 Prediction of future research trends 47
5.1 The model . . . 47
5.1.1 The architecture . . . 47
5.1.2 Link prediction model workflow . . . 48
5.2 Model performance . . . 52
5.2.1 The confusion matrix . . . 52
5.2.2 The confusion metrics . . . 53
5.2.3 Area Under the Curve (AUC) . . . 53
5.2.4 Results . . . 54
5.3 Personalised predictions for a Scientist . . . 57
5.3.1 Data preparation and processing . . . 57
5.3.2 Personalised predictions . . . 58

6 Conclusion and Outlook 63

List of Figures 65

List of Tables 69

Bibliography 70

Chapter 1

Introduction

1.1 Accelerated growth in scientific publications

Over the years, the number of publications in the field of Science has grown at a rapid rate. arXiv, which is a free distribution service and an open-access archive for scholarly articles in the fields of physics, mathematics, computer science, quantitative biology, quantitative finance, statistics, electrical engineering and systems science, and economics [1], in 2018, received 140,616 new submissions, a 14% increase from 2017 [1]. In 2019, the repository received 155,866 new submissions, an 11% increase from 2018 [2]. Based on the numbers published in the arXiv annual report for the year 2021 [3], there were 181,630 new submissions in total in 2021. The number of submissions per month was 15,500 [3].

After analyzing the research data published in the field of Optics on arXiv, a visualization as shown in figure 1.1 was obtained. The blue-colored plot indicates the fast growth in the number of published articles. There are many factors affecting the growth of publications. For example, new research fields coming up and the number of researchers has increased over the years, leading to a vast addition to the scientific literature. In figure 1.1, the orange-colored plot shows the growth trend in the number of authors.

With the current number of scientific publications, it can be safely assumed that this number is going to escalate more rapidly in the future with new emerging fields and new bright minds joining the research community. Therefore, a new methodology is needed to go through the scientific literature to pave the way for new research directions. This can lead to discoveries that might seem out of our imagination at the first glance.

[1] Source: https://arxiv.org/

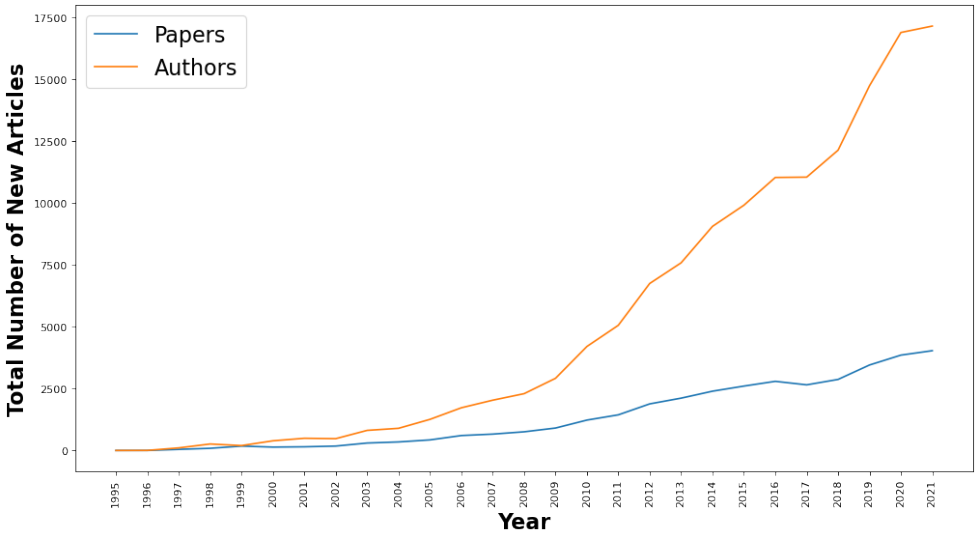


Figure 1.1: Growth in the number of scientists and publications in the field of Optics during the past century

A computer algorithm with access to a large corpus of published scientific research could potentially make genuinely new contributions to science [5]. This level of automation of science is more in the realm of science fiction than reality at present [5]. Algorithms capable of extracting semantic knowledge from the scientific literature can be employed to show researchers a guiding path. As an example, the evaluation of whether an idea is novel or surprising depends crucially on already-existing knowledge. Thus, a computer algorithm with the capability to propose new useful ideas or potential avenues of research will necessarily require access to published scientific literature—which forms at least partially the body of human knowledge in a scientific field [5]. Knowledge can be structured and represented using semantic networks that represent semantic relations between concepts in a network. Over the last few years, significant results have been obtained by automatically analyzing the large corpus of scientific literature, including the development of semantic networks in several scientific disciplines [5] [6] [7].

1.2 The idea of recommender systems

The way Google puts search results according to our query in front of us and then later provides us with all the relevant suggestions before we even finish typing the query is the best example of what magic structured stored information can create. When we open Instagram, the application suggests us to add people that are within our circle because of

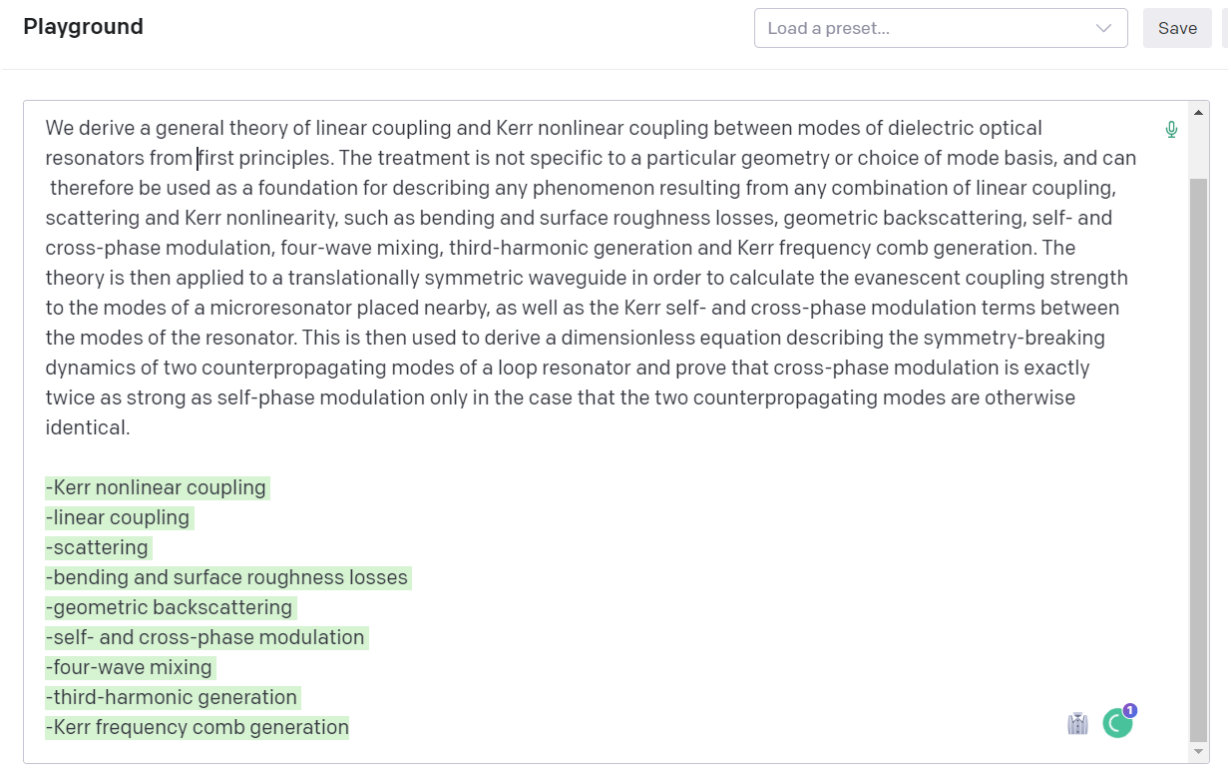


Figure 1.2: The user interface of the GPT 3 tool. Here, an abstract from a scientific text [4] is entered in order to get the keywords out of it. The words highlighted in green are the extracted words. It can be seen that some of the important words like *symmetry breaking*, *dielectric optical resonator*, *evanescent coupling* are not even extracted. This indicates that we might lose some important information discussed in the piece of text.

some mutual friends. When we open online shopping platforms like Amazon, the backend algorithm suggests products that might be interesting to us. Personal assistants like Siri, Alexa, or Cortana rely on Artificial Intelligence to transcribe our commands into text. The thing to think about here is the aspects of science that can be augmented by this kind of Artificial Intelligence algorithms. Modern tools can identify and personalize the papers that should be read. One of these which is widely used is Google Scholar [2] which enables searching of papers, viewing of paper statistics as well as citations and references, setting up alerts for new papers by following an author or a paper, and keeping a basic library with automatic recommendations. Another one is Semantic Scholar [3] which analyzes papers semantically with external material aggregation and suggests recommended papers. These new capabilities will help researchers expand the depth and quality of knowledge as well as help identify new research possibilities. For decision-makers in science, Artificial Intelligence could offer a more comprehensive horizon scanning capability, suggesting areas for strategic investment, and identifying ideas [8]. Publishers may also use such tools to identify which referees to seek for a manuscript or to automatically identify apparent flaws and inconsistencies in a manuscript, avoiding the need to bother human reviewers [8].

1.3 Generative models based on text data

There have been many new advances made in the field of applications of Artificial Intelligence in Science. Worth mentioning is GPT-3 where GPT stands for Generative Pre-trained Transformer. It is a machine learning model trained using internet data to generate any type of text. Developed by OpenAI, it requires a small amount of input text to generate large volumes of relevant and sophisticated machine-generated text. The GPT-3 model has over 175 billion machine learning parameters [9]. This model is trained to generate realistic human text based on the information from the internet [9]. GPT-3 has been used to create articles, poetry, stories, news reports, and dialogue using just a small amount of input text. It is also being used for automated conversational tasks, responding to any text that a person types into the computer with a new piece of text appropriate to the context. It can create anything with a text structure, and not just human language text. It can also automatically generate text summaries and even programming code [9]. In this work, extraction of important words from a piece of text is an important factor. After coming across such large language models like GPT3, the first thought that comes to the

[2]Source: https://scholar.google.com/intl/us/scholar/help.htmlcover
[3]Source: https://www.semanticscholar.org/

mind is using such tools to quickly extract the keywords. However, GPT-3 has no internal representation of what each of the words used by it mean [4]. It has no semantically-grounded model of the world or of the topics on which it discourses [4]. This implies that GPT-3 works with only statistical computations and does not work by understanding the input and output text's content. Thus, GPT-3 struggles in reasoning capabilities [5]. Figure 1.2 illustrates that we also lose some information if we process our text based on such models. Therefore, even though this large language model has high power and potential, it cannot be used to draw conclusions from a piece of scientific text by automatic information extraction and so it is not useful here as we are also interested in the context in which a particular piece of text has been mentioned.

1.4 Motivation and goal of the thesis

With the advances made in the field of Artificial Intelligence, the question that arises now is can such AI-powered systems come up with creative and better research ideas faster? It makes sense to use such tools to suggest new hypotheses and even new areas to explore. Some of the proven successful artificially intelligent tools have two critical ingredients. The very first and the heart of such a tool is a good and large amount of data and the second one is a clear method to analyze it for patterns [8]. The scientific community can benefit a lot from the potential the modern tools have to offer.

The vast and growing number of publications in the discipline of Optics cannot be comprehended by a single human researcher. It is a tedious process to go through all the papers that have been published and get a guiding path to have an overview of interesting research directions. Having all the published information in a structured format will not only make it easily accessible but also help scientists to discover interesting fields to investigate even if these fields don't fall under the current set of their research interests. Using network theoretical tools, we can suggest personalized, out-of-the-box ideas by identifying pairs of concepts, which have unique semantic network properties [5]. The strength of the recently developed machine learning algorithms can be used to mine hidden surprising facts from time-dependent data as well.

The goal is to develop a dynamic semantic graph. This semantic graph will consist of a number of vertices and the connections between every pair of vertices will be based on a common factor between the two. This semantic graph is then used to build a system that

[4]Source: https://www.univ.ai/post/the-limitations-of-gpt-3-and-its-impact-on-society

suggests relevant research topics that can be combined to provide interesting directions for future research in Optics. The next aim is to use this system to provide personalized research suggestions for Scientists to drive their research in Optics towards innovation and amazing breakthroughs. This target is achieved by building an efficient neural network to discover the patterns in the generated semantic graph and get the most probable topics that would be interesting to study together after a defined time interval.

Chapter 2

Background

2.1 Related work

Looking at structured and linked information collected from different sources can be highly informative rather than coming across the different pieces of information separately. Using connected research data can be highly beneficial to scientists. After analyzing the abstracts of millions of biomedical papers published during the period 1983 to 2008, scientists identified the names of biomolecules that were studied together in one paper [7]. This information was then stored in the form of connections between vertices where each vertex represented the name of a biomolecule and the connection between them was their co-appearance in the same paper [7]. This organized data turned out to be a powerful tool to infer the strategy scientists have to explore a novel chemical relationship. For example, figure 2.2 shows that scientists tend to explore the neighborhood of prominent chemicals. The crowded space in the graph indicates that multiple researchers focus on their investigations on a very congested neighborhood of the discoverable space rather than exploring the space of the unknown and unique pairs more broadly [7]. This work highlights the importance of co-appearance of topics in the same research article. Co-appearance indicates that the two connected entities were used to achieve a common goal in a particular research study. In this work, we also try to structure the information based on the idea of co-appearance of research fields in the title or abstract of the same research article.

In another study, future research trends in the field of Quantum Physics were predicted with semantic and neural networks [5]. Here, a method to build a semantic network from published scientific literature was demonstrated. The semantic network is used to predict future trends in research relevant to Quantum Physics. In the semantic network, scientific

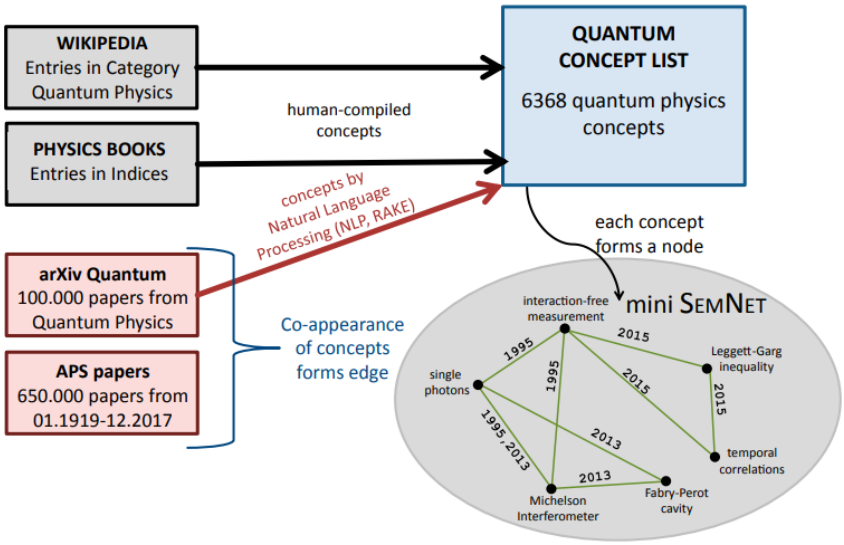


Figure 2.1: Diagrammatic inner working of SEMNET. Human-generated concept lists (from Wikipedia and books) are combined with automatically generated lists (with natural language processing, using RAKE on 100,000 arXiv articles) to generate a list of quantum physics concepts. Each concept forms a link in a semantic network. The edges are formed when two concepts coappear in a title or abstract of any of the 750,000 papers (from arXiv and APS). A mini-version of SEMNET is shown, using parts of three articles from APS. Edges carry temporal information of their formation year, which leads to an evolution of the semantic network SEMNET over time [5]. (Image from [5]).

knowledge is represented as an evolving network using the content of scientific papers published since 1919 [5]. The nodes of the network correspond to physical concepts, and links between two nodes are drawn when two physical concepts are concurrently studied in research articles. To form connections between different quantum physics concepts, 100,000 articles of quantum physics categories on arXiv and the dataset of all 650,000 articles ever published by the American Physical Society (APS) were extracted [5]. Thus, in total, 750,000 research articles were processed for this work [5]. These two data sources were chosen because the APS database contains peer-reviewed physics papers from the last 100 years (allowing for investigation of long-term trends), while the arXiv database contains specific quantum physics papers, allowing for more precise coverage of the quantum physics research trends. In this work, the list of keywords is generated using two independent methods. One set is a human-made list of physical concepts. These concepts are compiled from the indices of 13 quantum physics books which were available in a digital form. Other than this, titles of Wikipedia articles that are linked in a quantum physics category were used. This human-made collection contains around 5,000 entries of physical concepts.

Then, the human-generated list is extended with an automatically generated list of physical concepts obtained by processing the titles and abstracts of around 100,000 articles published in quantum physics categories on the arXiv [5]. The human- and machine-generated lists of concepts were combined and optimized to delete incorrectly identified concepts. Ultimately, this yielded a list of 6,300 terms [5]. Influential and prize-winning research topics from the past were identified inside the semantic Nnetwork which confirmed that the Network stored useful semantic knowledge [5]. A deep neural network was trained here using states of the past network, to predict future developments in quantum physics research, and confirm high quality predictions using historic data. The 17 properties for each unconnected concept pair in the semantic network are used by the neural network to estimate which pairs of quantum physics concepts are likely to be connected within 5 years and which are not [5]. The idea of extracting keywords and storing these in a structured form inside a semantic network is a crucial component of this work.

In another piece of work, Artificial Intelligence was used to predict the future research directions of Artificial Intelligence itself [6]. More than 100,000 research papers were used to build up a knowledge network with more than 64,000 concept nodes [6]. Ten diverse methods to tackle this task, ranging from pure statistical to pure learning methods were presented in this paper [6]. It was observed that the most powerful methods use a carefully curated set of network features [6]. Extracting the features based on the connection patterns in the semantic network is something that has been implemented in this work as well.

2.2 Network theory

2.2.1 Semantic networks

Network is a graph that represents symmetric or asymmetric relations between discrete objects. A semantic network is a graphic notation for representing knowledge in patterns of interconnected nodes [10]. The structural idea is that knowledge can be stored in the form of graphs, with nodes representing some real-world objects, and edges representing relationships between those objects. The edge labels have information about the relationships to provide the basic needed structure for organizing the knowledge. Some real-world examples of semantic networks have been depicted in figure 2.3.

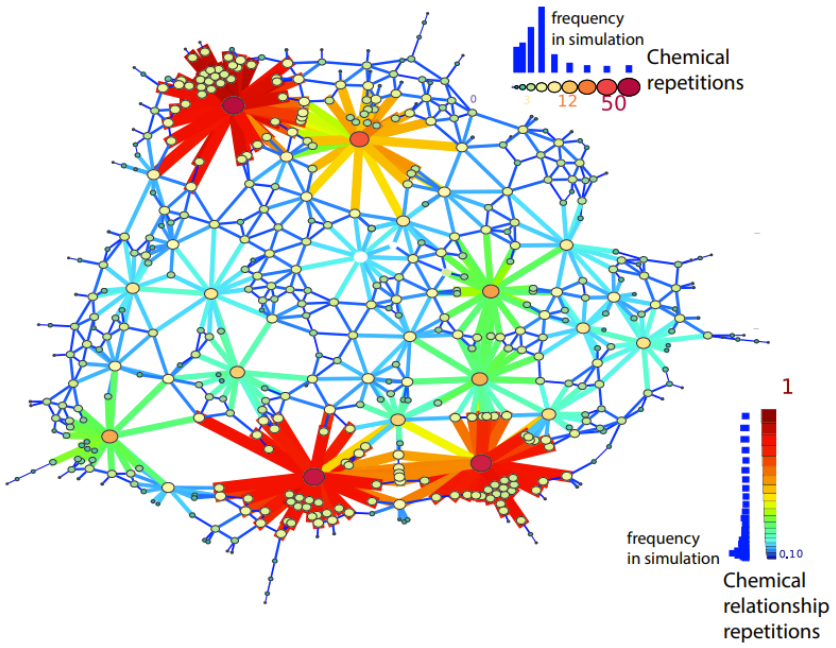


Figure 2.2: The semantic network in Biochemistry. The nodes represent the names of biomolecules. An edge is drawn to connect two biomolecules if these two have been studied together in one published article (Image from [7]).

