

**DEEP LEARNING TECHNIQUES FOR TEXT CLASSIFICATION**

**DIARDANO RAIHAN**

**SCHOOL OF ELECTRICAL AND ELECTRONIC**

**ENGINEERING**

**2021**

**DEEP LEARNING TECHNIQUES FOR TEXT CLASSIFICATION**

**DIARDANO RAIHAN**

**SCHOOL OF ELECTRICAL AND ELECTRONIC ENGINEERING**

**A DISSERTATION SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE IN COMPUTER CONTROL AND AUTOMATION**

**2021**

**Statement of Originality**

I hereby certify that the work embodied in this thesis is the result of original research, is free of plagiarised materials, and has not been submitted for a higher degree to any other University or Institution.

|  |  |  |
| --- | --- | --- |
| 23 April 2021 |  | ***[Signature]*** |
| Date |  | Diardano Raihan |

**Supervisor Declaration Statement**

I have reviewed the content and presentation style of this thesis and declare it is free of plagiarism and of sufficient grammatical clarity to be examined. To the best of my knowledge, the research and writing are those of the candidate except as acknowledged in the Author Attribution Statement. I confirm that the investigations were conducted in accord with the ethics policies and integrity standards of Nanyang Technological University and that the research data are presented honestly and without prejudice.

|  |  |  |
| --- | --- | --- |
| 23 April 2021 |  | ***[Signature]*** |
| Date |  | Prof. P. N. Suganthan |

**Authorship Attribution Statement**

This thesis contains material from [x number] paper(s) published in the following peer-reviewed journal(s) / from papers accepted at conferences in which I am listed as an author.

Chapter 4 is published as D.T. Murphy, S. Schmid, J.R. Hester, P.E.R. Blanchard, and W. Miiller. Coordination site disorder in spinel-type LiMnTiO4. *Inorganic* *Chemistry* **54**, 4636-4643 (2015). DOI: 10.1021/ic502747p.

The contributions of the co-authors are as follows:

A/Prof P. N. Suganthan provided the initial project direction and edited the manuscript drafts.

I prepared the manuscript drafts. The manuscript was revised by P. N. Suganthan.

I co-designed the study with A/Prof Siegbert Schmid and performed all the laboratory work at the School of Materials Science and Engineering and the Singapore Synchrotron Light Source. I also analyzed the data.

All microscopy, including sample preparation, was conducted by me in the Facility for Analysis, Characterization, Testing and Simulation.

Dr James Hester assisted in the collection of the neutron powder diffraction data.

Dr Peter Blanchard assisted in the interpretation of the X-ray absorption spectroscopy data and carried out the spectral interpretation.

Dr Wojciech Miiller assisted in the collection and provide guidance in the interpretation of the magnetic measurement data.

Chapter 5 is published as H. V Doan, B. Yao, Y. Fang, A. Sartbaeva, U. Hintermair, V. P Ting, Controlled Formation of Hierarchical Metal-Organic Frameworks using CO2 Expanded Solvent Systems. In press, *ACS Sustainable Chemistry &* *Engineering* (2017). DOI: 10.1021/acssuschemeng.7b01429.

The contributions of the co-authors are as follows:

Prof Ting suggested the materials area and edited the manuscript drafts.

I wrote the drafts of the manuscript. The manuscript was revised together with Dr. Sartbaeva and Dr. Yao.

I performed all the materials synthesis, collected X-ray diffraction patterns and visible light spectra, carried transmission electron microscopy, and conducted data evaluation.

Dr. Y. Fang conducted the Rietveld analysis of the powder X-ray diffraction data and single crystal structure determinations.

Dr U. Hintermair conducted the molecular dynamics simulations.

Ms. A. Sartbaeva prepared the samples for electron microscopy.

|  |  |  |
| --- | --- | --- |
| 23 April 2021 |  | ***[Signature]*** |
| Date |  | Diardano Raihan |

**Table of Contents**

[**Abstract** ii](#_Toc67084256)

[**Acknowledgements** iii](#_Toc67084257)

[**Acronyms** iv](#_Toc67084258)

[**Symbols** v](#_Toc67084259)

[**List of Figures** vi](#_Toc67084260)

[**List of Tables** vii](#_Toc67084261)

[**Chapter 1 Introduction** 1](#_Toc67084262)

[**1.1** **Motivation** 1](#_Toc67084263)

[**1.2** **Objectives and Scope** 2](#_Toc67084264)

