

APPLICATION OF DEEP LEARNING AND  
NATURAL LANGUAGE PROCESSING FOR  
SUSTAINABLE PROCESS DEVELOPMENT

AVAN KUMAR



DEPARTMENT OF CHEMICAL ENGINEERING  
INDIAN INSTITUTE OF TECHNOLOGY DELHI

MAY 2024

© **Indian Institute of Technology Delhi (IITD), New Delhi, 2024**

# Application of Deep Learning and Natural Language Processing for Sustainable Process Development

*by*

**Avan Kumar**

**Department of Chemical Engineering**

*Submitted*

*in fulfilment of the requirements of the degree of*

**Doctor of Philosophy**

*to the*



**INDIAN INSTITUTE OF TECHNOLOGY DELHI**

**May 2024**

# THESIS CERTIFICATE

This is to certify that the thesis titled **Application of Deep Learning and Natural Language Processing for Sustainable Process Development**, submitted by **Avan Kumar (2019CHZ8155)**, to the Indian Institute of Technology Delhi, for the award of the degree of **Doctor of Philosophy**, is a bonafide record of the research work done by him under my supervision. The contents of this thesis, in full or in parts, have not been submitted to any other Institute or University for the award of any degree or diploma.

**Prof. Hariprasad Kodamana**  
**Research Supervisor**

Dept. of Chemical Engineering &  
Yardi School of Artificial In-  
telligence, Indian Institute of  
Technology-Delhi, Hauzkhas,  
Delhi-110016

Date: May 01, 2024

## ACKNOWLEDGEMENTS

Foremost, I would like to express my sincere gratitude to my advisor, Prof. Hariprasad Kodamana, for his continuous support during my Ph.D. study and for his patience, motivation, enthusiasm, and immense knowledge. His guidance, timely advice, and scientific approach are a source of inspiration and helped me in all the time of research and writing of this thesis. His dedication and keen interest and, above all, his overwhelming attitude to always help his students had been mainly responsible for completing this thesis.

Besides my advisor, I would like to thank my research committee members, Prof. Manojkumar C. Ramteke, Prof. N. M. Anoop Krishnan, and Prof. Gaurav Goel, for their encouragement and insightful comments. I especially want to thank Prof. Bhavik R. Bakshi, Arizona State University (ASU), United States, for his valuable input, vast knowledge, and motivating guidance in this project. Many thanks to Prof. Sreedevi Upadhyayula and Prof. K. K. Pant for participating in my research by providing invaluable feedback and making this work possible.

In addition, I owe my thanks to my current fellow labmates and collaborators, Nikita, Jyoti, Umang, Ahtesham, Nasre, Deepak, Arjun, Harshitha, Vinayak, Pushp, and all of my past labmates, Tanuja, Reena, Divyanshi, Devansh, Abhyansh, and Mayuna in the Control & Analytics of Process Systems (CAPS) Lab, IITD for the stimulating discussions and the positive environment they have created in the lab. Furthermore, I would like to express my sincere thanks to all the office staff members of the Chemical Engineering Department, IITD.

Last but not least, I would like to thank the almighty god, my parents, family members, and friends for their unconditional love, prayers, and constant support throughout my life.



# ABSTRACT

The concentration of air pollutants and solid waste plastic is increasing, creating many severe problems for humankind. There are many causes for their generation, such as transportation, industries, household activities, etc. In this thesis, we try to address these problems and showcase some applications of deep learning and natural language processing tools that would help in making decision-support systems to keep the environment clean and sustainable. Specifically, we target the following application areas: harnessing solar energy, generating alternative green fuel methanol and hydrogen, and recycling waste plastic. In finding the solution, we also need to focus on the novel conversion process, where waste could be transformed into value-added products, e.g., alternative fuels and laboratory/industrial chemicals that can lower the dependency on non-renewable energy sources for developing sustainable systems. The work done in this thesis can be broadly segregated into two major parts: first, a data-driven approach for modeling, and second, data mining for extracting the information for designing decision support systems.