2.2.2 Mathematics of networks

The adjacency matrix

There are several ways of representing a network mathematically. Networks are graphs made up of nodes and edges. Representing a graph as a matrix model makes it easy for us to analyze it to get insights into the patterns hidden in the connections within the graph. One of the ways to encode graph information is 'The Adjacency Matrix'. This matrix gives information about the connections in the graph. The elements of the matrix indicate whether pairs of vertices are adjacent or not in the graph [13]. If the vertices are adjacent to each other, the adjacency matrix denotes whether the adjacent vertices are connected or not [13]. This can be understood better with the help of an example. Consider an undirected network with 6 vertices as shown in figure 2.4. The adjacency matrix A [13] of a simple graph is the matrix with elements A_{ij} such that:

- $A_{ij} = 1$ if an edge exists between vertices i and j
- $A_{ij} = 0$ if the vertices have no edge between them

(a) The protein interaction network of T.Pallidum. The nodes represent proteins and the edges are drawn if the 2 proteins interact with each other (Image from [11]).

(b) Empirical and simulated mobility network for France, Germany, and UK. The network depicts the flows between cities as observed in the period from 2000 to 2006. The thickness of the edges is proportional to the log of the number of moves between the two cities (Image from [12]).

Figure 2.3: Examples of semantic networks

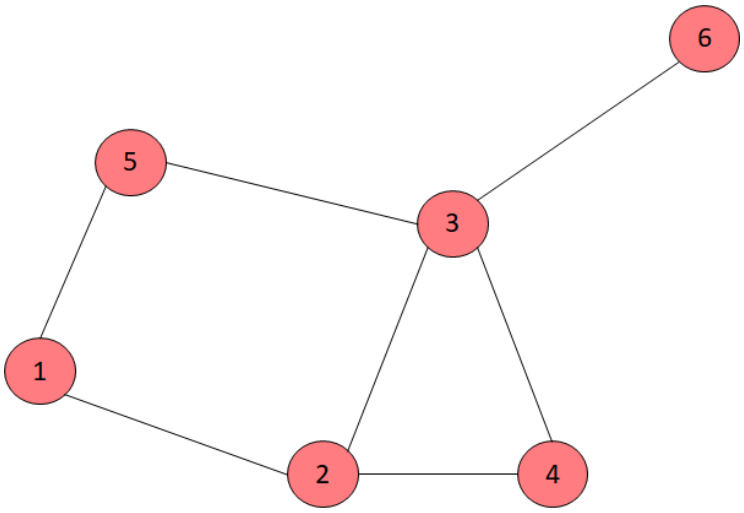


Figure 2.4: A simple graph of 6 nodes and 7 edges. The vertices are labeled as 1, 2, 3, 4, 5, and 6. Each label is unique so that the labels can be used to refer to any vertex unambiguously. If there is an edge between vertices i and j, this connection can be denoted as (i,j). In this way, the complete network can be specified by giving the number of edges and a list of all the edges. So, the graph in figure 2.3 has edges, (1,2), (1,5), (2,3), (2,4), (3,4), (3,5), and (3,6).

Two points to notice about the adjacency matrix are that, the diagonal matrix elements are all zero, and second that it is symmetric, since if there is an edge between i and j then there is an edge between j and i. Let us consider this with an example element A_{12} and A_{21}. The value of both these elements is 1 because nodes 1 and 2 are connected in the graph in figure 2.4. The explanation is similar to the other matrix elements.

$$\text{A (Adjacency Matrix of Graph in figure 2.4)} = \begin{bmatrix} 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix}$$

Cosine similarity

After defining a graph, it is important to find out the structural similarities hidden in it. One of the measures to do the same is counting the number of common neighbors the 2 vertices have. This is when we define the term, 'cosine similarity'. This value falls in the range of 0 and 1 [13]. A cosine similarity of 1 indicates that two vertices have exactly the

same neighbors [13]. A cosine similarity of zero indicates that they have none of the same neighbors [13] and also indicates that the vertices are far apart semantically.

Now, let us understand how we use this powerful concept in network analysis. In the previous section, we had an overview of the adjacency matrix. On the previous page, there is also a demonstration of how an adjacency matrix A is formed for a simple graph shown in figure 2.4. The element of the matrix is 1 if there is a connection between the two vertices and 0 otherwise. If we have a look at the elements with value 1 from a different perspective, it can be seen that you need to cover one edge to go from one node to another.

Now, if we consider the nodes that need two edges to reach from one to the other. This means that these two nodes have one node that you need to cross to reach the other. An example of such a scenario can be given with the help of the graph in figure 2.5. Consider node 1 and node 3. If you start at node 1 and you have to reach node 3, you have to either go past node 5 or node 2. In both cases, you need to cover two edges in total, that is, in total, two paths. You can either first go from node 1 to 2 and then node 2 to 3. The other option is going from 1 to 5 and then to 3. In the first case, the two paths are 1-2 and 2-3. In the second one, these are 1-5 and 5-3. Other pairs of nodes that have paths of length 2 between them are [(1, 4), (2, 3), (2, 4), (2, 5), (2, 6),(3, 4), (4, 5), (4, 6), (5, 6)]. The next step is to capture this information in the form of a matrix for speedy computation. Now, let us square the adjacency matrix. After squaring, the resulting matrix is given below. Let us call this matrix B.

$$\mathrm{B} = \begin{bmatrix} 2 & 0 & 2 & 1 & 0 & 0 \\ 0 & 3 & 1 & 1 & 2 & 1 \\ 2 & 1 & 4 & 1 & 0 & 0 \\ 1 & 1 & 1 & 2 & 1 & 1 \\ 0 & 2 & 0 & 1 & 2 & 1 \\ 0 & 1 & 0 & 1 & 1 & 1 \end{bmatrix}$$

If we have a closer look at the square of the adjacency matrix, we can draw the following conclusions:

- Every diagonal element denotes the total number of nodes connected to the node under consideration. If we consider, A_{33}, the value is 4. Now in figure 2.5, node 3 is connected to nodes 2, 4, 5, and 6, which is 4 nodes in total. The logic is same for A_{11}, A_{22}, A_{44}, A_{55} and A_{66}.

- Now, let us have an overview of the logic behind the values of the other matrix elements. If we have a look at the element A_{52}, the value is 2. This value comes

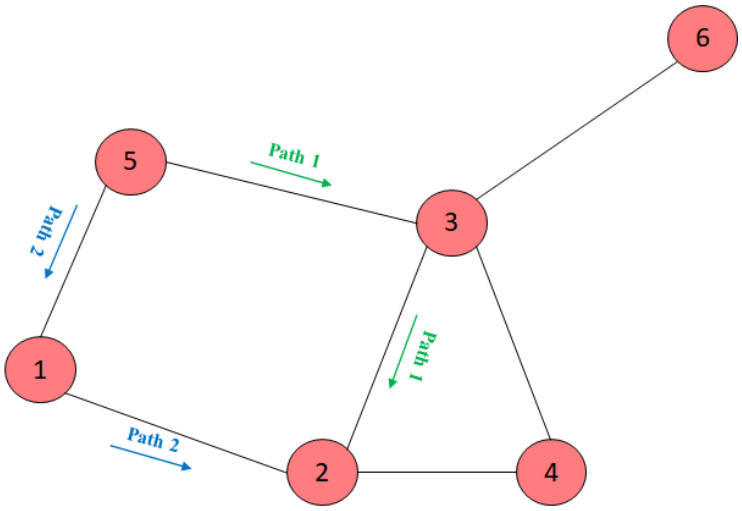


Figure 2.5: Number of paths of length 2. Here, nodes 2 and 5 are considered as an example. These nodes have 2 paths of length 2, which means there are two possibilities where we need to cross 2 edges to reach from one node to the other.

from the number of possible paths of length 2 to reach node 5 to node 2. This logic is illustrated in figure 2.5. The other matrix elements are calculated with the same approach and this is how we get this matrix.

- Thus, raising the adjacency matrix with a certain degree, that is, $[A^n]_{ij}$, means finding the total number of paths of length n which connect node i and node j. Matrix computations are time and cost-consuming so it is better to stop at n=2 as it is also efficient.

We now have all the building blocks to calculate the cosine similarity. Cosine similarity between node i and node j is expressed as given below [13].

$$Cosinesimilarity(i,j) = \frac{B_{ij}}{\sqrt{B_{ii} \times B_{jj}}}$$

So, suppose, if we want to find cosine similarity between node 1 and node 3, the approach would be as follows.

$$CosineSimilarity(1,3) = \frac{2}{\sqrt{2 \times 4}} = 0.707$$

The computation shown above can be performed on all pairs of nodes present in the graph to check how similar they are to proceed with further analysis.

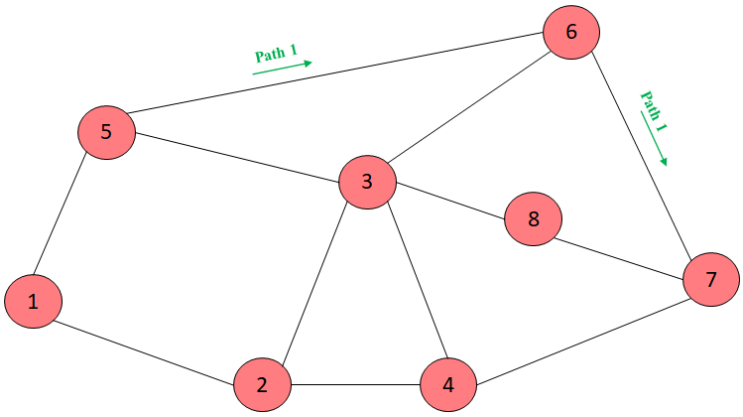


Figure 2.6: A simple graph (2)

Consider another graph shown in figure 2.6. Now, we have to compute the cosine similarity between node 5 and node 7 of this graph. As discussed in the section before, we need the degree of node 5, node 7, and the number of paths of length 2 to reach node 7 from node 5. Now, let us have a closer look again at the graph in figure 2.6. Node 5 is connected to 3 other nodes (node 1, node 3, and node 6) and so the degree of node 5 is 3. Node 7 is connected to 3 other nodes (node 4, node 6, and node 8) and so the degree of node 7 is 3. Now, in figure 2.6, the path highlighted in green color is the only one that has length 2 and the one that connects node 5 and node 7 with node 6 in between. So, the cosine similarity can be computed as given below now. The value of the cosine similarity is very less in this case which means that there is nothing much in common between node 5 and node 7.

$$Cosinesimilarity(5,7) = \frac{1}{\sqrt{3 \times 3}} = 0.333$$

2.3 Artificial neural networks

2.3.1 What is an artificial neural network?

An artificial neural network is an interconnected assembly of simple processing elements, units or nodes, whose functionality is loosely based on the animal neuron [14]. Similar to the human brain which has neurons interconnected to one another, artificial neural networks also have neurons that are interconnected to one another in various layers of the networks in a chain [15]. These neurons are known as nodes. The dendrites from biological neural networks represent inputs in artificial neural networks, the cell nucleus represents

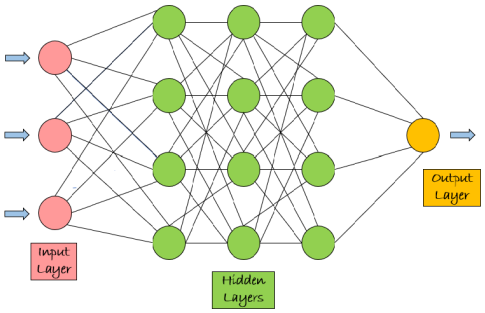


Figure 2.7: A general artificial neural network architecture

the nodes, the synapse represents the weights and the axon represents the output [14]. An artificial neural network in the field of Artificial Intelligence attempts to mimic the network of neurons that make up a human brain so that computers will have the option to understand things and make decisions in a human-like manner.

2.3.2 Feedforward neural networks

A feedforward neural network allows signals to travel in one direction only, from the input to the output. There are no feedback loops such that the output of some layer influences that same layer. Feedforward networks tend to be simple networks that associate inputs with outputs [1]. A general structure of a feedforward neural network is shown in figure 2.7.

- **Input layer:** This layer consists of the neurons that receive inputs and pass them on to the other layers. The number of neurons in the input layer should be equal to the number of features extracted from the dataset.

- **Hidden layer:** In between the input and output layers, there are hidden layers. Hidden layers contain a vast number of neurons that apply transformations to the inputs before passing them.

- **Output layer:** The output layer is the predicted feature.

- **Neuron weights:** Weights refer to the strength or amplitude of a connection between two neurons. As the network is trained, the weights are updated.

[1]Source: https://www.tutorialspoint.com/what-is-feed-forward-neural-networks

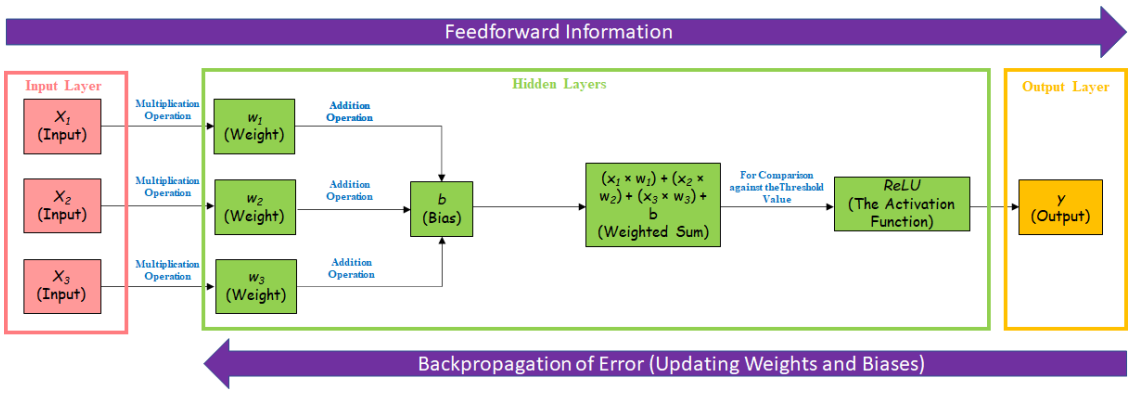


Figure 2.8: Computational model of a feedforward neural network

2.3.3 Training a feedforward neural network

Now, let us have a closer look at the different layers discussed in section 2.3.2 and how they come into the picture during training (refer to figure 2.8 for an illustration of the discussion in this section). This will be easier with the help of an example. Suppose, we want to work on a weather forecasting problem [2] to predict the chances of rain based on the following input data:

- Time (Day or Night)
- Temperature
- Month

Let us first store the data of these three parameters into three different variables - x_1 (Time (Day or Night)), x_2 (Temperature) and x_3 (Month). Then, let us assume the threshold value to be 20. This means that if the output value is higher than 20 then it will be raining, otherwise, it's a sunny day. Consider an input data tuple (x_1, x_2, x_3) as (1, 10, 6), initial weights of the feedforward network (w_1, w_2, w_3) as (4, 1, 2) and bias as 1. Following are the steps that will be followed to compute the output.

- **Multiplication of weights and inputs:** The input is multiplied by the assigned weight values. In this case, it would be as follows:

$$x_1 \times w_1 = 1 \times 4 = 4$$

$$x_2 \times w_2 = 10 \times 1 = 10$$

[2]Source: https://builtin.com/data-science/feedforward-neural-network-intro

$$x_3 \times w_3 = 6 \times 2 = 12$$

- **Adding the biases:** Every layer has a bias to determine whether or not the activation output from a neuron is going to be propagated forward through the network. With an activation output of zero, the neuron would not activate or in other words, would not fire. Thus, no information from these non-activated neurons will be passed forward to the rest of the network. Essentially, zero is the threshold here for the weighted sum in determining whether a neuron is firing or not. This is where bias comes into the picture to help us adjust the threshold. The bias gets added to the weighted sum before being passed to the activation function. Thus, the model now has increased flexibility in fitting the data since it now has a broader range of what values it considers as being activated or not. In this step, the product computed by multiplying weights and inputs is added to the bias. The modified inputs are then summed up to a single value.

$$\textit{weighted sum } (y) = (x_1 \times w_1) + (x_2 \times w_2) + (x_3 \times w_3) + b = 4 + 10 + 12 + 1 = 27$$

- **Activation:** An activation function is used to map the summed weighted input to the output of the neuron. It is called an activation function because it governs the inception at which the neuron is activated and the strength of the output signal. There are several activation functions for different use cases. The most commonly used activation functions are ReLU, tanh, and softmax. In this work, we use ReLU which stands for Rectified Linear Unit. ReLU is used because it saves a lot of computation time by accelerating the training speed as the derivative of ReLu is 1 for positive input [16]. Due to a constant value, neural networks do not need to take additional time for computing error terms during the training phase. ReLU, the activation function, determines the value of the output.

- **Output signal:** Finally, the weighted sum obtained is turned into an output signal by feeding the weighted sum into an activation function. As the weighted sum in our example is 27, which is greater than our threshold value of 20, the model predicts it to be a rainy day.

The method by which the inputs are tranformed to achieve the output signal is illustrated in figure 2.9.

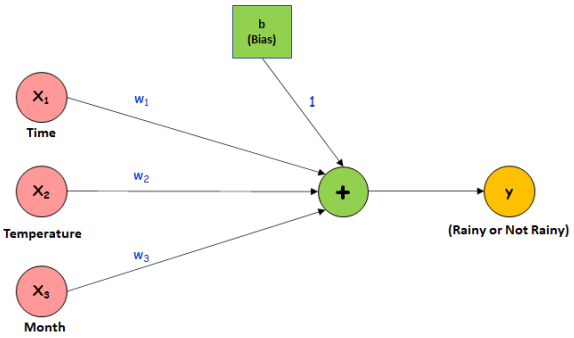


Figure 2.9: Computational model of a neuron. This illustrates the way in which inputs are transformed to predict an output. The illustration is based on the weather forecasting problem discussed in section 2.3.3.

Loss function

Now, we need a metric to detect the quality of the predicted output. A loss function compares the target and predicted output values. It measures how well the neural network models the training data. During training, we aim to minimize this loss between the predicted and target outputs. In this work, the loss function used is, 'mean squared error (MSE)'. The loss is the square of the difference between true and predicted values [**mse**]. It is the average of the calculated loss functions for all training examples in the training set. The loss function has an important job in that it must faithfully distill all aspects of the model down into a single number in such a way that improvements in that number are a sign of a better model. Mathematically speaking, it can be denoted as given below [**mse**].

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_{\mathrm{i}} - \hat{y}_{\mathrm{i}})^2$$

Here, $\hat{y}_{\mathrm{i}}$ is the predicted value and N is The number of samples we are testing against. In the case of mini-batches (discussed under the section 'Backpropagation'), N stands for the batch size after which the parameters are changed.