[**1.3** **Major Contribution of the Dissertation** 2](#_Toc67084265)

[**1.4** **Organisation of the Dissertation** 2](#_Toc67084266)

[**Chapter 2 Literature Review** 3](#_Toc67084267)

[**2. 1** **Text Preprocessing** 3](#_Toc67084268)

[**2.1.1** **Text Cleaning** 3](#_Toc67084269)

[**2.1.2** **Text Representation** 3](#_Toc67084270)

[**2. 2** **Deep Learning Techniques** 4](#_Toc67084271)

[**2.2.1** **Feed Forward Neural Networks** 4](#_Toc67084272)

[**2.2.2** **Recurrent Neural Networks** 5](#_Toc67084273)

[**2.2.3** **Convolutional Neural Networks** 5](#_Toc67084274)

[**2.2.4** **Hybrid Neural Networks** 6](#_Toc67084275)

[**2.3** **Model Optimization and Evaluation** 6](#_Toc67084276)

[**2.3.1** **Optimization** 6](#_Toc67084277)

[**2.3.2** **Evaluation** 7](#_Toc67084278)

[**Chapter 3 Deep Learning Techniques for Text Classification** 8](#_Toc67084279)

[**3.1** **Datasets and Experimental Setup** 8](#_Toc67084280)

[**3.1.1** **Datasets** 8](#_Toc67084281)

[**3.1.2** **Experimental Setup** 8](#_Toc67084282)

[**3.2** **Bag-of-Words based Models** 8](#_Toc67084283)

[**3.2.1** **Multi-Layer Perceptron** 9](#_Toc67084284)

[**3.2.2** **Ensemble Deep Random Vector Functional Link** 9](#_Toc67084285)

[**3.3** **Word Embedding / Word2Vec based Models** 9](#_Toc67084286)

[**3.3.1** **Multi Layer Perceptron** 9](#_Toc67084287)

[**3.3.2** **Ensemble Deep Random Vector Functional Link** 9](#_Toc67084288)

[**3.3.3** **Convolutional Neural Networks** 10](#_Toc67084289)

[**3.3.4** **Temporal Convolutional Neural Networks** 10](#_Toc67084290)

[**3.3.5** **Gated Recurrent Unit** 10](#_Toc67084291)

[**3.3.6** **Long Short-Term Memory** 10](#_Toc67084292)

[**3.3.7** **Hybrid (CNN + LSTM)** 11](#_Toc67084293)

[**Chapter 4 Evaluation** 12](#_Toc67084294)

[**4.1** **Results** 12](#_Toc67084295)

[**4.2** **Discussion** 12](#_Toc67084296)

[**Chapter 5 Conclusions and Future Work** 13](#_Toc67084297)

[**5.1** **Conclusions** 13](#_Toc67084298)

[**5.2** **Recommendation in Future Work** 13](#_Toc67084299)

# **Abstract**

# **Acknowledgements**

# **Acronyms**

|  |  |
| --- | --- |
| AI | Artificial Intelligence |
| NLP | Natural language Processing |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

# **Symbols**

svvdSv

# **List of Figures**

# **List of Tables**

# **Chapter 1 Introduction**

This chapter introduces the concept of Deep Learning techniques in a subfield of Artificial Intelligence (AI) namely Natural language Processing (NLP) for text classification task. We will discuss the motivation behind the project along with the objectives, the scope, and how we organize it.

## **Motivation**

Natural Language Processing (NLP) is the process of automating and/or manipulating human natural language (text and speech) by a program. NLP has many important tasks, such as:

* Text classification;
* Image Captioning;
* Language Modeling;
* Machine translation;
* Etc.

Text classification is one of the most popular tasks in NLP that allows a program to classify free-text document based on pre-defined classes. The classes can be based on topic, genre, or sentiment. As an illustration, the followings are applications commonly used nowadays:

|  |  |
| --- | --- |
| **Text Classification Applications** | **Illustration** |
| Sentiment Analysis | Identifying a product feedback either having positive, negative, or neutral sentiment. |
| Information Filtering | Selecting relevant information from a stream of text data. |
| Information Retrieval | Locating documents that comply with the information needed within large collections of documents. |
| Recommender System | Suggesting products to users based on its description and the user’s interests. |
| Document categorization | Detecting unwanted email as spam, categorizing a news article as politics, etc. |

The emerge of large amount of digital documents today makes this task more crucial than ever especially for companies to maximize their workflow or even profits.