In the first part, a data-driven approach has been implemented to enhance the photocatalytic synthesis reactions and methanol production from syngas. Solar energy is abundant in nature, and it can be utilized easily. A recent trend in chemical synthesis is photo-catalysis, which uses photo-active catalyst materials such as semiconductor materials. Its well-known electronic property is the band gap. Herein, we propose an integrated deep learning-based framework to classify the photo-active catalysts and predict their band gap using compositional features. It helps rationalize the synthesis of photo-active catalysts as an initial screening parameter. We propose (i) the 2D-CNN model as classification with an accuracy of 0.886 and (ii) prediction of a band gap using 1D-VGG +XGBoost regressor with  $R^2$  0.750 for the test dataset. This framework provides preliminary information on catalysts, such as a band gap value. This prior knowledge about any catalyst material will accelerate the designing and synthesis procedures. Further, this photo-catalytic reactor can be utilized in methanol production. It has gained considerable interest on the laboratory and industrial scale as it is a renewable fuel and an excellent hydrogen energy storehouse. The methanol

---

synthesis process has been modeled, optimized, and proposed as an interpretable Gaussian Process Regression (GPR) regressor and Multi-objective Objective Bayesian Optimization (MOBO), respectively. Our trained GPR models have  $R^2$  values in the range of 0.97 to 1.00, irrespective of predicting variables and datasets. We have interpreted the models using "SHAP" technique and found the most important input features, as follows: inlet mole fraction of  $CO$  ( $Y(CO, in)$ ) and net inlet flow rate ( $Fin(nl/min)$ ). GPR and MOBO's performance was excellent compared to other machine learning models. The computational time to train the above-mentioned ML/DL models is approximately 5-10 minutes for each run. The configuration utilized by the machine is as follows: 16GB RAM, 4GB GPU Card.

The second part of the thesis aims to extract crucial information from available literature using Natural Language Process (NLP) tools that build knowledge extractor language models to help make rational decisions. In Hydrogen production, we developed a deep learning model, Extend-SciBERT (Ex-SciBERT), which extracts the process and catalyst information from the literature. It has two layers of screen-like classification of sentences and is followed by a NER task with accuracy values of 0.890 and 0.997 for the test dataset, respectively. With the same literature data,  $H_2 - BERT$ , is also developed that extracts the catalyst information using the task of Q&A with an accuracy of 0.823 for random abstracts. Such knowledge would be crucial for enhancing Hydrogen production as it is a renewable fuel that helps in driving towards a sustainable environment.

Similarly, waste plastic recycling technologies are also a supporting factor in making sustainable and circular system systems. A vast amount of knowledge about waste plastic recycling is stored in literature. We develop an NLP model, namely, Recycle-BERT, that employs fine-tuned BERT models, including Class-BERT for screening and a Q&A module, and four models specifying reactant, methodology, recycled product, and catalyst details for efficient recycling information extraction. The classification model achieved an accuracy of 0.974 on the test dataset. For the Q&A task, F1-score values for Catalyst-BERT, Method-BERT, Reactant-BERT, and Product-BERT were 0.7646, 0.8014, 0.8221, and 0.8512, respectively. Recycle-BERT performed well and extracted reliable information from the literature.

As an extension, an NLP-based framework has been proposed that processes the recycling of polyethylene terephthalate (PET) plastic waste-based literature to study the evo-

---

lution of recycling technologies and methodologies. It comprises the three approaches as follows: (i) Time Series Knowledge Graphs (TSKGs), which extract the three pieces of information such as reactant, technology name, and product formed, and shown in knowledge graph; (ii) Dynamic Transformer-based Topic Modeling (DTTM), which is to model the topics using a transformer based approach and followed by series of steps to extract the methodologies of PET-based plastic waste recycling, (iii) which are further quantified using popularity index calculation. With it, one can quickly identify and select an efficient and suitable recycling method. This combined study shows the potential to choose the efficient and trending approach in the recycling field. The computational time to train all above-mentioned transformer-based models is at least 4-5 hours for each run. The configuration utilized by the machine is as follows: 32GB RAM, 32GB GPU Card. In short, the aforementioned techniques and created models demonstrate their capacity to contribute to the establishment of sustainable processes and environments in a complete manner.

KEYWORDS: data-driven models; text mining; deep learning; natural language processing; solar energy; photo-active catalyst; methanol; hydrogen production; transformer; Ex-SciBERT; sustainable processes; waste plastic recycling; Recycle-BERT; TSKGs; DTTM, PI