Optimization

Gradient descent is the optimization technique used in this work. The term, gradient, refers to the quantity change of output obtained from a neural network when the inputs change a little. Gradient Descent incrementally adjusts the parameters based on the gradient of the

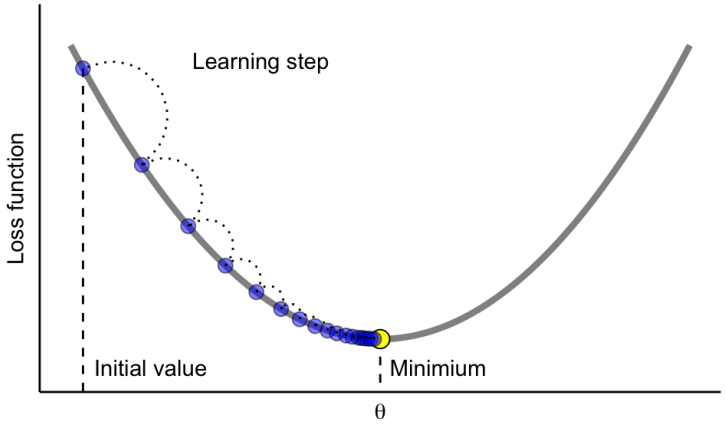


Figure 2.10: The Gradient descent algorithm. The approach here is to change the training parameters at a suitable learning rate to find the minimum loss point (Image from [17]).

parameters. The gradient descent algorithm has the central equation as given below.

$$\theta_{i+1} = \theta_i - \mu \times \nabla\theta_i$$

Here, θ_i is the parameter, μ is the learning rate and i is the number of the ongoing iteration over the samples under consideration. The learning rate indicates how fast the parameters should be changed as the machine learns [17]. To summarize, the gradient descent method is used to minimize the loss. It is like going down a slope where you have to determine how fast you should walk to reach the foot of the slope [17]. This idea is illustrated in figure 2.10.

Backpropagation

The backpropagation algorithm tells how a machine should change its internal parameters [18] like weights and biases that are used to compute the representation in each layer from the representation in the previous layer [18] in order to improve the network's performance. We cannot change the activations directly as we have control only over the weights and biases. The activation is the weighted sum of all the activations from the previous layers. In the example of rain prediction, this value is 27. In case this value comes out to less than 20, which is the threshold that we set earlier, we need to change the weights and biases such that the activation increases. The same backpropagation process has to be implemented for every other training example, recording how the weights and biases should be adjusted for each one of them and then an average is computed together with all those desired changes.

The weights and biases are optimized based on the Gradient Descent Algorithm discussed in the immediately previous section. It is a time-consuming computation to add up the influences of every single training example. So, to speed up the process, the training data is randomly shuffled and divided into several mini-batches. Then the Gradient Descent is computed according to each mini-batch rather than the entire set of training examples. Each mini-batch gives a good approximation of the gradient of the loss function [3]. At the end of the batch, the predictions are compared to the expected output variables and an error is calculated. This error is then propagated back within the whole network, one layer at a time [3]. We start at the output layer and propagate backward, updating weights and biases for each layer, except the input layer, to have the desired output in the output layer. In this way, by repeating the process, the model improves its performance and gets trained to solve the task more efficiently for which it is designed.

2.4 Computation tools

This section will give a brief overview of some important tools that were used to build up the computer program for this thesis. There are many other required back-end tools as well but this section aims at highlighting only the major game changers.

Pytorch

The neural network architecture that is discussed in section 5.1.1 is developed in the Pytorch [19] framework. The training and testing of the neural network are completely done with this tool. PyTorch is an open-source machine learning framework. Pytorch is widely used in academia and industry for applications such as computer vision and natural language processing [4]. This powerful framework was originally developed by Adam Paszke, Sam Gross, Soumith Chintala, and Gregory Chanan, under the umbrella of Meta AI [20].

- ***GitHub source repository link:*** *https://github.com/pytorch/pytorch*

Natural language toolkit

The major part of this thesis is processing text data generated from research papers and extracting relevant information from these papers. The Natural language toolkit (NLTK)

[3]Source: https://www.3blue1brown.com/lessons/backpropagation
[4]Source: https://dl4nlp.info/en/latest/

```
from nltk.stem import WordNetLemmatizer
print(WordNetLemmatizer().lemmatize('waves'))
print(WordNetLemmatizer().lemmatize('systems'))
print(WordNetLemmatizer().lemmatize('sources'))
```

```
wave
system
source
```

Figure 2.11: Lemmatizing words using NLTK

```
import nltk
nltk_stop_list=nltk.corpus.stopwords.words('english')
print(nltk_stop_list)
```

```
['i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'ourselves', 'you', "you're", "you've", "you'll", "you'd", 'your', 'yours', 'y
ourself', 'yourselves', 'he', 'him', 'his', 'himself', 'she', "she's", 'her', 'hers', 'herself', 'it', "it's", 'its', 'itself',
'they', 'them', 'their', 'theirs', 'themselves', 'what', 'which', 'who', 'whom', 'this', 'that', "that'll", 'these', 'those',
'am', 'is', 'are', 'was', 'were', 'be', 'been', 'being', 'have', 'has', 'had', 'having', 'do', 'does', 'did', 'doing', 'a', 'a
n', 'the', 'and', 'but', 'if', 'or', 'because', 'as', 'until', 'while', 'of', 'at', 'by', 'for', 'with', 'about', 'against', 'b
etween', 'into', 'through', 'during', 'before', 'after', 'above', 'below', 'to', 'from', 'up', 'down', 'in', 'out', 'on', 'of
f', 'over', 'under', 'again', 'further', 'then', 'once', 'here', 'there', 'when', 'where', 'why', 'how', 'all', 'any', 'both',
'each', 'few', 'more', 'most', 'other', 'some', 'such', 'no', 'nor', 'not', 'only', 'own', 'same', 'so', 'than', 'too', 'very',
's', 't', 'can', 'will', 'just', 'don', "don't", 'should', "should've", 'now', 'd', 'll', 'm', 'o', 're', 've', 'y', 'ain', 'ar
en', "aren't", 'couldn', "couldn't", 'didn', "didn't", 'doesn', "doesn't", 'hadn', "hadn't", 'hasn', "hasn't", 'haven', "have
n't", 'isn', "isn't", 'ma', 'mightn', "mightn't", 'mustn', "mustn't", 'needn', "needn't", 'shan', "shan't", 'shouldn', "should
n't", 'wasn', "wasn't", 'weren', "weren't", 'won', "won't", 'wouldn', "wouldn't"]
```

Figure 2.12: Generation of stop words in English language using NLTK

[21] comes into the picture now. NLTK [21] is a well-known python package to deal with human language data. It is a very efficient library to analyze natural language. In this work, we use this package to lemmatize the words in the text. This helps us to convert all the plural forms to the respective singular forms. An implementation example is illustrated in figure 2.11. NLTK provides easy-to-use interfaces to over 50 corpora and lexical resources such as WordNet [21]. Thus, in the very initial phase of data processing, we use NLTK to generate the common list of stop words (refer the Step 1 of the section, 'Workflow of the keyword extraction algorithm' on page 29 to understand what stop words are) in the English language. This makes it easy to clean the text in the first round of data processing. An example of a code snippet to generate these stop words is illustrated in figure 2.12.

- ***GitHub source repository link:*** *https://github.com/nltk/nltk*

Rapid Automatic Keyword Extraction (RAKE)

The algorithm used for the extraction of keywords is RAKE [22]. The methodology to extract keywords from the research data is discussed in detail in section 3.1.4.

- ***GitHub source repository link:*** *https://github.com/csurfer/rake-nltk*

NumPy

NumPy [23] is the fundamental package for scientific computing in Python. It is a Python library that provides a multidimensional array object, and various derived objects such as masked arrays and matrices. This package is an easy way for performing logical operations on arrays. At the core of the NumPy package, is the ndarray object. This encapsulates n-dimensional arrays of homogeneous data types, with many operations being performed in compiled code for performance. Vectorization going on in the back end in the form of a pre-compiled C code leads to the absence of any explicit looping and indexing in the code [23]. Thus NumPy makes computations fast and thus, it has been used a lot in this work. The adjacency matrix operations discussed in section 2.2.2 have been performed using NumPy. Other than this, the performance of the model discussed in section 5.1, during training and testing is also stored using NumPy. Numpy comes into the picture for several other small computations as well in this work.

- ***GitHub source repository link:*** *https://github.com/numpy/numpy*

SciPy

SciPy [24] is a free and open-source Python library used for scientific and technical computing. SciPy contains modules for optimization, linear algebra, integration, interpolation, special functions, FFT, signal and image processing, ODE solvers, and other tasks common in science and engineering. The SciPy package is currently distributed under the BSD license and its development is sponsored and supported by an open community of developers. The basic data structure used by SciPy is a multidimensional array provided by the NumPy package which we have already discussed in the previous section. NumPy also provides some functions for linear algebra, Fourier transforms and random number generation but in a generalized manner like the equivalent functions in SciPy. In this work, sparse matrices are used to store the data of future edges. This is later used to create the historic data to train the neural network discussed in section 5.1.1.

- ***GitHub source repository link::*** *https://github.com/scipy/scipy*

Gephi

We use the tool, Gephi [25], for the visualization of the semantic network in Optics. Gephi is a free and open-source data exploration and visualization software for all kinds of graphs and networks. This software package is written in Java on the Netbeans platform. The

plots of the semantic network (discussed in detail in chapter 3) developed in this thesis are shown in figure 3.8. These plots have been made using Gephi. This makes it easier to understand the connected research topics by quickly looking at the visualization.

- ***GitHub source repository link:*** *https://github.com/gephi/gephi*

Chapter 3

Creation of the semantic network

The main aim of the semantic network that we will discuss in detail in this chapter is to store the huge amount of research data in the field of Optics, in the form of a graph. This will give insights into the fields that have already been investigated together and will make us think about the potential connections which are unconnected as of now. The semantic network created in this chapter and the methodology followed is based on the idea of the semantic network of Quantum Physics [5] and the semantic network of Artificial Intelligence [6]. Figure 3.3 gives the overview of the steps followed to prepare and process the data in a snapshot. After implementing every block illustrated in figure 3.3, we get a summary of the entire set of research articles (published on arXiv [1], discussed in section 3.1.1) relevant to Optics in the form of a list of important words appearing in the article.

3.1 The dataset

3.1.1 Data source (arXiv)

The arXiv [1], introduced by Paul Ginsparg, is an online repository for self-archived, so-called e-prints of scientific papers covering different fields of physics, science, mathematics, and biology. arXiv began in the print-only era in 1991 [26]. Started at Los Alamos National Laboratory, and known as xxx.lanl.gov until 1998, it was intended to level the global research playing field by providing equal-time access to the latest research results [26]. This was before the World Wide Web, and publishers and librarians at the time were skeptical about any near-term transition to digital content [26]. In the early 1990s, arXiv played a pioneering

[1]Source: https://arxiv.org/

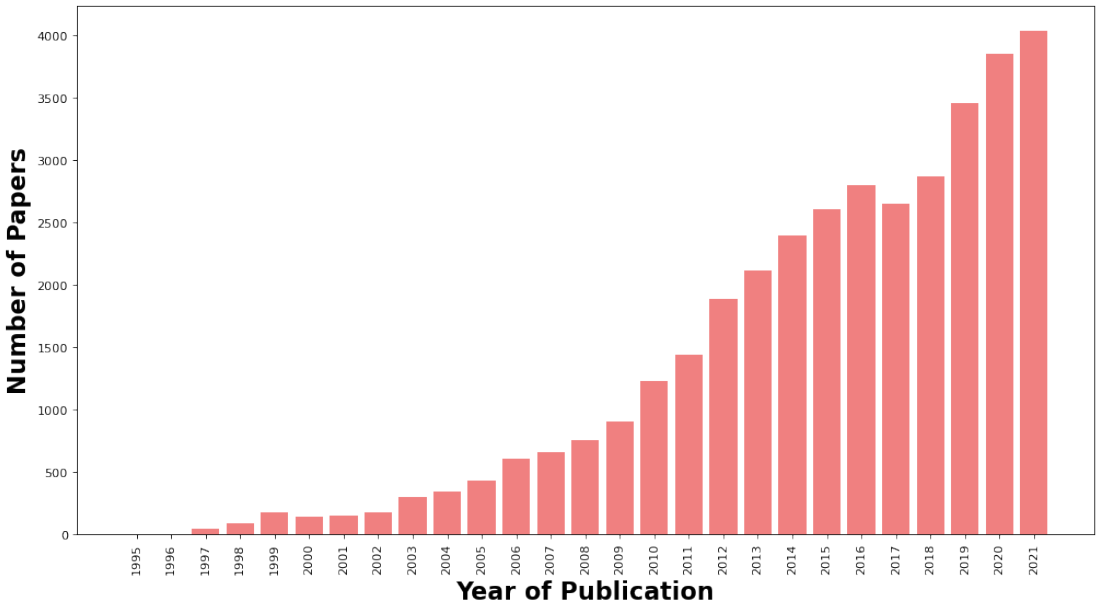


Figure 3.1: Growth in papers published under the category of Optics

role as an automated repository and was the first to use an abstract-landing web page for articles, with links to associated resources, including full-text postscript, and then pdf [26]. Authors upload their pre-prints (before peer review) or post-prints (after peer review) in source format (mostly TEX), which are automatically converted into postscript or PDF files. Since its foundation, it showcased the possibility of distributing scientific documents freely over the internet, which led to the current revolution in scientific publishing, known as the open access movement. As of April 16, 2022, arXiv consists of 2,051,413 papers. So it's not astonishing that it is the largest centralized Open Access archive available today. In arXiv, a system of endorsements of other authors leads intrinsically to better data quality. After analyzing data from the source, the graphical representation shown in figure 3.1 is obtained. This graph gives a clear vision of how rapidly the publications on arXiv relevant to the field of Optics are growing.

3.1.2 Source data structure

We use the data from Kaggle [2] (figure 3.2) which is a mirror of the original arXiv data because the full dataset is rather large (1.1TB and growing), this dataset provides only

[2]Source: https://www.kaggle.com/datasets/Cornell-University/arxiv

{'id': '1712.02892', 'submitter': 'Xuemei Gu', 'authors': 'Xuemei Gu, Mario Krenn, Manuel Erhard, Anton Zeilinger', 'title': 'G
ouy Phase Radial Mode Sorter for Light: Concepts and Experiments', 'comments': 'main text: 7 pages, 5 figures. Supplementary In
formation: 5 pages, 4\n figures', 'journal-ref': 'Phys. Rev. Lett. 120, 103601 (2018)', 'doi': '10.1103/PhysRevLett.120.10360
1', 'report-no': None, 'categories': 'quant-ph physics.optics', 'license': 'http://arxiv.org/licenses/nonexclusive-distrib/1.
0/', 'abstract': ' We present an in principle lossless sorter for radial modes of light, using\naccumulated Gouy phases. The e
xperimental setups have been found by a computer\nalgorithm, and can be intuitively understood in a geometric way. Together wit
h\nthe ability to sort angular-momentum modes, we now have access to the complete\n2-dimensional transverse plane of light. The
device can readily be used in\nmultiplexing classical information. On a quantum level, it is an analog of the\nStern-Gerlach ex
periment -- significant for the discussion of fundamental\nconcepts in quantum physics. As such, it can be applied in high-dime
nsional and\nmulti-photonic quantum experiments.\n', 'versions': [{'version': 'v1', 'created': 'Thu, 7 Dec 2017 23:51:53 GMT'},
{'version': 'v2', 'created': 'Wed, 29 Aug 2018 14:40:41 GMT'}], 'update_date': '2018-08-30', 'authors_parsed': [['Gu', 'Xueme
i', ''], ['Krenn', 'Mario', ''], ['Erhard', 'Manuel', ''], ['Zeilinger', 'Anton', '']]}

Figure 3.2: Paper data Example [27]

key data of papers in a file in the *json* format. This file contains an entry of each paper's information broken down into categories as given below.

- *id*: arXiv ID (can be used to access the paper)
- *submitter*: Who submitted the paper
- *authors*: Authors of the paper
- *title*: Title of the paper
- *comments*: Additional info, such as number of pages and figures
- *journal-ref*: Information about the journal the paper was published in
- *doi*: DOI of the paper
- *abstract*: The abstract of the paper
- *categories*: Categories / tags in the arXiv system
- *versions*:The version history

3.1.3 Data preprocessing

Good quality data is the heart of any machine learning problem. The data quality directly impacts the accuracy of the results that our neural network delivers. We have the main database from arXiv but this is not the final dataset that will be used to train the machine learning model. We have to do some data mining from the source data from arxiv. For the current problem statement, all the papers published under the category *physics.optics* are extracted from the source and processed for further use. This is a total of 35,717 papers.

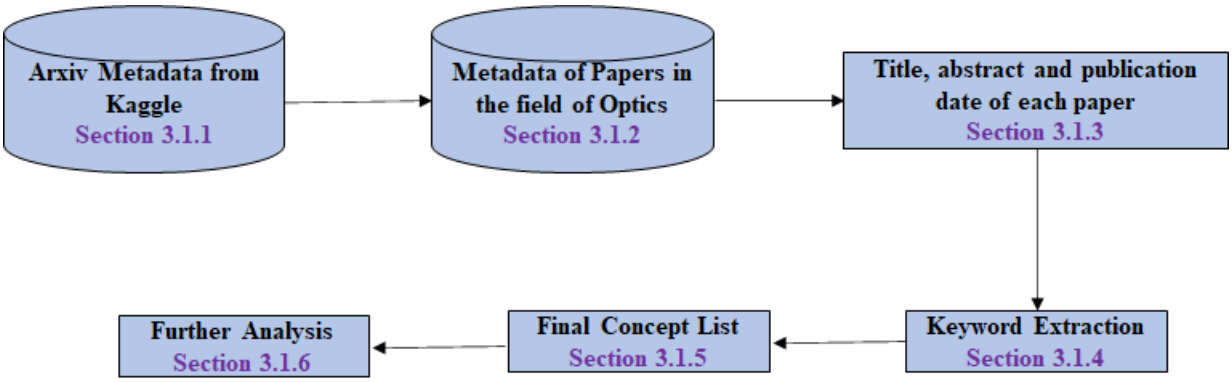


Figure 3.3: Data preprocessing workflow. This block diagram illustrates the steps and the sequence that is followed to get the research articles, process them and generate a list of important words that highlight the topic of the article.

The main concern now is transforming the data in such a way that the machine can process it efficiently. We are now interested in the titles and abstracts of all the papers because these sections give us an overview of what the paper is about and which areas relevant to Optics have been investigated in that particular publication. The data is in the form of raw text as the titles and abstracts of all the papers, which is human-written text and thus, can't be used directly for computing insights. The data that is relevant for us is only some main/ technical words that are used in the field of Optics and the rest of the text can be discarded. This is because there are some words in every piece of text that can throw light on what has been discussed in the text without actually having to read it as sentences.

3.1.4 Keyword extraction

Keyword extraction is a text analysis technique to summarize the content of texts and recognize the main topics discussed. Keyword extraction breaks down human language so that it can be processed and analyzed by machines. This is one of the crucial steps as the result of this step will give us an idea of how our dataset will look like.

Why is keyword extraction important?

Keyword extraction helps to find the most important words and phrases within massive sets of data (like new articles, papers, or journals) without having to read the entire content.

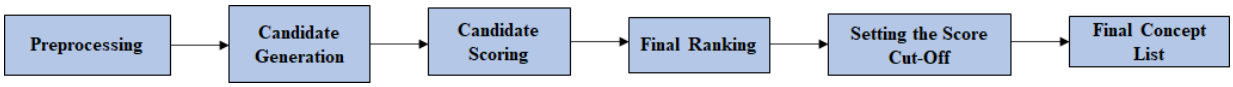


Figure 3.4: Workflow of the RAKE Algorithm

This helps to automatically index data, summarize a text, or generate tag clouds with the most representative keywords.

Workflow of the keyword extraction algorithm

The algorithm used for extracting keywords is Rapid Automatic Keyword Extraction (RAKE) [22]. One of the critical points made by the creator of RAKE is that keywords frequently contain multiple words but rarely contain punctuation, stop words, or other words with minimum lexical meaning [22]. Once we have the text corpus, RAKE splits the text into a list of words, removing stop words from the same list. The return list is known as Content Words. Ignoring them will make our main corpus lean and clean [22]. This can be explained better with the help of an example. Suppose, we have a piece of text as given below that needs to be processed to extract keywords.

"Deep Learning is a subfield of AI. It is very useful."

- **Step 1: Preprocessing and candidate generation**
  The very first step is to split the text into a list of words and remove stop words from that list. Stop words are words that do not add much meaning to a sentence. Ignoring these words does not make any difference to the meaning of the sentence. This step returns a list of what is known as content words [22].