Recently, the progress of NLP research on text classification has arrived at the state-of-the-art, showing Deep Learning methods as the cutting-edge technology to perform such task. Deep Learning is a subfield of Machine Learning techniques that uses layered representation of data known as Neural Networks and the word “deep” indicates a large or deep neural network. The increase of deep learning popularity nowadays simply because the models have achieved terrific results to perform many tasks in NLP, one of which is text classification.

The need to evaluate the performance of the state-of-the-art deep learning models for text classification is essential not only for academic purpose, but also for AI practitioners or professionals that need guidance and benchmark on similar projects.

## **Objectives and Scope**

Vaubviaubvpiaubv[aiuvb[aiuvba[viuba[v

## **Major Contribution of the Dissertation**

Sab;iabva;iubva’uvba[‘vba’vbv’vb’v

## **Organisation of the Dissertation**

Diluvabi;vubavdiuba[vdiub[vb

# **Chapter 2 Literature Review**

|  |
| --- |
| Then comes the main part of your work. To lay the ground, there should first be a chapter on what has been done before on the problem - a Literature Review. This is an important section because it shows that you do not narrowly focus only on what you do, but are aware of the related work elsewhere, some of which might be instructive to your solving the problem. It can also explain why you are taking the direction you do. |

Iubaiuba[ivuba[uvba[vuab[vab Dlviuabvdiauvb[aiubv[viuebav[ib[

## **Text Preprocessing**

Text classification is one of the most popular tasks in NLP that allows a program to classify free-text document based on pre-defined classes. The classes can be based on topic, genre, or sentiment. For example, a news article can be categorized as politics, sports, scientific, or entertainment; an email can be identified as spam or ham; a movie review can be judged having positive, negative, or neutral sentiment.

### **Text Cleaning**

### **Text Representation**

#### **Bag-of-Words (BoW)**

Text classification is one of the most popular tasks in NLP that allows a program to classify free-text document based on pre-defined classes. The classes can be based on topic, genre, or sentiment. For example, a news article can be categorized as politics, sports, scientific, or entertainment; an email can be identified as spam or ham; a movie review can be judged having positive, negative, or neutral sentiment.

#### **Word Embedding**

Text classification is one of the most popular tasks in NLP that allows a program to classify free-text document based on pre-defined classes. The classes can be based on topic, genre, or sentiment. For example, a news article can be categorized as politics, sports, scientific, or entertainment; an email can be identified as spam or ham; a movie review can be judged having positive, negative, or neutral sentiment.

## **Deep Learning Techniques**

Text classification is one of the most popular tasks in NLP that allows a program to classify free-text document based on pre-defined classes. The classes can be based on topic, genre, or sentiment. For example, a news article can be categorized as politics, sports, scientific, or entertainment; an email can be identified as spam or ham; a movie review can be judged having positive, negative, or neutral sentiment.

### **Feed Forward Neural Networks**

Text classification is one of the most popular tasks in NLP that allows a program to classify free-text document based on pre-defined classes. The classes can be based on topic, genre, or sentiment. For example, a news article can be categorized as politics, sports, scientific, or entertainment; an email can be identified as spam or ham; a movie review can be judged having positive, negative, or neutral sentiment.

#### **Multi-Layer Perceptrons (MLP)**

Text classification is one of the most popular tasks in NLP that allows a program to classify free-text document based on pre-defined classes. The classes can be based on topic, genre, or sentiment. For example, a news article can be categorized as politics, sports, scientific, or entertainment; an email can be identified as spam or ham; a movie review can be judged having positive, negative, or neutral sentiment.

#### **Ensemble Deep Random Vector Functional Link (edRVFL)**

Text classification is one of the most popular tasks in NLP that allows a program to classify free-text document based on pre-defined classes. The classes can be based on topic, genre, or sentiment. For example, a news article can be categorized as politics, sports, scientific, or entertainment; an email can be identified as spam or ham; a movie review can be judged having positive, negative, or neutral sentiment.

### **Recurrent Neural Networks**

#### **Standard RNN**

Text classification is one of the most popular tasks in NLP that allows a program to classify free-text document based on pre-defined classes. The classes can be based on topic, genre, or sentiment. For example, a news article can be categorized as politics, sports, scientific, or entertainment; an email can be identified as spam or ham; a movie review can be judged having positive, negative, or neutral sentiment.