  Suppose our list of stopwords and phrase delimiters look like these:
  stopwords = [is, a, of, it, very]
  delimiters = [.]

  After the stopwords and delimiters are removed, the leftover words are the 'Candidates'.
  Candidates = [deep, learning, subfield, ai, useful]

  This means that every word from the list above is a potential keyword from the text we processed.

- **Step 2: Candidate scoring**
  In this step, we will score the candidates and then choose the ones with the highest scores. Let us now try to understand what scoring means here.

Words	*deep*	*learning*	*subfield*	*ai*	*useful*
Word Frequency: freq(w)	1	1	1	1	1

Table 3.1: Word frequency. The number corresponding to every word in this table indicates the number of times that particular word appears in the piece of text that is being processed.

Word	*deep*	*learning*	*subfield*	*ai*	*useful*
deep	1	1	0	0	0
learning	1	1	0	0	0
subfield	0	0	1	0	0
ai	0	0	0	1	0
useful	0	0	0	0	1
Degree: deg (w)	***1 + 1 = 2***	***1 + 1 = 2***	***1***	***1***	***1***

Table 3.2: Word co-occurrence count. The co-occurrence count is 1 if we pair every word with itself. When a word is paired with other words in a sentence and if these words appear next to each other, the count for this pair is 1. In this case, the words, 'deep' and 'learning' appear one after the other in the text and thus, both of these have a co-occurrence count of 1 when paired with each other. This is important to extract 'deep learning' as one keyword and not separately as two keywords, 'deep' and 'learning'.

1. The first step to begin with candidate scoring is determining the word frequency [22]. This is the number of times that word appears in the text we are processing. This data can be seen in table 3.1.
2. The next step is to get the word co-occurrence count and the degree for each word which is the total sum [22]. This metric identifies words that occur often in longer candidate keywords. This data can be seen in table 3.2.
3. Next, we divide the degree by the frequency for each word to get a final score [22]. This score identifies words that occur more in longer candidate keywords than individually. The score for each candidate is shown in table 3.3.

Word	*deep*	*learning*	*subfield*	*ai*	*useful*
Score = deg (w)/freq(w)	2/1 = 2	2/1 = 2	1/1 = 1	1/1 = 1	1/1 = 1

Table 3.3: Word scores. This number gives an overview of every word's appearance with other words in the text. This gives an idea of the word quality. A higher number here means that the particular word appears with many other words and is a word of high importance.

Keyword	Score	Remarks
deep learning	4	score (deep) + score (learning) = 2 + 2 = 4
subfield	1	score (subfield) = 1
ai	1	score (ai) = 1
useful	1	score (useful) = 1

Table 3.4: Scores of candidate keywords. This number indicates the quality of the whole keyword. The higher the score, the more important is the keyword to help throw light on the gist of the text.

```
['dual frequency vertical external cavity surface emitting laser', 'scattering type scanning near field optical microscopy',
 'periodically poled potassium titanyl phosphate crystal', 'dielectric loaded surface plasmon polariton waveguide',
 'laser interferometer gravitational wave observatory', 'synchronously pumped optical parametric oscillator',
 'continuous variable quantum information processing', 'phase sensitive optical time domain reflectometry',
 'atomically thin transition metal dichalcogenides', 'high resolution transmission electron microscopy',
 'incoherent point source', 'dispersion compensation', 'small optical thickness', 'far field approximation',
 'plasmonic nanoparticles', 'nanoimprint lithography', 'polarization dependence', 'high numerical aperture',
 'coherent frequency comb', 'ponderomotive potential']
```

Figure 3.5: Concept list snapshot. These are some randomly chosen concepts from the final concept list to give a brief idea of how the list looks like.

- **Step 3: Final ranking**
  Now, from the processing done in the previous steps, we already have the scores for all the words. In this step, we rank the words based on their scores. This data can be seen in table 3.4 in which the keywords are sorted in the descending order of their score value.

3.1.5 Final concept list

So, now we know how the RAKE [22] algorithm works. The text we pass in our case is the title and abstract of each paper one by one. Then all the 3 steps mentioned in the previous section are implemented to get the keywords ranked based on their score computed from the text analysis. It is up to us how many first 'n' number of words we would like to pick from this by setting a cut-off value to discard the rest and get only good quality final words. These keywords finally form our concept list. figure 3.5 gives some idea of how the keywords look in our case. In our case, once we have the list of concepts, we sort these in the descending order of their length as a string, meaning based on the number of characters the concept has. Thus, the initial concepts have more letters than the last ones.

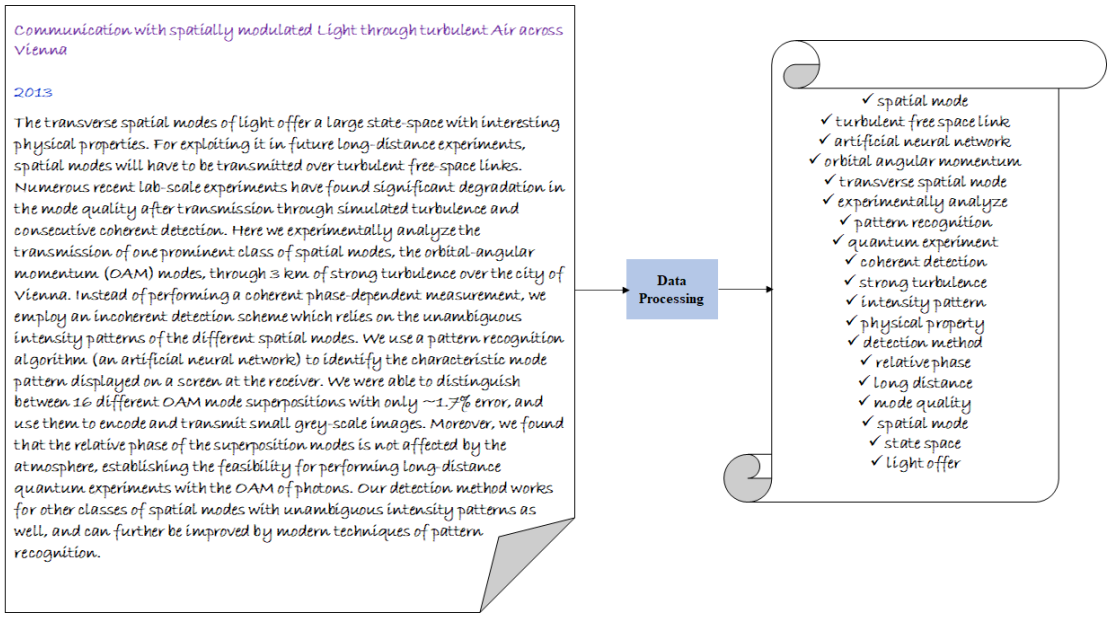


Figure 3.6: Concept list extraction example (The paper title and abstract shown in this image is from [28])

3.1.6 Further analysis

The creation of the concept list requires an ample amount of time and vigilance to get good-quality words that give you an overview of the field, in this case, Optics. Sometimes, there still might be some generalized words that do not necessarily throw light on a research field in Optics and can be used in any context. Knowledge of a specific field plays an important role here to decide, which is a good concept and which is not. Let us take an example of the concept, *light source*. Now, when it comes to the field of Optics, this is a very generic term and it is fine to not treat this as a concept. Some other words are the ones that have the words *high*, *low*, or *average* in them. The concepts like *high power*, *low power*, or *average power* do not convey the necessary information relevant to Optics. However, there are not many cases like this if we compare with the whole number of concepts as the keyword extraction algorithm discussed earlier efficiently does its job.

3.2 Creation of the semantic network

We had an overview of what a semantic network is in Section 2.2.1. Here, we will discuss in detail the underlying logic used to develop the semantic network of Optics. The various building blocks for such networks and their analogy to our work will also be presented.

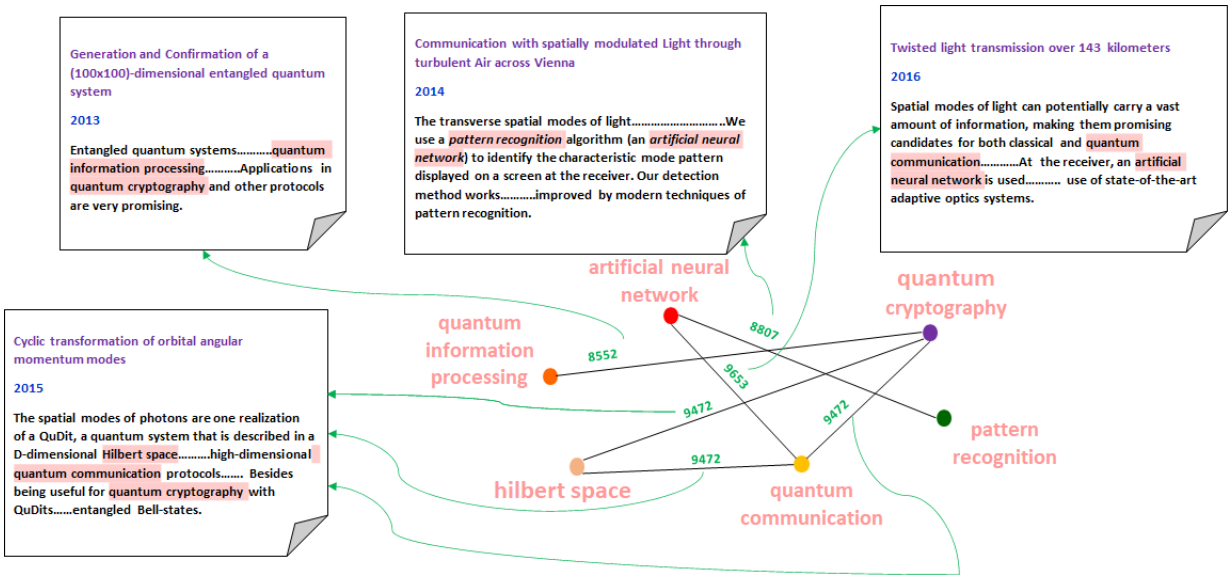


Figure 3.7: Logic behind the semantic network creation (image from [28], [29], [30], [31])

3.2.1 Nodes of the semantic network of Optics

We had an overview of what a graph node means in general in section 2.2.1. In our case, each node is equal to each concept in our concept list. We have 15,750 concepts that we have extracted from the research papers in the field of Optics from arXiv. Thus, we have 15,750 nodes in our semantic graph. Using these words directly will make it harder to process this data further. To avoid further problems, we assign a unique id to every concept. This creates a dictionary of all the concepts with their unique ids. The first concept in the list has id '0' and the last one has '15,749'. Thus, while dealing with these concepts further, we will use their unique id instead of the concept in words, directly. That means each node in the semantic network has one of the numbers from 0 to 15,749.

3.2.2 Edges of the semantic network of Optics

We had an overview of what a graph node means in general in section 2.2.1. In the case of our semantic network, we define the edge as the co-occurrence of the two concepts in the same paper. Suppose, we have two concepts under consideration, concept A and concept B which appeared together in the title and abstract of a research paper in 2014, we draw an edge connecting the nodes assigned to these concepts. This can be better understood with the illustration shown in figure 3.7. In addition to this, we also add a label to every edge which is a time stamp, that is, the publication date. We add the time stamp as a

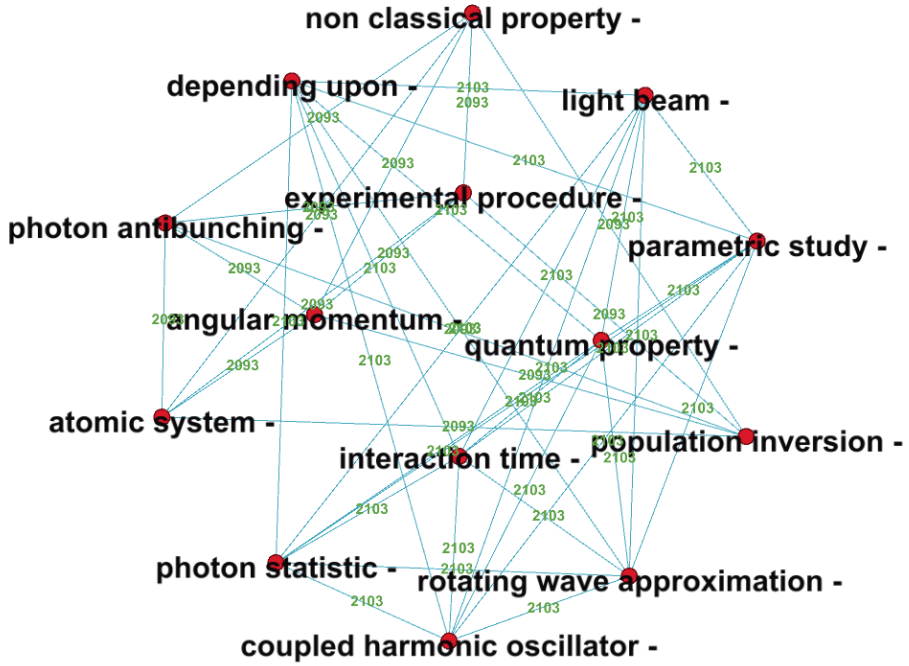
	Year 1995	Year 2021
Number of Nodes	14	353
Number of Edges	43	190

Table 3.5: Node and edge number for the networks shown in figure 3.8

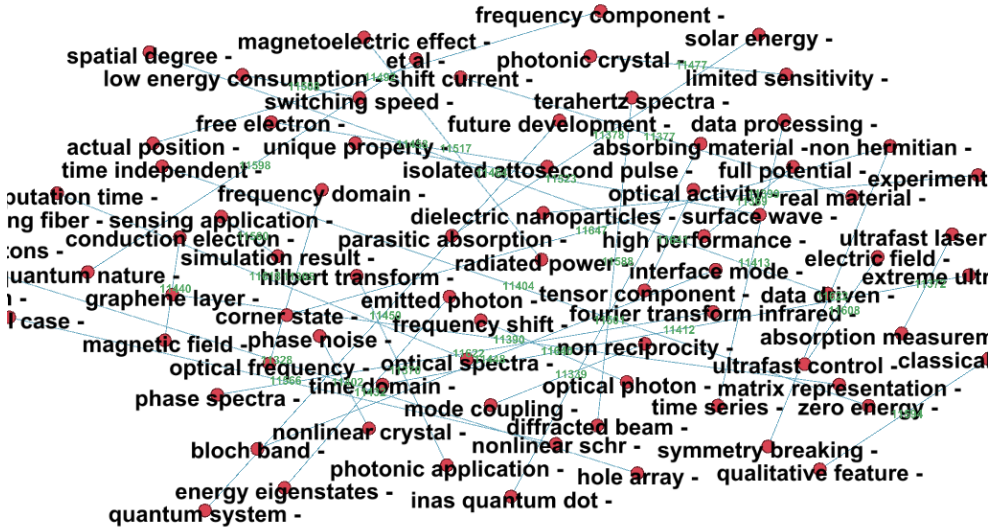
day count from January 1, 1990, to have a uniform baseline. So, we count the number of days from our baseline to the publication date when the concepts appeared together. As an assumption, if the exact date of our dummy publication is November 12, 2014, our time stamp will be 9081. The final semantic network has 2,907,489 edges.

3.2.3 An attempt at synonym detection

We use cosine similarity as the measurement to decide if the nodes are similar or not. The methodology of cosine similarity has been discussed in section 2.2.2. We calculate the values for every pair of nodes in the semantic network. Based on the values, we decide if one of the nodes can be removed or merged with other nodes or not. However, in some cases, the pairs recognized as synonyms were antonyms. One of the most common recurring examples is of the concepts, *linear optics* and *nonlinear optics*. The context in which two concepts are used or the other concepts with which these two concepts appear might be common to a greater extent but these two fields are not the same. These are antonyms because these are the two different branches of Optics. Also in some cases, the concept pairs detected as synonyms got detected because these two concepts were used together multiple times by the same author. This highlights a key issue that research graph analytics is affected by the way different researchers use different words to refer to a common scientific term. So, we decided to not remove nodes based on the conclusions from cosine similarity Calculations. This in turn, also prevents us from distorting the graph. This analysis also helps us realize the importance of the co-occurrence of concepts. After all, we are not just interested in how two nodes are connected but also in how these two nodes are connected to other nodes in the network. If we merge some nodes based on cosine similarity and form clusters, we can end up losing information about individual connections of every concept in the cluster. At this stage, cluster analysis can make computations complicated and time-consuming so this can be omitted at the cost of no loss in further analysis. So, even though we don't use cosine similarity to filter nodes, this is a very helpful filter to have an optimized list



(a) Connected concepts in the year 1995. Here, the red circles denote nodes and the respective concepts. The blue lines represent the edges. The numbers in green are the labels which are the timestamps. The timestamps are calculated as the total number of days from a uniform baseline date until the publication date of the article in which the two concepts connected by the edge appeared together.



(b) 0.01% of the connected concepts in the year 2021

Figure 3.8: Growth in the co-occurrence of concepts from 1995 to 2021

of predictions in the later stage. This is also one of the important factors to develop the predictive model which is discussed in detail in Chapter 5.

3.2.4 Visualizing the semantic network

After all the data processing and connection generation has been accomplished, the next step is to visualize this data. Since this is a huge graph, we have used the tool Gephi [25], discussed in section 2.4, to plot the giant web of Optics. We plot the network such that it is force directed. The tool Gephi simulates a physical system to spatialize a network. Nodes repel each other like charged particles, while edges attract their nodes, like springs. These forces create a movement that converges to a balanced state. This final configuration is expected to help the interpretation of the data. We have used the ForceAtlas2 package [32] of Gephi for this purpose. The force-directed drawing places each node depending on the other nodes. This process depends only on the connections between nodes.

The plot of the entire semantic network is difficult to visualize on a piece of paper. So, the visualizations for the evolution are shown in Figure 3.8 that compares the semantic network of the concepts used in the year 1995 and just 0.01% of the connections of the concepts used in 2021. This depicts the enormous growth in the way concepts are investigated. Some key information about these networks is shown in table 3.5.

Chapter 4

The evolution of concepts in Optics

4.1 Data analysis of popular concepts

To predict future research trends, it is very essential to have a look at the trends in the past as well. This can play a crucial role in developing new methods to have insights into what is coming ahead. To have the idea of these emerging concepts, data is analyzed over different periods to investigate which concepts from the concept list appeared in a paper on arXiv, published under the category, 'physics.optics' for the first time. The next thing to be understood is which of these concepts remained popular in the next few years as well. The data is analyzed over 5 years, 3 years, and 1 year. To understand this better, let us pick up a period, say, 5 years. Suppose, the start year is set to 2016, the concept that first time appeared in a paper published in 2016 and has the most number of papers collectively in the years from 2016 to 2020 needs to be identified. The same logic applies to other periods and starting years. The concepts that turned out to be investigated frequently over the different periods are discussed in the next sections.

4.1.1 Concepts evolved over 5 years

This section discusses the fields that were most popular among scientists over five years by appearing in the most number of the papers. This trend can be visualized with the illustration shown in figure 4.1. It would be good to highlight here that the concepts like *electromagnetic field* and *atomic system* go back to the early 19^{th} and 20^{th} centuries respectively. However, these turn out be the rapidly emerging concepts for the years 1995 and 1996 respectively. The reason why this happens here is we have used only arXiv as

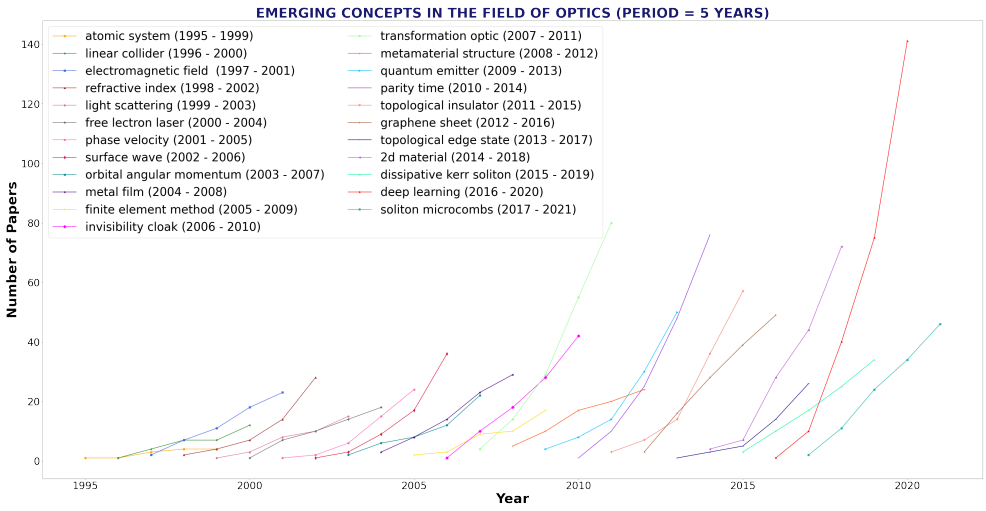


Figure 4.1: Plot of the evolving concepts [period = 5 Years]

the source of the research papers to perform this analysis and other tasks discussed in this work. Thus, the emerging concepts are based only on arXiv data and don't indicate the actual discovery period of a particular field.

Atomic system (1995 - 1999)

An atom is the smallest particle into which an element can be divided without losing its chemical identity [1]. This concept first appeared in the paper titled, 'Squeezing in the interaction of radiation with two-level atoms' [33] in 1995 on arXiv. The method proposed in this paper decreases the uncertainties of the angular-momentum quadratures representing the two-level atomic system in the interaction of the two-level atoms with quantized radiation [33].

Linear collider (1996 - 2000)

A collider is a type of particle accelerator that brings two opposing particle beams together such that the particles collide [2]. This concept first appeared in the paper titled, 'Laser cooling of electron beams for linear colliders' [34] in 1996 on arXiv. In this paper, a novel method of electron beam cooling is considered which can be used for linear colliders [34].

[1]Source: https://atomic.lindahall.org/what-is-an-atom.html
[2]Source: https://news.fnal.gov/2013/08/fixed-target-vs-collider/

Electromagnetic field (1997 - 2001)

An electromagnetic field is the physical field extending throughout space that delivers electric and magnetic effects [35]. Electromagnetic fields obey the quantum version of Maxwell's equations [35] and most of the optical effects are explained based on this. This concept first appeared in the paper titled, 'A 1D Model for N-Level Atoms Coupled to an EM Field' [36] in 1997 on arXiv. A model for n-level atoms coupled to quantized electromagnetic fields in a rod-like geometry of diameter ranging from 10-100 nanometer was proposed in this paper [36].

Refractive index (1998 - 2002)

The refractive index of an optical medium is a dimensionless number that indicates the light-bending ability of that medium. This concept first appeared in the paper titled, 'A simple method for the determination of slowly varying refractive index profiles from in situ spectrophotometric measurements' [37] in 1998 on arXiv. Here, the refractive index of different optical films were calculated using several methods. [37].

Light scattering (1999 - 2003)

Light scattering is a phenomenon that occurs when light changes its direction after hitting a small particle causing optical phenomena such as the blue color of the sky, and halos. This concept first appeared in the paper titled, 'Differential light scattering: probing the sonoluminescence collapse' [38] in 1999 on arXiv. Here, a light scattering technique was proposed that is capable of information retrieval without the need of fast electronic equipment [38].

Free electron laser (2000 - 2004)

A free electron laser is a light source producing extremely brilliant and short pulses of radiation. This concept first appeared in the paper titled, 'Photon collider at TESLA' [39] in 2000 on arXiv where the status of a photon collider based at TESLA was discussed. The key element in photon colliders is a very powerful laser system such as in a free electron laser and thus, this laser system was one of the considered approaches for the TESLA project [39].

Phase velocity (2001 - 2005)

Phase velocity is the speed at which a point of constant phase travels as the wave propagates [40]. This concept first appeared in the paper titled, 'Light Propagation For Accelerated Observers' [41] in 2001 on arXiv. In this study, it is shown that for a moving observer, a linearly polarized plane wave has two modes of propagation in a stationary, homogeneous and isotropic medium according to Hertz's version of Maxwell's theory. Of the two modes, the second mode has a phase velocity that is controlled by the motion of the observer and some applications of this second mode in emerging technologies are outlined in this paper [41].

Surface wave (2002 - 2006)

A surface wave is a mechanical wave that propagates along the interface between differing media. This concept first appeared in the paper titled, 'Photonic Approach to Making a Left-Handed Material' [42] in 2002 on arXiv. It was noticed during this study that the surface waves localized at the dielectric interfaces can be either surface plasmons or phonons [42].

Orbital angular momentum (2003 - 2007)

The orbital angular momentum of light is the component of the angular momentum of a light beam that is dependent on the field spatial distribution. This concept first appeared in the paper titled, 'Molecular chirality and the orbital angular momentum of light' [43] in 2003 on arXiv. Optical beams with different types of helicity have orbital angular momentum [43]. Here, the wave-front surface of the electromagnetic fields assume helical form. This study assesses what new features, if any, can be expected when such beams are used to interrogate a chiral system.

Metal film (2004 - 2008)

A metal film is a layer of metallic material ranging from fractions of a nanometer (monolayer) to several micrometers in thickness. This concept first appeared in the paper titled, 'Fano-type interpretation of red shifts and red tails in hole array transmission spectra' [44] in 2004 on arXiv. In this paper, an opinion is presented to understand the spectral features reported in the past experiments performed on holes in metal films [44].

Finite element method (2005 - 2009)

The finite element method is used for numerically solving differential equations arising in engineering and mathematical modeling. This concept first appeared in the paper titled, 'Numerical Investigation of Light Scattering off Split-Ring Resonators' [45] in 2005 on arXiv. In this paper, numerical solutions to the time-harmonic Maxwell's equations by using advanced finite-element methods (FEM) have been presented [45].

Invisibility cloak (2006 - 2010)

Cloaking in general refers to hiding objects from the eye, and in particular, the radar [46]. This concept first appeared in the paper titled, 'Calculation of material properties and ray tracing in transformation media' [47] in 2006 on arXiv. The method to calculate the material properties associated with a coordinate transformation (changing the coordinate system) has been shown here and demonstrated using spherical and cylindrical shaped invisibility cloaks by performing ray tracing on them [47].

Transformation optic (2007 - 2011)

Transformation optics is a branch of Optics that uses coordinate transformations. This concept first appeared in the paper titled, 'Design of Electromagnetic Cloaks and Concentrators Using Form-Invariant Coordinate Transformations of Maxwell's Equations' [48] in 2007 on arXiv. The material design of a square-shaped cloak and an electromagnetic field concentrator were presented here by using coordinate transformation (changing the coordinate system). [48].

Metamaterial structure (2008 - 2012)

Metamaterials are artificially engineered materials designed to induce customized properties in a material that originally does not exist [49]. This concept first appeared in the paper titled, 'Optical Activity of Planar Achiral Metamaterials' [50] in 2008 on arXiv. Optical activity and circular dichroism are linked to chirality (helicity) of organic molecules, proteins and inorganic structures, can also be observed in non-chiral artificial media [50]. The metamaterial structure used for this study yields a strong resonant optical activity and thus, has been used here to report this classical phenomena [50].

Quantum emitter (2009 - 2013)

A quantum emitter is generally defined as a quantum system that is capable of radiative optical transitions [3]. This concept first appeared in the paper titled, 'Orientation-dependent spontaneous emission rates of a two-level quantum emitter in any nanophotonic environment' [51] in 2009 on arXiv. In this paper, a theoretical study of the spontaneous emission rate of a two-level quantum emitter in any nanophotonic system has been presented.

Parity time (2010 - 2014)

A parity-time (PT) symmetric system is a special non-Hermitian system of which its Hamiltonian possesses real eigenvalues [52]. This concept first appeared in the paper titled, 'Nonlinear suppression of time-reversals in PT-symmetric optical couplers' [53] in 2010 on arXiv. The effect of nonlinearity-induced PT-symmetry breaking is described analytically and demonstrated numerically in this paper [53].

Topological insulator (2011 - 2015)

A topological insulator is a material whose interior behaves like an electrical insulator while its surface behaves like an electrical conductor. This concept first appeared in the paper titled, 'Spatially resolved femtosecond pump-probe study of topological insulator Bi2Se3' [54] in 2011 on arXiv. The phonon dynamics in the topological insulator, Bi2Se3 crystals are studied in this paper [54].

Graphene sheet (2012 - 2016)

Graphene sheets comprise carbon atoms attached in hexagonal shapes and every carbon molecule covalently sticks to three other carbon atoms [4]. This concept first appeared in the paper titled, 'Superradiance mediated by Graphene Surface Plasmons' [55] in 2012 on arXiv. In this paper, it is demonstrated that the interaction between two emitters can be controlled using the efficient excitation of modes in graphene surface plasmons supported by two-dimensional graphene sheets [55].

[3] Source: https://www.physik.hu-berlin.de/de/nano/forschung/quantum$_e$*mitters*

[4] Source: https://nanografi.com/popular-products/graphene-sheet-size-10-cm-x-10-cm-thickness-35-m-highly-conductive/

Topological edge state (2013 - 2017)

An edge state consists of pairs of states which have opposite spins and propagate in opposite directions [56]. This concept first appeared in the paper titled, 'Imaging topological edge states in silicon photonics' [57] in 2013 on arXiv. The topological edge states of light in a two-dimensional system were noticed during this research [57].

2d material (2014 - 2018)

If only one of the three dimensions of a material is nano-sized, it would be a 2d material, resembling a large, but very thin sheet like a piece of paper [5]. This concept first appeared in the paper titled, 'Single-molecule study for a graphene-based nano-position sensor' [58] in 2014 on arXiv. Here, several experiments were performed with dibenzoterrylene (DBT) molecules to show a genuine manifestation of a dipole interacting with a 2D material.

Dissipative kerr soliton (2015 - 2019)

Dissipative Kerr solitons are self-organized optical waves arising from the interplay between the Kerr effect and dispersion [59]. This concept first appeared in the paper titled, 'Dissipative Kerr solitons in optical microresonators' [60] in 2015 on arXiv. This paper describes the discovery and stable generation of temporal dissipative Kerr solitons in continuous-wave (CW) laser-driven optical microresonators [60].

Deep learning (2016 - 2020)

Deep learning is a machine learning technique that teaches computers to do what comes naturally to humans, that is, learn by example. This concept first appeared in the paper titled, 'Deep Learning with Coherent Nanophotonic Circuits' [61] in 2016 on arXiv. Here, a new architecture for a neural network has been proposed that uses unique advantages of optics and promises an increase in the computational speed for conventional learning tasks [61].

Soliton microcombs (2017 - 2021)

Solitons are nonlinear waves that maintain their shape while propagating at a constant velocity [6]. Soliton microcombs are phase-locked microcavity frequency combs. This concept

[5]Source: https://www.ossila.com/en-eu/pages/introduction-2d-materials
[6]Source: https://www.tu-chemnitz.de/physik/KSND/abb/node6.html

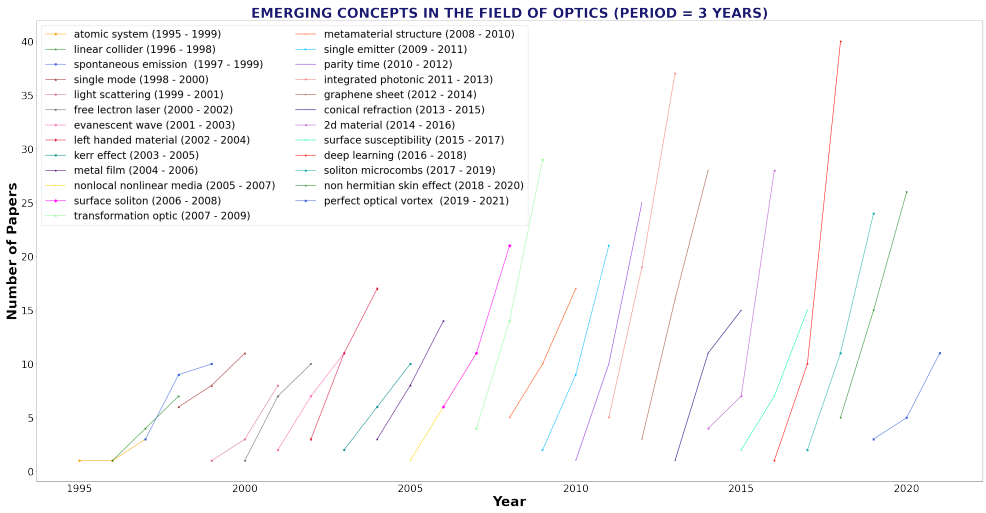


Figure 4.2: Plot of the evolving concepts [period = 3 Years]

first appeared in the paper titled, 'Towards Visible Soliton Microcomb Generation' [62] in 2017 on arXiv. Soliton microcombs at 778 nm and 1064 nm using on-chip high-Q silica resonators have been demonstrated here [62].

4.1.2 Concepts evolved over 3 years

It is not necessary that a concept popular for five years is popular for three years as well. Figure 4.2 shows the concepts that remained popular over a three year period. Some of the concepts like *graphene sheet*, *deep learning* remained popular over both the periods under analysis.

4.1.3 Popular concept over 1 year

As mentioned at the beginning of the previous section, the popularity of concepts can differ over different periods. The visualization of popular concepts for every year can be visualized in the graph depiction in figure 4.3.

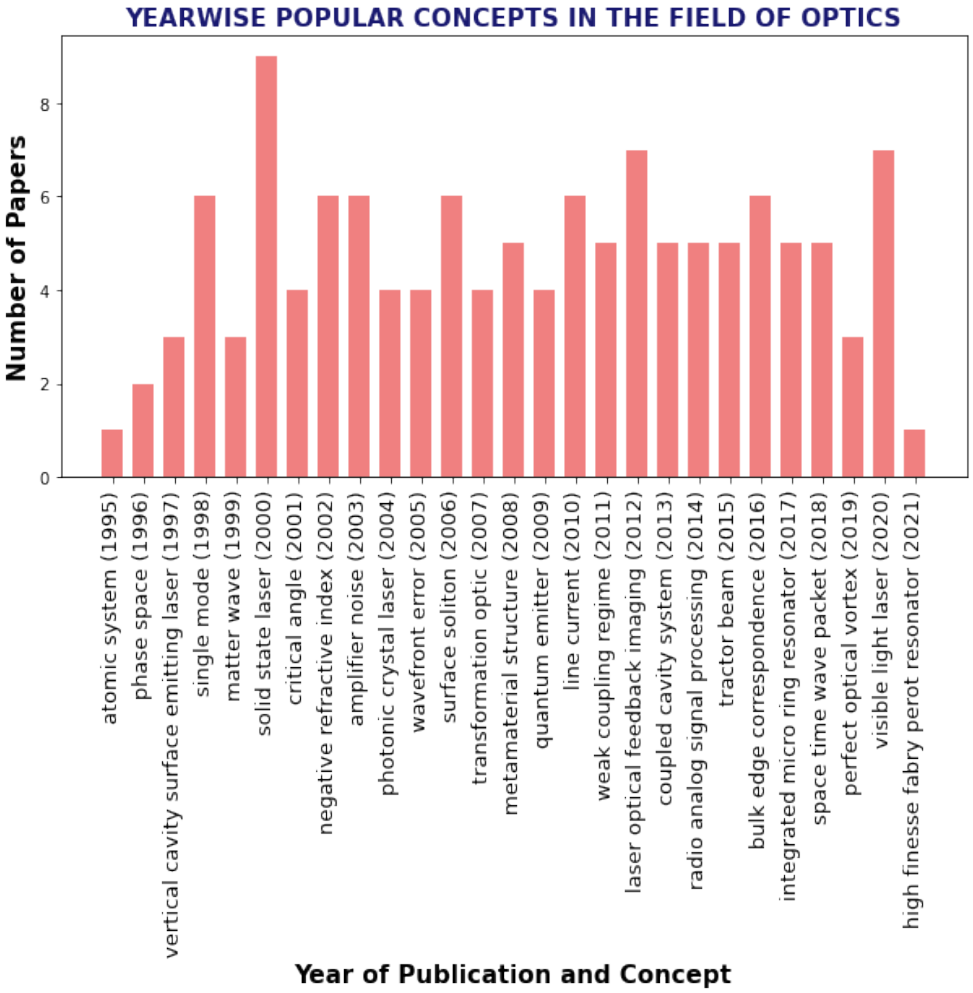


Figure 4.3: Plot of the popular concepts [period = 1 Year]

Chapter 5

Prediction of future research trends

In the last chapters, we had an overview of what the original data looks like and how we create a semantic network out of it by multiple data processing methods. This is the final dataset that we are going to use for the procedure henceforth to predict the future research trends in Optics by using an artificial neural network. This chapter will give an overview of the architecture of the neural network, various parameters, and the algorithm followed.

5.1 The model

5.1.1 The architecture

To solve the prediction task, we employ a feedforward neural network as described in section 2.3.2. The model used has one input layer, three hidden layers, and one output layer. The input layer has 15 neurons. These 15 neurons are nothing but the 15 extracted features from the semantic network we created earlier. The logic behind this feature extraction is discussed on pages, 49 and 50. The consecutive hidden layers after this have 100, 100, and 10 neurons each. The last one is the output layer with 1 neuron. Between every layer, there is an activation function layer. The activation function used here is 'Rectified Linear Unit (ReLU)' which we discussed in section 2.3.3. During the training process discussed in section 5.1.2, the model is continuously optimized using Gradient Descent and Backpropagation (discussed in section 2.3.3) to minimize the loss as the data is passed from one layer to another. The loss function used is 'Mean Squared Error' (please refer to section 2.3.3), which helps us determine the quality of the model. The layers of the model used here are

```
ff_network(
  (semnet): Sequential(
    (0): Linear(in_features=15, out_features=100, bias=True)
    (1): ReLU()
    (2): Linear(in_features=100, out_features=100, bias=True)
    (3): ReLU()
    (4): Linear(in_features=100, out_features=10, bias=True)
    (5): ReLU()
    (6): Linear(in_features=10, out_features=1, bias=True)
  )
)
```

Figure 5.1: Neural network layers

summarised in figure 5.1 and figure 5.2 shows a generic structure of the feedforward neural network.

5.1.2 Link prediction model workflow

In this section, the procedure to determine new connections after the specified number of years will be discussed. Before we proceed further, let us have an overview of the problem our trained neural network needs to solve. In Chapter 3, we discussed the process we followed to develop the semantic network. From this network, we already know what is connected. However, there are still many unconnected pairs, and finding this out can be fast and efficient using a neural network. This is the exact problem our model needs to tackle. The model needs to find out the most probable connections between 2 nodes in the future three years which are not been connected until now.