#### **Gated Recurrent Unit (GRU)**

Text classification is one of the most popular tasks in NLP that allows a program to classify free-text document based on pre-defined classes. The classes can be based on topic, genre, or sentiment. For example, a news article can be categorized as politics, sports, scientific, or entertainment; an email can be identified as spam or ham; a movie review can be judged having positive, negative, or neutral sentiment.

#### **Long Short-Term Memory (LSTM)**

Text classification is one of the most popular tasks in NLP that allows a program to classify free-text document based on pre-defined classes. The classes can be based on topic, genre, or sentiment. For example, a news article can be categorized as politics, sports, scientific, or entertainment; an email can be identified as spam or ham; a movie review can be judged having positive, negative, or neutral sentiment.

### **Convolutional Neural Networks**

Text classification is one of the most popular tasks in NLP that allows a program to classify free-text document based on pre-defined classes. The classes can be based on topic, genre, or sentiment. For example, a news article can be categorized as politics, sports, scientific, or entertainment; an email can be identified as spam or ham; a movie review can be judged having positive, negative, or neutral sentiment.

#### **Standard CNN**

Text classification is one of the most popular tasks in NLP that allows a program to classify free-text document based on pre-defined classes. The classes can be based on topic, genre, or sentiment. For example, a news article can be categorized as politics, sports, scientific, or entertainment; an email can be identified as spam or ham; a movie review can be judged having positive, negative, or neutral sentiment.

#### **Temporal Convolutional Network (TCN)**

Text classification is one of the most popular tasks in NLP that allows a program to classify free-text document based on pre-defined classes. The classes can be based on topic, genre, or sentiment. For example, a news article can be categorized as politics, sports, scientific, or entertainment; an email can be identified as spam or ham; a movie review can be judged having positive, negative, or neutral sentiment.

### **Hybrid Neural Networks**

Text classification is one of the most popular tasks in NLP that allows a program to classify free-text document based on pre-defined classes. The classes can be based on topic, genre, or sentiment. For example, a news article can be categorized as politics, sports, scientific, or entertainment; an email can be identified as spam or ham; a movie review can be judged having positive, negative, or neutral sentiment.

## **Model Optimization and Evaluation**

### **Optimization**

Text classification is one of the most popular tasks in NLP that allows a program to classify free-text document based on pre-defined classes. The classes can be based on topic, genre, or sentiment. For example, a news article can be categorized as politics, sports, scientific, or entertainment; an email can be identified as spam or ham; a movie review can be judged having positive, negative, or neutral sentiment.

### **Evaluation**

Text classification is one of the most popular tasks in NLP that allows a program to classify free-text document based on pre-defined classes. The classes can be based on topic, genre, or sentiment. For example, a news article can be categorized as politics, sports, scientific, or entertainment; an email can be identified as spam or ham; a movie review can be judged having positive, negative, or neutral sentiment.

# **Chapter 3 Deep Learning Techniques for Text Classification**

## **Datasets and Experimental Setup**

### **Datasets**

Text classification is one of the most popular tasks in NLP that allows a program to classify free-text document based on pre-defined classes. The classes can be based on topic, genre, or sentiment. For example, a news article can be categorized as politics, sports, scientific, or entertainment; an email can be identified as spam or ham; a movie review can be judged having positive, negative, or neutral sentiment.

### **Experimental Setup**

Text classification is one of the most popular tasks in NLP that allows a program to classify free-text document based on pre-defined classes. The classes can be based on topic, genre, or sentiment. For example, a news article can be categorized as politics, sports, scientific, or entertainment; an email can be identified as spam or ham; a movie review can be judged having positive, negative, or neutral sentiment.

## **Bag-of-Words based Models**

Text classification is one of the most popular tasks in NLP that allows a program to classify free-text document based on pre-defined classes. The classes can be based on topic, genre, or sentiment. For example, a news article can be categorized as politics, sports, scientific, or entertainment; an email can be identified as spam or ham; a movie review can be judged having positive, negative, or neutral sentiment.

### **Multi-Layer Perceptron**

Text classification is one of the most popular tasks in NLP that allows a program to classify free-text document based on pre-defined classes. The classes can be based on topic, genre, or sentiment. For example, a news article can be categorized as politics, sports, scientific, or entertainment; an email can be identified as spam or ham; a movie review can be judged having positive, negative, or neutral sentiment.

### **Ensemble Deep Random Vector Functional Link**

Text classification is one of the most popular tasks in NLP that allows a program to classify free-text document based on pre-defined classes. The classes can be based on topic, genre, or sentiment. For example, a news article can be categorized as politics, sports, scientific, or entertainment; an email can be identified as spam or ham; a movie review can be judged having positive, negative, or neutral sentiment.