Initially, we train the model to get the predictions related to the past research work from the past data. So, let us understand this with the help of an example that has been implemented in this thesis. In this thesis, we are interested in the dynamics of the semantic network in 2024. For faster computations, we consider the minimal vertex degree to be used in the predictions as 5 and the minimal edges in the range from 0 to 3 that have a chance to form. We have already prepared a semantic network with the help of research data published on arXiv until 2021. Now, first, we have to build up a model such that it can predict for 2021 from the past data. We call this historic training data. The model is trained based on the properties of the nodes in the year 2013, 2014 and 2015. After this, the model first learns to predict for the year 2018. Once we have the trained model, it is very important to evaluate the efficiency of the model. In order to this, the evaluation data

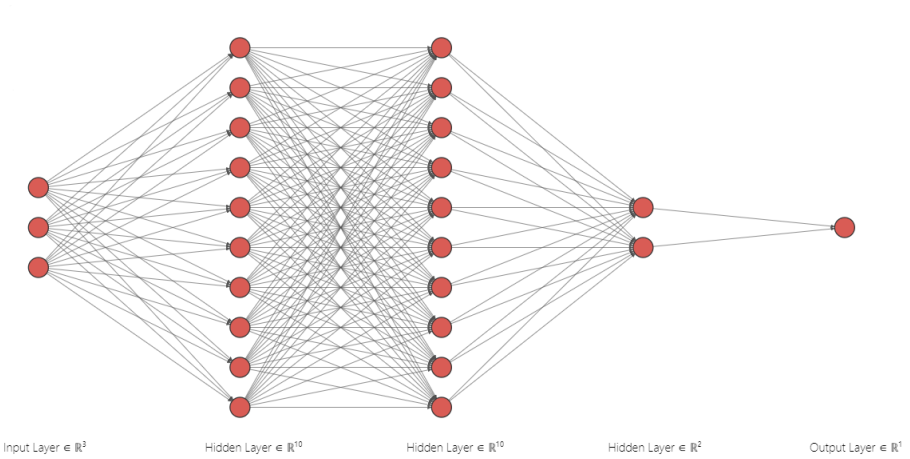
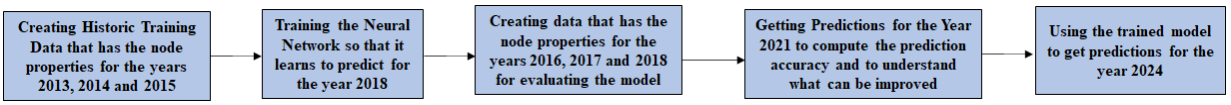


Figure 5.2: Sequence of the layers in the network used to build the link prediction model. For better visualization to give an overview of the interconnections, the number of neurons in each layer has been reduced. In the real network, the first, second, third and fourth layers have 15, 100, 100 and 10 neurons respectively. The fifth layer that where we receive the output as a prediction connection score (discussed in section 5.2.4), has one neuron. This plot is made using the NN SVG tool [63]

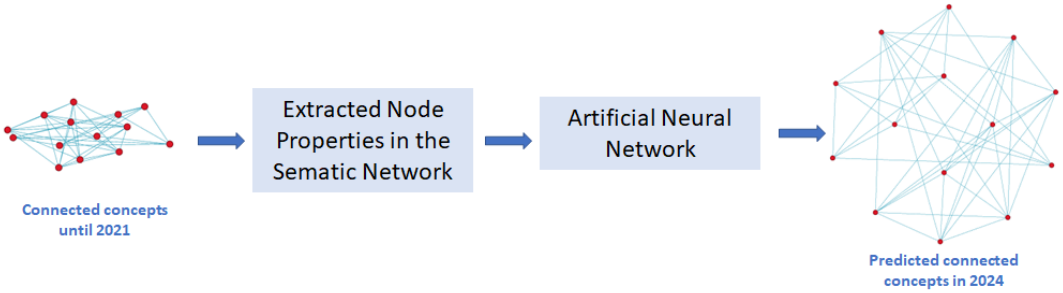
is prepared based on the properties of the nodes in the year 2016, 2017 and 2018. Then, we get the predictions for the year 2021. This is a very important step to tune our model to get the best out of it. It is already known to us what is going on in the field of research in 2021 as we already have all the required material at hand. This helps us to judge the model's accuracy. This explanation can be understood better with the help of figure 5.3(a). Another simple illustration shown in figure 5.3(b) summarizes the basic idea of the link prediction task.

Another very important part of training an artificial neural network is finding hidden patterns in our dataset and extracting features to get good-quality predictions. In this case, these features are generated based on the semantic network we created earlier. These features are nothing but the properties that describe the different nodes and edges of the network. In total, we have to generate fifteen features. These features are computed for every vertex pair (v1, v2). Let us now have a closer look at the fifteen features which are as given below.

- **First set (6 features):** The degree of both the vertices, v1 and v2. The degree of every concept represents the number of times it has appeared with other concepts and thus, there can also be multi edges between the same two concepts. This degree



(a) Methodology of extracting data for different time periods to train the model for link prediction



(b) Graphical illustration of the link prediction model. The red circles are the nodes assigned to every concept and the blue color lines are the edges to show the connected concepts.

Figure 5.3: Link prediction model workflow

is found for the current year and the previous two years. Thus, for every node, we have node degrees for three years each.

- **Second set (6 features):** The number of shared neighbours in total for v1 and v2 for the current year and previous two years. The number of shared neighbors of every concept represents the number of connections to the concept by a path of length 1, that is, there is one direct path connecting the two concepts.

- **Third set (3 features):** The last set of three properties is the total number of shared numbers between v1 and v2 in the current year and the previous two years.

In total, we have 15750 x 15750 edges as the total number of concepts we have is 15750 which equals to the number of nodes. This amounts to around 2.48 million edges and out of which, only around 10% are connected. That means, we have around 2.23 million unconnected edges. Computing the set of fifteen features is very time and memory consuming if it is done for the entire set of edges to train the model. Therefore, to make the computation faster and less expensive, we choose a random subset consisting of 10^7 edges. These fifteen features are then passed as input to the different layers discussed in section 5.1.1 and illustrated in figure 5.1. So, 99% of the unconnected edges are not even considered in training. There are more cases of connected edges in the training dataset. The benefit of doing this is observed in the feature extraction process. Since we have more connected edges in the dataset, we can efficiently compute all the possible properties that throw light on the

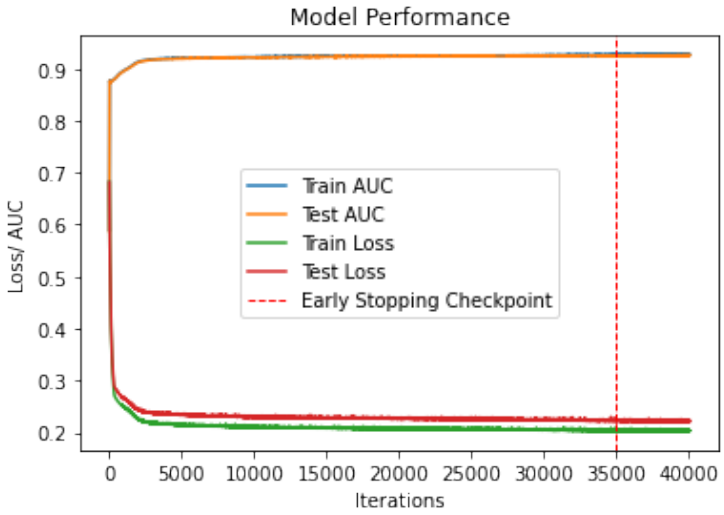


Figure 5.4: Early stopping checkpoint

fact that why certain nodes are connected. This helps us to predict the connections in the future. After all, we are more interested to see the possible edges instead of the impossible ones. Thus, we significantly improve the precision and spend less time on computation. The model is trained for 100,000 iterations with early stopping criteria. Setting this number up is decided based on the quality of predictions after every training round. It was noticed that when this number is smaller, the top predicted interesting combinations were nothing but the most common keywords in the field of Optics. Examples of such common concepts are 'linear optics' and 'nonlinear optics'. The same problem occurs when we train the model for a high number of iterations. We start getting the error due to the common problem of generalization because the model overfits as these two errors are closely related. Overfitting occurs when the learned function becomes sensitive to the noise in the sample. As a result, the function will perform well on the training set but not perform well on other data. Thus, the more the overfitting, the larger the generalization error [??]. During the training process, the model tries to chase the loss function crazily on the training data, by tuning the parameters. Now, we keep another set of data as the test set and as we go on training, we keep a record of the loss function on the test data, and when we see that there is no improvement on the test set, we stop. This strategy is called early stopping. Figure 5.4 shows the early stopping checkpoint for our model. The test loss average doesn't fluctuate much and remains in the same value range as earlier after this point and so the training is stopped here instead of using the complete number of set iterations to avoid overfitting and degradation of the prediction quality due to generalization.

	Actual Positive	Actual Negative
Predicted Positive	TP	FP
Predicted Negative	FN	TN

Table 5.1: The Confusion matrix parameters

5.2 Model performance

Before we proceed on discussing the performance of the model, let us have a brief overview of some important parameters that are required to measure the same.

5.2.1 The confusion matrix

The Confusion Matrix is a table that gives us an overview of the performance of our model. This matrix gives us an idea of how accurately the model interprets the data. Therefore, this matrix is also known as an Error Matrix. This is a contingency table with two dimensions, 'actual' and 'predicted' as illustrated in table 5.1. This table captures the True Positives, True Negatives, False Positives, and False Negatives which are explained below.

True Positives (TP)

This is the total number of correctly labeled positive samples [1]. Analogically, in our case, the number of connected vertex pairs that are predicted as connected and these are connected in reality.

True Negatives (TN)

This is the total number of correctly labeled negative samples [1]. Analogically, in our case, the number of unconnected vertex pairs that are predicted as unconnected and these are unconnected in reality.

[1]Source: https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62

False Positives (FP)

This is the total number of negative samples incorrectly labeled as positive [1]. Analogically, in our case, the number of unconnected vertex pairs that are predicted as connected and these are unconnected in reality.

False Negatives (FN)

This is the total number of positive samples incorrectly labeled as negative [1]. Analogically, in our case, the number of connected vertex pairs that are predicted as unconnected and these are connected in reality.

5.2.2 The confusion metrics

In the previous section, we had an overview of the concept of the Confusion Matrix. In this section, we will have a look at the inferences we can draw from this matrix that directly point toward the model performance.

True Positive Rate (TPR)

The true positive rate is calculated as the total number of true positives divided by the sum of the true positives and the false negatives [2]. The true positive rate is also known as the sensitivity or the recall [2].

$$TPR = \frac{TP}{TP + FN}$$

False Positive Rate (FPR)

The false positive rate is calculated as the total number of false positives divided by the sum of the false positives and true negatives [2].

$$FPR = \frac{FP}{FP + TN}$$

5.2.3 Area Under the Curve (AUC)

In machine learning, performance measurement is an essential task. AUC is one of the most important evaluation metrics for a binary classifier to check its performance as it

[2]Source: https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5

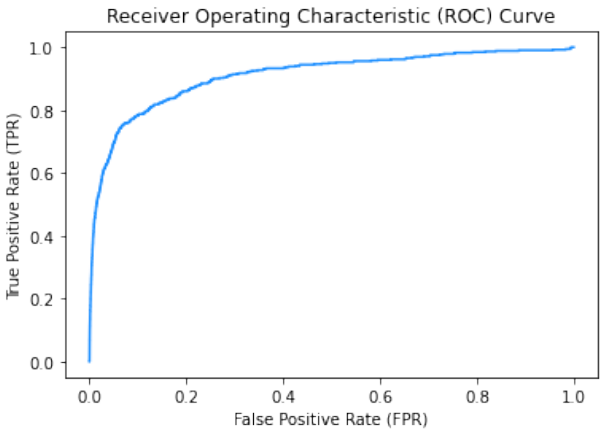


Figure 5.5: Receiver Operating Characteristic (ROC) Curve

represents the measure of separability between two classes. It is also written as AUROC (Area Under the Receiver Operating Characteristic). Receiver Operating Characteristic (ROC) is a probability of the chance that a random true element is ranked higher than a random false one. The ROC curve depicts the rate of true positives concerning the rate of false positives, therefore highlighting the sensitivity of the classifier model [64]. The higher the AUC, the better the model is at predicting. The ROC curve is plotted with TPR against the FPR where TPR is on the Y-axis and FPR is on the X-axis [64]. AUC measures the entire two-dimensional area underneath the entire ROC curve [64]. Thus, these metrics are computed based only on two outcomes, a positive or a negative prediction. The link prediction model workflow discussed in section 5.1.2 also is also trained only on two possible outcomes - connected or unconnected and this is why we use the AUC to identify the performance of our model. The ROC curve for the model in this thesis is shown in figure 5.5. The AUC value is 0.91 for our model which means that there is a 91% chance that the model ranks a random case of connected nodes higher than an unconnected case.

5.2.4 Results

The output of the trained model is pairs of vertices sorted based on the connection score coming out of the neural network. The 15 properties extracted for each unconnected concept pair (discussed in section 5.1.2) are used by the neural network to estimate which pair of concepts is likely to be connected within 3 years. The first ones in the list have high chances of connecting in the upcoming years and the last ones have the least chance. The prediction quality is identified by plotting a ROC curve and then measuring the AUC. We

discussed ROC and AUC in section 5.2.3. In the ROC plot, The Y - axis shows the True Positive Rate (TPR). This is the rate of concept pairs that have been correctly identified to be connected within 3 years. In the ROC plot, the X- axis shows the False Positive Rate (FPR). This is the rate of the concept pairs that have falsely been predicted to be connected. A perfect neural network would have a TPR of 1 and an FPR of 0. Thus, the AUC for a perfect neural network is 1. The AUC can be interpreted as the probability that the neural network will rank a randomly chosen true instance higher than a randomly chosen negative instance [3]. The computed AUC of the trained model is 0.91 which indicates that the neural network can learn to predict future research interests in the field of Optics based on historical information with good accuracy.

Table 5.2 summarizes the most and the least probable predicted future connection. The highest predicted connection is between the concepts *beat note* and *deep learning*. The concept, *beat note* has the degree of 649, which means it appears 649 times with other concepts in different research papers published on arXiv under the category of *physics.optics*. The number of neighbors of this concept in the semantic network is 497, which means, you can reach 497 concepts directly by path of length 1 starting from the node of this concept. The degree of the concept, *deep learning* is 2415 and it has 1332 neighbors. These concepts have a cosine similarity value of 0.092 which is very less indicating that these concepts do not have significant number of shared neighbors. The predicted connection score for these two concepts is 1.21 which is the highest among all the score values. This score value can be greater than or even less than 1 as this is the weighted sum received at the output neuron of the neural network after the input data undergoes multiple transformations in the various layers of the neural network. Higher is this number, higher is the chance that the two involved concepts will form a connection in the future. The highest predicted connection is between the concepts *test setup* and *cold cesium atom*. The two concepts have the degrees, 64 and 86 respectively. The number of neighbors are 58 and 73 respectively. The value of cosine similarity is 0. The predicted connection score value for this pair is -0.15 which is the lowest of the scores of all the pairs.

The important network data of the predicted connections has been summarized in the tables 5.3, 5.4 and 5.5. C_1 stands for one node in the predicted connection and C_2 stands for the second node. These tables have been formulated based on different boundary conditions for the value of the cosine similarity.

[3]Source: https://www.hpl.hp.com/techreports/2003/HPL-2003-4.pdf

Predicted Position	C_1	Degree (C_1)	Neighbors (C_1)	C_2	Degree (C_2)	Neighbors (C_2)	C_1 - C_2 (Number of paths of length 2)	Cosine Similarity	Predicted Connection Score
Highest predicted Connection	beat note	649	497	deep learning	2415	1332	75	0.092	1.21
Lowest predicted Connection	test setup	64	58	cold cesium atom	86	73	0	0	-0.15

Table 5.2: The highest and the lowest predicted connections based on the connection score value

C_1	Degree (C_1)	Neighbors (C_1)	C_2	Degree (C_2)	Neighbors (C_2)	C_1 - C_2 (Number of paths of length 2)	Cosine Similarity	Predicted Connection Score
beat note	649	497	deep learning	2415	1332	75	0.092	1.21
neural network	2418	1428	nano optic	1092	917	166	0.145	1.18
electromagnetically induced transparency	4403	2400	space time	1958	1312	386	0.218	1.17
orbital angular momentum	8624	3642	wave optic	739	643	305	0.199	1.15
light matter interaction	8036	4255	lorentz law	332	153	78	0.097	1.149
phase noise	2263	1264	magnetic field	8385	4248	469	0.202	1.147
quantum information processing	6110	3157	anti stokes	1225	941	427	0.248	1.13
fiber link	574	364	electromagnetic field	10946	5110	140	0.103	1.12

Table 5.3: Top predicted connections without any filter

C_1	Degree (C_1)	Neighbors (C_1)	C_2	Degree (C_2)	Neighbors (C_2)	C_1 - C_2 (Number of paths of length 2)	Cosine Similarity	Predicted Connection Score
neural network	2418	1428	zero phonon line	760	509	75	0.088	1.082381
allan deviation	606	423	deep learning	2415	1332	67	0.089	1.0792067
st wave packet	162	81	wide range	9943	5335	44	0.067	1.0714669
fiber link	574	364	angular momentum	4183	2280	72	0.079	1.0558507
photonic crystal cavity	2448	1470	bulk band	433	306	58	0.086	1.0420154
low loss	7904	3931	3d object	247	208	75	0.083	1.0390215
fixed value	110	103	harmonic generation	8964	4002	42	0.065	1.0366176
lithium niobate	3225	1712	gamma ray	358	299	59	0.082	1.0345842

Table 5.4: Top predicted connections with a filter of cosine similarity. The cosine similarity of the predicted connection has to be less than 0.09 in this case.

C_1	Degree (C_1)	Neighbors (C_1)	C_2	Degree (C_2)	Neighbors (C_2)	C_1 - C_2 (Number of paths of length 2)	Cosine Similarity	Predicted Connection Score
squeezed light	1298	878	average power	1585	958	151	0.16	1.1090078
topological phenomenon	708	500	chern number	1223	705	181	0.30	1.0903783
carrier envelope phase	1540	981	gas jet	439	330	141	0.25	1.0585827
laser diode	1391	969	te mode	1177	985	161	0.16	1.0540853
single pas	1175	883	high purity	985	749	149	0.18	1.0480801
quasi phase matching	1205	768	harmonic field	386	331	84	0.17	1.0437386
carrying orbital angular momentum	976	713	paraxial beam	493	360	76	0.15	1.040059
timing jitter	1307	840	pump field	1212	959	127	0.14	1.0339742

Table 5.5: Top predicted connections with a filter of the number of neighbors. The number of neighbors for both the concepts should be less than 1000 here.

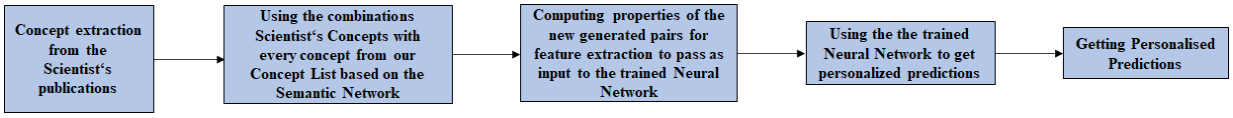


Figure 5.6: Personalised predictions model workflow

5.3 Personalised predictions for a Scientist

Now, it's time to put the trained model to good use. In the previous section, we saw the process to develop the model to propose new research topics. However, it would be interesting if we use this model to suggest interesting research topic combinations based on the research interests of one particular scientist. We have attempted to tackle this issue with the help of our trained neural network.

5.3.1 Data preparation and processing

Before we use our trained model, we need the data of which edges are unconnected when it comes to the Scientist of our interest. The very first element that we need to proceed further is the papers published by the Scientist. As we already processed multiple papers earlier from arXiv to extract concepts, we already have a good amount of keywords that highlight diffrent research directions in the field of Optics. So, now, we don't need to constrain ourselves to arXiv to mine the Scientist specific papers. This time, the papers are extracted manually from Google Scholar. Google Scholar is nothing less than a gold mine for researchers. It is a freely accessible web search engine that stores the full text or

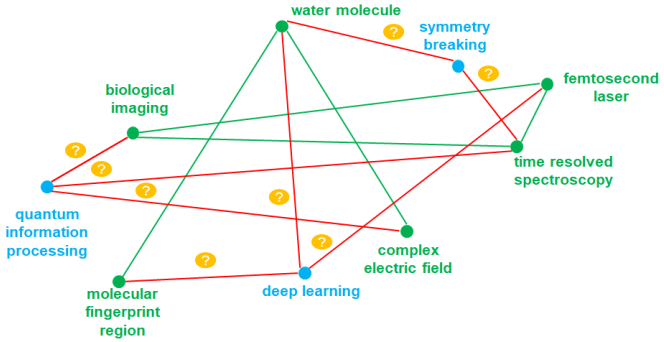


Figure 5.7: The idea behind personalized predictions. Concepts in green color are extracted from the Scientist's publication and the ones in blue come from the other research articles published in the field of Optics. The green color edges are the topics that have been investigated by the Scientist earlier. The ones in red are something which the Scientist has not tackled together yet and thus, makes it a potential combination for personalized predictions.

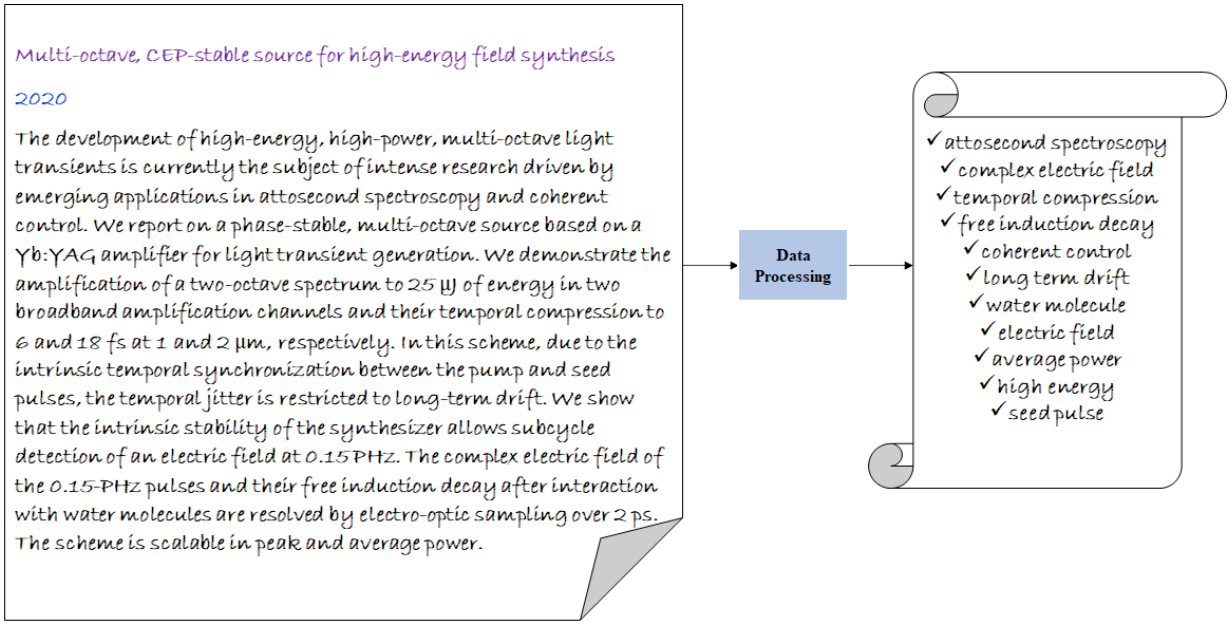
metadata of scholarly literature across multiple disciplines. Released in beta in November 2004, the Google Scholar index includes peer-reviewed online academic journals and books, conference papers, theses and dissertations, preprints, abstracts, technical reports, and other scholarly literature, including court opinions and patents [4]. This experiment was performed on the research papers of the Scientist, Prof. Dr. Hanieh Fattahi. So, in total, we extract 53 papers for our Scientist from the Google Scholar database.

Once we have all the papers at hand, the next step is to look for the concepts from our list and if it is present in the Scientist's publication. A total of 278 concepts are extracted. The next step is to begin the preparation of the combination of potential future research topics for the Scientist. The first step is pairing every concept from the list we created in this section with every concept in the list that we created in section 3.1.5. After this, we end up with a list of around 4.37 million combinations.

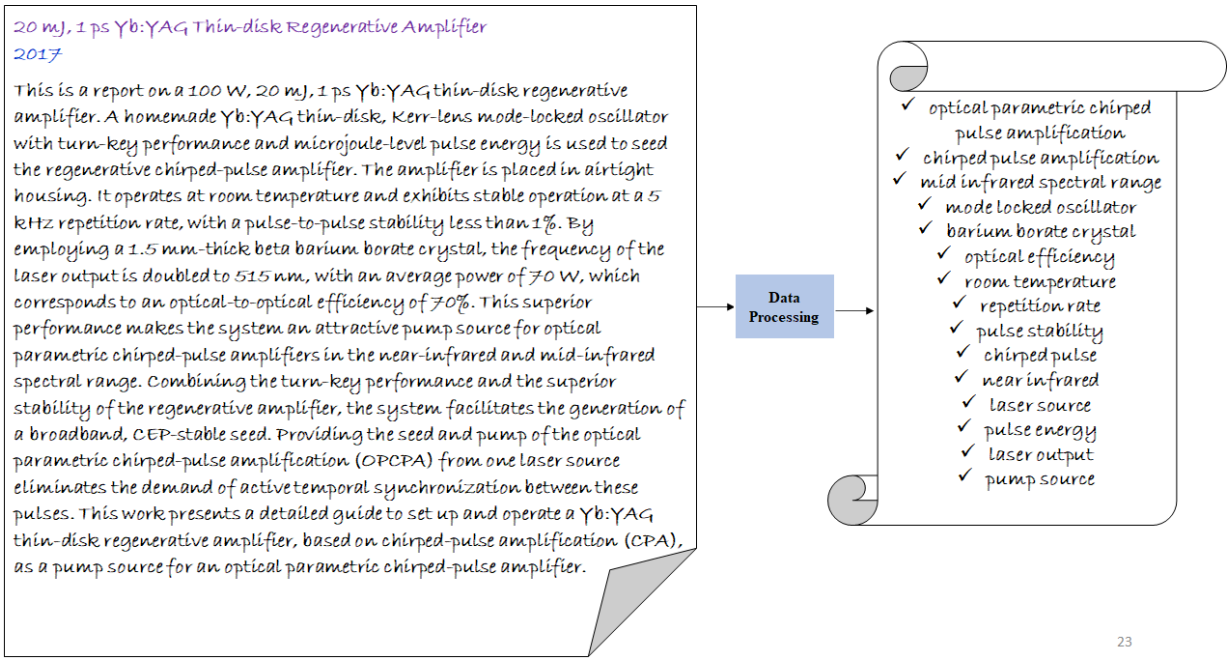
5.3.2 Personalised predictions

The complete personalized prediction process workflow is illustrated in figure 5.6. The semantic network visualization in figure 5.7 illustrates the logic behind making the connections personalized. It is similar to the generalized prediction process discussed in section 5.1.2. We get personalized predictions using our trained neural network to predict the most

[4]Source: https://scholar.google.com/intl/us/scholar/help.htmlcover



(a) (The paper title and abstract shown in this image is from [65].



(b) (The paper title and abstract shown in this image is from [66].

Figure 5.8: Concept extraction from the papers of the chosen Scientist

Figure 5.9: Insights into the research interests of the Scientist based on the extracted research material from the Scientist. A darker shade indicates that the chosen Scientist investigates that particular concept a lot compared to the entire scientific community of Optics. A lighter shade denotes the opposite.

C_1	Degree (C_1)	Neighbors (C_1)	C_2	Degree (C_2)	Neighbors (C_2)	C_1 - C_2 (Number of paths of length 2)	Cosine Similarity	Predicted Connection Score
low loss	7904	3931	room temperature	12143	5052	2094	0.47	1.31
high order harmonic generation	1683	957	attosecond pulse	1544	911	374	0.40	1.26
quantum information processing	6110	3157	linear optic	1400	1037	526	0.29	1.24
light matter interaction	8036	4255	low energy	1693	1356	763	0.32	1.21
wide range	9943	5335	time domain	4769	3044	1683	0.42	1.20
single photon	10534	4322	noise ratio	5439	2989	1341	0.37	1.19
neural network	2418	1428	label free	1738	1126	297	0.23	1.171
low noise	3142	1748	numerical simulation	14328	6472	1045	0.31	1.170

Table 5.6: Top personalized predicted connections without any filter

and least probable connections from the data we generated here. Once, we have this list, we can analyze it by applying several filters like cosine similarity, node degree, or even by directly querying the data with a particular concept to understand the prediction strength. The Scientist, thus, in the end, gets all the interesting combined research topics that he/ she initially might not have thought of. This list can spark some interest among the research fraternity to investigate new things together and can also trigger collaborations to unravel the mysteries of Optics. Some interesting suggestions were given by the model we trained earlier. The results are summarized in the tables 5.6, 5.7 and 5.8. The summary in these tables is similar to the tables discussed in section 5.2.4 where the data was segregated based on the cosine similarity value. The only difference here is that the name of the second concept (C_2) comes from the concept list extracted only from the research material of the chosen Scientist. The first concept (C_1) comes from the complete research material on arXiv published in the field of Optics.

C_1	Degree (C_1)	Neighbors (C_1)	C_2	Degree (C_2)	Neighbors (C_2)	C_1 - C_2 (Number of paths of length 2)	Cosine Similarity	Predicted Connection Score
intensity noise	520	419	linear optic	1400	1037	56	0.085	1.079
neural network	2418	1428	thz generation	450	320	60	0.089	1.072
deep learning	2415	1332	driving laser	648	513	70	0.085	1.06
channel loss	230	181	harmonic generation	8964	4002	61	0.072	1.049
inverse design	1748	1092	spectral phase	709	537	64	0.084	1.047
wide range	9943	5335	soft tissue	139	123	72	0.089	1.03
output power	2342	1404	small particle	903	716	90	0.090	1.016
thin film	8115	4042	reference laser	198	183	73	0.085	1.0002

Table 5.7: Top personalized predicted connections with a filter of cosine similarity. The cosine similarity of the predicted connection has to be less than 0.09 in this case.

C_1	Degree (C_1)	Neighbors (C_1)	C_2	Degree (C_2)	Neighbors (C_2)	C_1 - C_2 (Number of paths of length 2)	Cosine Similarity	Predicted Connection Score
high order harmonic generation	1683	957	attosecond pulse	1544	911	374	0.40	1.26
squeezed light	1298	878	long time	972	853	124	0.14	1.11
intensity noise	520	419	fs pulse	1577	986	109	0.17	1.08
laser diode	1391	969	cycle pulse	1345	878	161	0.17	1.06
comb generation	1164	764	average power	1585	958	179	0.21	1.05
superconducting nanowire single photon detector	1199	703	timing jitter	1307	840	287	0.37	1.04
squeezed state	1412	954	broad spectra	1028	875	129	0.14	1.02
quantum light	1212	883	photon flux	824	675	142	0.18	1.01

Table 5.8: Top personalized predicted connections with a filter of the number of neighbors. The number of neighbors for both the concepts should be less than 1000 here.

Chapter 6

Conclusion and Outlook

This thesis aimed at kick-starting the work that utilizes the power of Semantic Networks to pave the way for generating new, interesting, and surprising research topics in the field of Optics. By processing published papers in the field of Optics, a dataset of keywords that occur in the publications in this field was successfully created as a part of this work. This dataset throws light on various scientific terminologies used in the field of Optics and also on different subfields of research in Optics. The extracted data was stored in the form of a semantic network such that the fields that have been investigated together in one article are connected in the network and the connection has a timestamp that throws light on the publication date. The developed semantic network has very important information stored in it that can be accessed to have an overview of the prominent research topics during a particular period. The neural network proposed in this thesis makes it easier to determine the dynamics of emerging research topics and to find hidden patterns in the connections in the past. This neural network was then deployed to predict interesting research directions for one particular Scientist. The only required input, in the beginning, would be relevant research papers and then the entire proposed workflow in this thesis can be executed to process the papers, extract keywords, create the semantic network, train the neural network and get personalized predictions according to your interest at the end.

The foundation on which this thesis is built is the research papers. As data processing and concept extraction is a time-consuming process, we have limited ourselves and only used the papers published on arXiv to propose the idea. Thus, some of the predicted combinations can be such that these have already been tackled together in one paper earlier but this paper has been published on other platforms. There are various other platforms where papers are published in the domain of Optics. If research papers from all the possible

platforms are collected, converted to a uniform format, and processed, such a tool has the potential to create wonders in the field of scientific research as our model performs very well on the arXiv data that we feed to it so it has a great potential to find hidden patterns.

The semantic connection criteria that we have used in this work is the paper publication date. Future research can focus on other impacting parameters as well, such as the number of citations, authors, collaborations among scientists, and other relevant factors as required. The ongoing developments relevant to this can be used to achieve the target. One of the best examples is OpenAlex, a free and open catalog of the world's scholarly papers, researchers, journals, and institutions. This catalog also gives an overview of the way these factors are connected [67]. Research data can be efficiently managed by tracking citation impact, spotting promising new research areas, and identifying and promoting work from underrepresented groups. The more factors are considered, the higher is the possibility to unravel the information hidden in past research work and guide us to the future. As we read in the previous chapters, in every single operation, the keywords, which we have called concepts are required. We have extracted these concepts from the research papers by deploying multiple natural language processing algorithms considering the scope of the work. Future research can focus on developing a Machine Learning Model to generate the concept list itself. Also, if the clusters of concepts in the semantic network are investigated, there is a possibility that all the clustered concepts point toward a whole new research domain which would be very interesting to investigate. One important factor to be considered in future work can be finding the importance and impact of the new connection. As a tool for high-quality suggestions, the computation of a 'metric-of-success', for example, estimated citation numbers of the new link or the rate of citation growth over time can be helpful. The collected data can be stored using modern Graph Database Algorithms available today and using these for the storage of research data can only enhance the decision-making process of the researchers as getting the required information at the right time at high speed by simply querying these giant graphs is the feature that makes these tools powerful.

If these further improvements fall into place with the help of collaboration between Optics Researchers and Computer Scientists, the power of the tool developed in this thesis can be amplified to create wonders in research in the field of Optics.

List of Figures

1.1 Growth in the number of scientists and publications in the field of Optics during the past century 2

1.2 The user interface of the GPT 3 tool. Here, an abstract from a scientific text [4] is entered in order to get the keywords out of it. The words highlighted in green are the extracted words. It can be seen that some of the important words like *symmetry breaking*, *dielectric optical resonator*, *evanescent coupling* are not even extracted. This indicates that we might lose some important information discussed in the piece of text. 3

2.1 Diagrammatic inner working of SEMNET. Human-generated concept lists (from Wikipedia and books) are combined with automatically generated lists (with natural language processing, using RAKE on 100,000 arXiv articles) to generate a list of quantum physics concepts. Each concept forms a link in a semantic network. The edges are formed when two concepts coappear in a title or abstract of any of the 750,000 papers (from arXiv and APS). A mini-version of SEMNET is shown, using parts of three articles from APS. Edges carry temporal information of their formation year, which leads to an evolution of the semantic network SEMNET over time [5]. (Image from [5]). 8

2.2 The semantic network in Biochemistry. The nodes represent the names of biomolecules. An edge is drawn to connect two biomolecules if these two have been studied together in one published article (Image from [7]). . . . 10

2.3 Examples of semantic networks 11

2.4 A simple graph of 6 nodes and 7 edges. The vertices are labeled as 1, 2, 3, 4, 5, and 6. Each label is unique so that the labels can be used to refer to any vertex unambiguously. If there is an edge between vertices i and j, this connection can be denoted as (i,j). In this way, the complete network can be specified by giving the number of edges and a list of all the edges. So, the graph in figure 2.3 has edges, (1,2), (1,5), (2,3), (2,4), (3,4), (3,5), and (3,6). 12
2.5 Number of paths of length 2. Here, nodes 2 and 5 are considered as an example. These nodes have 2 paths of length 2, which means there are two possibilities where we need to cross 2 edges to reach from one node to the other. 14
2.6 A simple graph (2) 15
2.7 A general artificial neural network architecture 16
2.8 Computational model of a feedforward neural network 17
2.9 Computational model of a neuron. This illustrates the way in which inputs are transformed to predict an output. The illustration is based on the weather forecasting problem discussed in section 2.3.3. 19
2.10 The Gradient descent algorithm. The approach here is to change the training parameters at a suitable learning rate to find the minimum loss point (Image from [17]). 20
2.11 Lemmatizing words using NLTK 22
2.12 Generation of stop words in English language using NLTK 22

3.1 Growth in papers published under the category of Optics 26
3.2 Paper data Example [27] 27
3.3 Data preprocessing workflow. This block diagram illustrates the steps and the sequence that is followed to get the research articles, process them and generate a list of important words that highlight the topic of the article. . . 28
3.4 Workflow of the RAKE Algorithm 29
3.5 Concept list snapshot. These are some randomly chosen concepts from the final concept list to give a brief idea of how the list looks like. 31
3.6 Concept list extraction example (The paper title and abstract shown in this image is from [28]) 32
3.7 Logic behind the semantic network creation (image from [28], [29], [30], [31]) 33
3.8 Growth in the co-occurrence of concepts from 1995 to 2021 35

4.1 Plot of the evolving concepts [period = 5 Years] 38
4.2 Plot of the evolving concepts [period = 3 Years] 44
4.3 Plot of the popular concepts [period = 1 Year] 45

5.1 Neural network layers 48
5.2 Sequence of the layers in the network used to build the link prediction model. For better visualization to give an overview of the interconnections, the number of neurons in each layer has been reduced. In the real network, the first, second, third and fourth layers have 15, 100, 100 and 10 neurons respectively. The fifth layer that where we receive the output as a prediction connection score (discussed in section 5.2.4), has one neuron. This plot is made using the NN SVG tool [63] 49
5.3 Link prediction model workflow 50
5.4 Early stopping checkpoint 51
5.5 Receiver Operating Characteristic (ROC) Curve 54
5.6 Personalised predictions model workflow 57
5.7 The idea behind personalized predictions. Concepts in green color are extracted from the Scientist's publication and the ones in blue come from the other research articles published in the field of Optics. The green color edges are the topics that have been investigated by the Scientist earlier. The ones in red are something which the Scientist has not tackled together yet and thus, makes it a potential combination for personalized predictions. . . 58
5.8 Concept extraction from the papers of the chosen Scientist 59
5.9 Insights into the research interests of the Scientist based on the extracted research material from the Scientist. A darker shade indicates that the chosen Scientist investigates that particular concept a lot compared to the entire scientific community of Optics. A lighter shade denotes the opposite. . . . 60

List of Tables

3.1 Word frequency. The number corresponding to every word in this table indicates the number of times that particular word appears in the piece of text that is being processed. 30

3.2 Word co-occurrence count. The co-occurrence count is 1 if we pair every word with itself. When a word is paired with other words in a sentence and if these words appear next to each other, the count for this pair is 1. In this case, the words, 'deep' and 'learning' appear one after the other in the text and thus, both of these have a co-occurrence count of 1 when paired with each other. This is important to extract 'deep learning' as one keyword and not separately as two keywords, 'deep' and 'learning'. 30

3.3 Word scores. This number gives an overview of every word's appearance with other words in the text. This gives an idea of the word quality. A higher number here means that the particular word appears with many other words and is a word of high importance. 30

3.4 Scores of candidate keywords. This number indicates the quality of the whole keyword. The higher the score, the more important is the keyword to help throw light on the gist of the text. 31

3.5 Node and edge number for the networks shown in figure 3.8 34

5.1 The Confusion matrix parameters 52

5.2 The highest and the lowest predicted connections based on the connection score value 56

5.3 Top predicted connections without any filter 56

5.4 Top predicted connections with a filter of cosine similarity. The cosine similarity of the predicted connection has to be less than 0.09 in this case. 56

5.5 Top predicted connections with a filter of the number of neighbors. The number of neighbors for both the concepts should be less than 1000 here. . 57
5.6 Top personalized predicted connections without any filter 61
5.7 Top personalized predicted connections with a filter of cosine similarity. The cosine similarity of the predicted connection has to be less than 0.09 in this case. 62
5.8 Top personalized predicted connections with a filter of the number of neighbors. The number of neighbors for both the concepts should be less than 1000 here. 62

Bibliography

[1] *arXiv Annual Update, January 2019*. URL: https://arxiv.org/about/reports/2019_update (cited on p. 1).

[2] *arXiv Annual Update, January 2020*. URL: https://arxiv.org/about/reports/2020_update (cited on p. 1).