## **Word Embedding / Word2Vec based Models**

Text classification is one of the most popular tasks in NLP that allows a program to classify free-text document based on pre-defined classes. The classes can be based on topic, genre, or sentiment. For example, a news article can be categorized as politics, sports, scientific, or entertainment; an email can be identified as spam or ham; a movie review can be judged having positive, negative, or neutral sentiment.

### **Multi Layer Perceptron**

Text classification is one of the most popular tasks in NLP that allows a program to classify free-text document based on pre-defined classes. The classes can be based on topic, genre, or sentiment. For example, a news article can be categorized as politics, sports, scientific, or entertainment; an email can be identified as spam or ham; a movie review can be judged having positive, negative, or neutral sentiment.

### **Ensemble Deep Random Vector Functional Link**

Text classification is one of the most popular tasks in NLP that allows a program to classify free-text document based on pre-defined classes. The classes can be based on topic, genre, or sentiment. For example, a news article can be categorized as politics, sports, scientific, or entertainment; an email can be identified as spam or ham; a movie review can be judged having positive, negative, or neutral sentiment.

### **Convolutional Neural Networks**

Text classification is one of the most popular tasks in NLP that allows a program to classify free-text document based on pre-defined classes. The classes can be based on topic, genre, or sentiment. For example, a news article can be categorized as politics, sports, scientific, or entertainment; an email can be identified as spam or ham; a movie review can be judged having positive, negative, or neutral sentiment.

### **Temporal Convolutional Neural Networks**

Text classification is one of the most popular tasks in NLP that allows a program to classify free-text document based on pre-defined classes. The classes can be based on topic, genre, or sentiment. For example, a news article can be categorized as politics, sports, scientific, or entertainment; an email can be identified as spam or ham; a movie review can be judged having positive, negative, or neutral sentiment.

### **Gated Recurrent Unit**

Text classification is one of the most popular tasks in NLP that allows a program to classify free-text document based on pre-defined classes. The classes can be based on topic, genre, or sentiment. For example, a news article can be categorized as politics, sports, scientific, or entertainment; an email can be identified as spam or ham; a movie review can be judged having positive, negative, or neutral sentiment.

### **Long Short-Term Memory**

Text classification is one of the most popular tasks in NLP that allows a program to classify free-text document based on pre-defined classes. The classes can be based on topic, genre, or sentiment. For example, a news article can be categorized as politics, sports, scientific, or entertainment; an email can be identified as spam or ham; a movie review can be judged having positive, negative, or neutral sentiment.

### **Hybrid (CNN + LSTM)**

Text classification is one of the most popular tasks in NLP that allows a program to classify free-text document based on pre-defined classes. The classes can be based on topic, genre, or sentiment. For example, a news article can be categorized as politics, sports, scientific, or entertainment; an email can be identified as spam or ham; a movie review can be judged having positive, negative, or neutral sentiment.

# **Chapter 4 Evaluation**

## **Results**

Text classification is one of the most popular tasks in NLP that allows a program to classify free-text document based on pre-defined classes. The classes can be based on topic, genre, or sentiment. For example, a news article can be categorized as politics, sports, scientific, or entertainment; an email can be identified as spam or ham; a movie review can be judged having positive, negative, or neutral sentiment.

## **Discussion**

Text classification is one of the most popular tasks in NLP that allows a program to classify free-text document based on pre-defined classes. The classes can be based on topic, genre, or sentiment. For example, a news article can be categorized as politics, sports, scientific, or entertainment; an email can be identified as spam or ham; a movie review can be judged having positive, negative, or neutral sentiment.

# **Chapter 5 Conclusions and Future Work**

## **Conclusions**

Text classification is one of the most popular tasks in NLP that allows a program to classify free-text document based on pre-defined classes. The classes can be based on topic, genre, or sentiment. For example, a news article can be categorized as politics, sports, scientific, or entertainment; an email can be identified as spam or ham; a movie review can be judged having positive, negative, or neutral sentiment.

## **Recommendation in Future Work**

Text classification is one of the most popular tasks in NLP that allows a program to classify free-text document based on pre-defined classes. The classes can be based on topic, genre, or sentiment. For example, a news article can be categorized as politics, sports, scientific, or entertainment; an email can be identified as spam or ham; a movie review can be judged having positive, negative, or neutral sentiment.