[3] *arXiv Annual Report 2021*. URL: https://static.arxiv.org/static/arxiv.marxdown/0.1/about/reports/2021_arXiv_annual_report.pdf (cited on p. 1).

[4] Jonathan M. Silver and Pascal Del'Haye. "Generalized theory of optical resonator and waveguide modes and their linear and Kerr nonlinear coupling". In: *Phys. Rev. A 105, 023517* (Feb. 2022). URL: https://journals.aps.org/pra/abstract/10.1103/PhysRevA.105.023517 (cited on p. 3).

[5] Mario Krenn and Anton Zeilinger. "Predicting research trends with semantic and neural networks with an application in quantum physics". In: *The Proceedings of the National Academy of Sciences (PNAS)* (Jan. 2020). URL: https://www.pnas.org/doi/full/10.1073/pnas.1914370116 (cited on pp. 2, 5, 7–9, 25).

[6] Mario Krenn, Lorenzo Buffoni, Bruno Coutinho, Sagi Eppel, Jacob Gates Foster, Andrew Gritsevskiy, Harlin Lee, Yichao Lu, Joao P. Moutinho, Nima Sanjabi, Rishi Sonthalia, Ngoc Mai Tran, Francisco Valente, Yangxinyu Xie, Rose Yu, and Michael Kopp. *Predicting the Future of AI with AI: High-Quality link prediction in an exponentially growing knowledge network*. Version 1. Sept. 2022. URL: https://arxiv.org/abs/2210.00881 (cited on pp. 2, 9, 25).

[7] Andrey Rzhetskya, Jacob G. Fosterd, Ian T. Fosterb, and James A. Evansb. "Choosing experiments to accelerate collective discovery". In: *Proceedings of the National Academy of Sciences* (Nov. 2015). URL: https://www.pnas.org/doi/full/10.1073/pnas.1509757112 (cited on pp. 2, 7, 10).

[8] Dashun Wang and Albert-László Barabási. *The Science of Science*. Cambridge University Press, 2021. DOI: 10.1017/9781108610834 (cited on pp. 4, 5).

[9] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. "Language Models are Few-Shot Learners". Version 4. In: (July 2020). URL: https://arxiv.org/abs/2005.14165 (cited on p. 4).

[10] P. He, B. Akhgar, and H. Arabnia. "Emerging Trends in ICT Security (pages 455-467)". In: *Sciencedirect* (2014). URL: https://www.sciencedirect.com/science/article/pii/B978012411474600027X (cited on p. 9).

[11] Björn Titz, Seesandra V. Rajagopala, Johannes Goll, Roman Häuser, Matthew T. McKevitt, Timothy Palzkill, and Peter Uetz. "The Binary Protein Interactome of Treponema pallidum The Syphilis Spirochete". In: *Plos One* (May 2008). URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0002292 (cited on p. 11).

[12] Giacomo Vaccario, Luca Verginer, and Frank Schweitzer. "Reproducing scientists' mobility: A data-driven model". In: *Scientific Reports* (May 2021). URL: https://www.nature.com/articles/s41598-021-90281-9 (cited on p. 11).

[13] M. E. J. Newman. *Networks: an introduction*. Oxford; New York: Oxford University Press, Mar. 2010. URL: https://academic.oup.com/book/27303 (cited on pp. 10, 12–14).

[14] Kevin Gurney. *An introduction to neural networks*. Taylor Francis, Inc., Mar. 1997. URL: https://dl.acm.org/doi/10.5555/523781 (cited on pp. 15, 16).

[15] Juergen Schmidhuber. "Deep Learning in Neural Networks: An Overview". In: *Neural Networks, Vol 61, pp 85-117, Jan 2015* (Jan. 2015). URL: https://www.sciencedirect.com/science/article/abs/pii/S0893608014002135?via%3Dihub (cited on p. 15).

[16] Vinod Nair and Geofrey E. Hinton. "Rectified Linear Units Improve Restricted Boltzmann Machines". In: *Association for Computing Machinery* (June 2010). URL: https://dl.acm.org/doi/10.5555/3104322.3104425 (cited on p. 18).

[17] Brad Boehmke and Brandon M. Greenwell. *Hands-On Machine Learning with R*. Chapman Hall, Nov. 2019. URL: https://www.routledge.com/Hands-On-Machine-Learning-with-R/Boehmke-Greenwell/p/book/9781138495685 (cited on p. 20).

[18] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. "Deep Learning". In: *Nature* (May 2015). URL: https://www.nature.com/articles/nature14539 (cited on p. 20).

[19] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. *PyTorch: An Imperative Style, High-Performance Deep Learning Library*. Version 1. Dec. 2019. URL: https://arxiv.org/abs/1912.01703 (cited on p. 21).

[20] Mo Patel. "When two trends fuse: PyTorch and recommender systems". In: *O'Reilly Media* (Dec. 2017). URL: https://www.oreilly.com/content/when-two-trends-fuse-pytorch-and-recommender-systems/ (cited on p. 21).

[21] Steven Bird, Ewan Klein, and Edward Loper. *Natural Language Processing with Python*. Oxford; New York: O'Reilly Media, 2009. URL: https://www.nltk.org/book/ (cited on p. 22).

[22] Stuart Rose, Dave Engel, Nick Cramer, and Wendy Cowley. "Automatic Keyword Extraction from Individual Documents". In: *Wiley Online Library* (Mar. 2010). URL: https://onlinelibrary.wiley.com/doi/10.1002/9780470689646.ch1 (cited on pp. 22, 29–31).

[23] Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. "Array programming with NumPy". In: *Nature* 585.7825 (Sept. 2020), pp. 357–362. URL: https://doi.org/10.1038/s41586-020-2649-2 (cited on p. 23).

[24] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C

J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python". In: *Nature Methods* 17 (2020), pp. 261–272. DOI: 10.1038/s41592-019-0686-2 (cited on p. 23).

[25] Mathieu Bastian, Sebastien Heymann, and Mathieu Jacomy. "Gephi: An Open Source Software for Exploring and Manipulating Networks". In: *International AAAI Conference on Weblogs and Social Media* (Mar. 2009). URL: https://ojs.aaai.org/index.php/ICWSM/article/view/13937/13786 (cited on pp. 23, 36).

[26] Paul Ginsparg. "Lessons from arXiv's 30 years of information sharing". In: *Nature Reviews Physics* (Aug. 2021). URL: https://www.nature.com/articles/s42254-021-00360-z (cited on pp. 25, 26).

[27] Xuemei Gu, Mario Krenn, Manuel Erhard, and Anton Zeilinger. "Gouy Phase Radial Mode Sorter for Light: Concepts and Experiments". Version 1. In: *Physical Review Letters* (Dec. 2017). URL: https://journals.aps.org/prl/abstract/10.1103/PhysRevLett.120.103601 (cited on p. 27).

[28] Mario Krenn, Robert Fickler, Matthias Fink, Johannes Handsteiner, Mehul Malik, Thomas Scheidl, Rupert Ursin, and Anton Zeilinger. "Communication with spatially modulated Light through turbulent Air across Vienna". In: *New Journal of Physics* (Nov. 2014). URL: https://iopscience.iop.org/article/10.1088/1367-2630/16/11/113028 (cited on pp. 32, 33).

[29] Mario Krenn, Marcus Huber, Robert Fickler, Radek Lapkiewicz, Sven Ramelow, and Anton Zeilinger. "Generation and Confirmation of a (100x100)-dimensional entangled Quantum System". Version 1. In: *The Proceedings of the National Academy of Sciences (PNAS)* (June 2013). URL: https://www.pnas.org/doi/full/10.1073/pnas.1402365111 (cited on p. 33).

[30] Mario Krenn, Johannes Handsteiner, Matthias Fink, Robert Fickler, Rupert Ursin, Mehul Malik, and Anton Zeilinger. "Twisted Light Transmission over 143 kilometers". In: *The Proceedings of the National Academy of Sciences (PNAS)* (June 2016). URL: https://www.pnas.org/doi/full/10.1073/pnas.1612023113 (cited on p. 33).

[31] Florian Schlederer, Mario Krenn, Robert Fickler, Mehul Malik, and Anton Zeilinger. "Cyclic transformation of orbital angular momentum modes". Version 1. In: *New Journal of Physics* (Apr. 2016). URL: https://iopscience.iop.org/article/10.1088/1367-2630/18/4/043019 (cited on p. 33).

[32] Mathieu Jacomy, Tommaso Venturini, Sebastien Heymann, and Mathieu Bastian. "ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software". In: *PLOS ONE* (June 2014). URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0098679 (cited on p. 36).

[33] Abir Bandyopadhyay and Jagdish Rai. "Squeezing in the interaction of radiation with two-level atoms". Version 2. In: *International Symposium on Atomic Coherence and Inversion-less Amplification (ISAMP, Changchun, China)* (Oct. 1995). URL: https://arxiv.org/abs/atom-ph/9509005 (cited on p. 38).

[34] Valery Telnov. "Laser cooling of electron beams for linear colliders". Version 2. In: *Phys.Rev.Lett. 78 (1997) 4757-4760; Erratum-ibid. 80 (1998) 2747* (Oct. 1996). URL: https://journals.aps.org/prl/abstract/10.1103/PhysRevLett.78.4757 (cited on p. 38).

[35] Christopher S. Baird. *Electromagnetic field*. Mc Graw Hill, Aug. 2021. URL: https://www.accessscience.com/content/article/a222300 (cited on p. 39).

[36] Z. S. Bassi and A. LeClair. "A one-dimensional model for n-level atoms coupled to an electromagnetic field". Version 1. In: *Journal of Mathematical Physics* (Mar. 1997). URL: https://aip.scitation.org/doi/10.1063/1.532922 (cited on p. 39).

[37] Daniel Poitras and Ludvik Martinu. "A simple method for the determination of slowly varying refractive index profiles from in situ spectrophotometric measurements". Version 1. In: *Optica Publishing Group (formerly OSA)* (Apr. 1998). URL: https://opg.optica.org/ao/abstract.cfm?uri=ao-37-19-4160 (cited on p. 39).

[38] G. Vacca, R. D. Morgan, and R. B. Laughlin. "Differential light scattering: probing the sonoluminescence collapse". Version 1. In: *J Phys. Rev. E 60 (6), R6303-6306* (Dec. 1999). URL: https://journals.aps.org/pre/abstract/10.1103/PhysRevE.60.R6303 (cited on p. 39).

[39] Valery Telnov. "Photon collider at TESLA". Version 1. In: *Nucl.Instrum.Meth.A472:43-60,2001* (Oct. 2000). URL: https://doi.org/10.1016/S0168-9002(01)01161-5 (cited on p. 39).

[40] Steven W. Ellingson. *Electromagnetics, Volume 2*. Virginia Tech, Jan. 2020. URL: https://vtechworks.lib.vt.edu/handle/10919/93253 (cited on p. 40).

[41] A. I. A. Adewole. *Light Propagation For Accelerated Observers*. Version 3. Oct. 2001. URL: https://arxiv.org/abs/physics/0104069 (cited on p. 40).

[42] Gennady Shvets. *Photonic Approach to Making a Left-Handed Material*. Version 3. Oct. 2002. URL: https://arxiv.org/abs/physics/0206004 (cited on p. 40).

[43] David L. Andrews, Luciana C. Davila Romero, and Mohamed Babiker. "Molecular chirality and the orbital angular momentum of light". Version 1. In: *Optics Communications* (May 2003). URL: https://arxiv.org/abs/physics/0305002 (cited on p. 40).

[44] Cyriaque Genet, Martin P. van Exter, and J.P. Woerdman. "Fano-type interpretation of red shifts and red tails in hole array transmission spectra". Version 1. In: *Optics Communications* (Jan. 2004). URL: https://arxiv.org/abs/physics/0401054 (cited on p. 40).

[45] S. Burger, L. Zschiedrich, R. Klose, A. Schädle, F. Schmidt, C. Enkrich, S. Linden, M. Wegener, and C. M. Soukoulis. "Numerical Investigation of Light Scattering off Split-Ring Resonators". Version 1. In: *Proc. SPIE Vol. 5955 (2005) 595503* (Oct. 2005). URL: https://www.spiedigitallibrary.org/conference-proceedings-of-spie/5955/1/Numerical-investigation-of-light-scattering-off-split-ring-resonators/10.1117/12.622184.short (cited on p. 41).

[46] Maha Intakhab Alam, Muhammad Amin, and Irfan Majid. "The Physics of Invisibility Cloak". In: *2020 17th International Bhurban Conference on Applied Sciences and Technology (IBCAST)* (2020). URL: https://ieeexplore.ieee.org/document/9044535 (cited on p. 41).

[47] D. Schurig, J.B. Pendry, and D.R. Smith. "Calculation of material properties and ray tracing in transformation media". Version 1. In: *Optica Publishing Group (formerly OSA)* (July 2006). URL: https://opg.optica.org/oe/fulltext.cfm?uri=oe-14-21-9794&id=116380 (cited on p. 41).

[48] Marco Rahm, David Schurig, Daniel A. Roberts, Steven A. Cummer, David R. Smith, and John B. Pendry. "Design of Electromagnetic Cloaks and Concentrators Using Form-Invariant Coordinate Transformations of Maxwell's Equations". Version 1. In: *Photon. Nanostruct.: Fundam. Applic. 6, 87 (2008)* (June 2007). URL: https://doi.org/10.1016/j.photonics.2007.07.013 (cited on p. 41).

[49] Jasgurpreet S. Chohan and Rupinder Singh. *Encyclopedia of Materials: Plastics and Polymers (Chapter: Thermosetting Polymer Application as Meta Materials.* Elsevier Inc., 2022. URL: https://doi.org/10.1016/B978-0-12-820352-1.00159-0 (cited on p. 41).

[50] E. Plum, V. A. Fedotov, and N. I. Zheludev. "Optical Activity of Planar Achiral Metamaterials". Version 1. In: *Phys. Rev. Lett., vol. 102, page 113902 (2009)* (July 2008). URL: https://aip.scitation.org/doi/10.1063/1.3021082 (cited on p. 41).

[51] Willem L. Vos, A. Femius Koenderink, and Ivan S. Nikolaev. "Orientation-dependent spontaneous emission rates of a two-level quantum emitter in any nanophotonic environment". Version 2. In: *Phys. Rev. A 80, 053802 (2009)* (Dec. 2009). URL: https://journals.aps.org/pra/abstract/10.1103/PhysRevA.80.053802 (cited on p. 42).

[52] Lingzhi Li, Yuan Cao, Yanyan Zhi, Jiejun Zhang, Yuting Zou, Xinhuan Feng, Bai-Ou Guan, and Jianping Yao. "Polarimetric parity-time symmetry in a photonic system". In: *Light: Science Applications (LSA)* (Sept. 2020). URL: https://www.nature.com/articles/s41377-020-00407-3 (cited on p. 42).

[53] Andrey A. Sukhorukov, Zhiyong Xu, and Yuri S. Kivshar. "Nonlinear suppression of time-reversals in PT-symmetric optical couplers". Version 1. In: *Phys. Rev. A 82, 043818* (Oct. 2010). URL: https://journals.aps.org/pra/abstract/10.1103/PhysRevA.82.043818 (cited on p. 42).

[54] Nardeep Kumar, Brian A. Ruzicka, N. P. Butch, P. Syers, K. Kirshenbaum, J. Paglione, and Hui Zhao. "Spatially resolved femtosecond pump-probe study of topological insulator Bi2Se3". Version 1. In: *Phys. Rev. B 83, 235306 (2011)* (Apr. 2011). URL: https://journals.aps.org/prb/abstract/10.1103/PhysRevB.83.235306 (cited on p. 42).

[55] Paloma A. Huidobro, Alexey Y. Nikitin, Carlos González-Ballestero, Luis Martín-Moreno, and Francisco J. García-Vidal. "Superradiance mediated by Graphene Surface Plasmons". Version 1. In: *Phys. Rev. B 85, 155438* (Jan. 2012). URL: https://journals.aps.org/prb/abstract/10.1103/PhysRevB.85.155438 (cited on p. 42).

[56] Shuichi Murakami. "Two-dimensional topological insulators and their edge states". In: *Journal of Physics: Conference Series and International Symposium Nanoscience and Quantum Physics 2011"(nanoPHYS'11) 26–28 January 2011, Minato-ku, Tokyo,*

Japan (2011). URL: https://iopscience.iop.org/article/10.1088/1742-6596/302/1/012019 (cited on p. 43).

[57] S. Mittal M. Hafezi, A. Migdall J. Fan, and J. Taylor. "Imaging topological edge states in silicon photonics". Version 1. In: *Nature Photonics 7, 1001 (2013) (v2 extended version)* (Feb. 2013). URL: https://www.nature.com/articles/nphoton.2013.274 (cited on p. 43).

[58] G. Mazzamuto, A. Tabani, S. Pazzagli, S. Rizvi, A. Reserbat-Plantey, K. Schädler, G. Navickaité, L. Gaudreau, F. S. Cataliotti, F. Koppens, and C. Toninelli. "Single-molecule study for a graphene-based nano-position sensor". In: *New Journal of Physics, Volume 16* (July 2014). URL: https://iopscience.iop.org/article/10.1088/1367-2630/16/11/113007 (cited on p. 43).

[59] Bo Meng, Matthew Singleton, Johannes Hillbrand, Martin Franckié, Mattias Beck, and Jérôme Faist. "Dissipative Kerr solitons in semiconductor ring lasers". In: *Nature Photonics* (Dec. 2021). URL: https://www.nature.com/articles/s41566-021-00927-3 (cited on p. 43).

[60] Tobias Herr, Michael L. Gorodetsky, and Tobias J. Kippenberg. *Dissipative Kerr solitons in optical microresonators*. Version 1. Aug. 2015. URL: https://arxiv.org/abs/1508.04989 (cited on p. 43).

[61] Yichen Shen, Nicholas C. Harris, Scott Skirlo, Mihika Prabhu, Tom Baehr-Jones, Michael Hochberg, Xin Sun, Shijie Zhao, Hugo Larochelle, Dirk Englund, and Marin Soljacic. "Deep Learning with Coherent Nanophotonic Circuits". Version 1. In: *Nature Photonics* (Oct. 2016). URL: https://www.nature.com/articles/nphoton.2017.93 (cited on p. 43).

[62] Myoung-Gyun Suh and Kerry Vahala. "Soliton Microcomb Range Measurement". Version 3. In: *Science 23 Feb 2018: Vol. 359, Issue 6378, pp. 884-887* (June 2017). URL: https://www.science.org/doi/10.1126/science.aao1968 (cited on p. 44).

[63] Alexander LeNail. "NN-SVG: Publication-Ready Neural Network Architecture Schematics". In: *The Journal of Open Source Software* (Jan. 2019). URL: https://joss.theoj.org/papers/10.21105/joss.00747 (cited on p. 49).

[64] Victor Dey. *Understanding the AUC-ROC Curve in Machine Learning Classification*. Sept. 2021. URL: https://analyticsindiamag.com/understanding-the-auc-roc-curve-in-machine-learning-classification/ (cited on p. 54).

[65] Ayman Alismail, Haochuan Wang, Gaia Barbiero, Najd Altwaijry, Syed Ali Hussain, Volodymyr Pervak, Wolfgang Schweinberger, Abdallah M Azzeer, Ferenc Krausz, and Hanieh Fattahi. "Multi-octave, CEP-stable source for high-energy field synthesis". In: *Science Advances* (Feb. 2020). URL: https://www.science.org/doi/full/10.1126/sciadv.aax3408 (cited on p. 59).

[66] Ayman Alismail, Haochuan Wang, Jonathan Brons, and Hanieh Fattahi. "20 mJ, 1 ps Yb: YAG Thin-disk Regenerative Amplifier". In: *JoVE (Journal of Visualized Experiments)* (July 2017). URL: https://www.jove.com/t/55717/20-mj-1-ps-ybyag-thin-disk-regenerative-amplifier (cited on p. 59).

[67] Jason Priem, Heather Piwowar, and Richard Orr. *OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts*. Version 2. June 2022. URL: https://arxiv.org/abs/2205.01833 (cited on p. 64).