## सारांश

वायु प्रदूषकों और ठोस अपशिष्ट प्लास्टिक की सांद्रता बढ़ रही है, जिससे कई गंभीर स्थितियां पैदा हो रही हैं मानव जाति के लिए समस्याएँ. इनके उत्पन्न होने के कई कारण हैं, जैसे परिवहन-व्यवसाय, उद्योग, घरेलू गतिविधियाँ, आदि। इस थीसिस में, हम इन समस्याओं का समाधान करने का प्रयास करते हैं और गहन शिक्षण और प्राकृतिक भाषा प्रसंस्करण उपकरणों के कुछ अनुप्रयोगों का प्रदर्शन करेंगे पर्यावरण को स्वच्छ और सतत बनाए रखने के लिए निर्णय-समर्थन प्रणाली बनाने में मदद मिलेगी धारणीय. विशेष रूप से, हम निम्नलिखित अनुप्रयोग क्षेत्रों को लक्षित करते हैं: सौर ऊर्जा का दोहन, वैकल्पिक हरित ईंधन मेथनॉल और हाइड्रोजन का उत्पादन, और अपशिष्ट प्लास्टिक का पुनर्चक्रण। में समाधान ढूंढते हुए, हमें नवीन रूपांतरण प्रक्रिया पर भी ध्यान केंद्रित करने की आवश्यकता है, जहां बर्बादी हो सकती है मूल्य-वर्धित उत्पादों, जैसे वैकल्पिक ईंधन और प्रयोगशाला/औद्योगिक में परिवर्तित किया जा सकता है ऐसे रसायन जो विकास के लिए गैर-नवीकरणीय ऊर्जा स्रोतों पर निर्भरता को कम कर सकते हैं टिकाऊ प्रणालियाँ। इस थीसिस में किए गए कार्य को मोटे तौर पर दो प्रमुख भागों में विभाजित किया जा सकता है भाग: पहला, मॉडलिंग के लिए डेटा-संचालित दृष्टिकोण, और दूसरा, निकालने के लिए डेटा माइनिंग निर्णय समर्थन प्रणाली डिज़ाइन करने के लिए जानकारी।

पहले भाग में, फोटोकैटलिटिक को बढ़ाने के लिए डेटा-संचालित दृष्टिकोण लागू किया गया है संश्लेषण प्रतिक्रियाएं और सिनगैस से मेथनॉल उत्पादन। सौर ऊर्जा प्रचुर मात्रा में है प्रकृति, और इसका उपयोग आसानी से किया जा सकता है। रासायनिक संश्लेषण में एक हालिया प्रवृत्ति फोटो-कैटलिसिस है, जो अर्धचालक सामग्री जैसे फोटो-सक्रिय उत्प्रेरक सामग्री का उपयोग करता है। यह सर्वविदित है इलेक्ट्रॉनिक संपत्ति बैंड गैप है। इसमें, हम एक एकीकृत गहन शिक्षण-आधारित का प्रस्ताव करते हैं फोटो-सक्रिय उत्प्रेरकों को वर्गीकृत करने और कंपो का उपयोग करके उनके बैंड गैप की भविष्यवाणी करने की रूपरेखा स्थितिगत विशेषताएँ. यह प्रारंभिक रूप से फोटो-सक्रिय उत्प्रेरक के संश्लेषण को तर्कसंगत बनाने में मदद करता है स्क्रीनिंग पैरामीटर. हम सटीकता के साथ वर्गीकरण के रूप में (i) 2डी-सीएनएन मॉडल का प्रस्ताव करते हैं 0.886 का और (ii) आर2 0.750 के साथ 1डी-वीजीजी +एक्सजीबूस्ट रिग्रेसर का उपयोग करके बैंड गैप की भविष्यवाणी परीक्षण डेटासेट के लिए. यह ढाँचा उत्प्रेरकों पर प्रारंभिक जानकारी प्रदान करता है, जैसे बैंड गैप मान के रूप में। किसी भी उत्प्रेरक सामग्री के बारे में यह पूर्व ज्ञान गति प्रदान करेगा डिजाइनिंग और संश्लेषण प्रक्रियाएं। इसके अलावा, इस फोटो-कैटलिटिक रिएक्टर का उपयोग किया जा सकता है मेथनॉल उत्पादन. इसने प्रयोगशाला और औद्योगिक पैमाने पर काफी रुचि प्राप्त की है क्योंकि यह एक नवीकरणीय ईंधन और एक उत्कृष्ट हाइड्रोजन ऊर्जा भंडारगृह है। मेथनॉल

संश्लेषण प्रक्रिया को एक व्याख्या योग्य गाऊसी के रूप में मॉडलिंग, अनुकूलित और प्रस्तावित किया गया है प्रोसेस रिग्रेसन (जीपीआर) रिग्रेसर और मल्टी-ऑब्जेक्टिव ऑब्जेक्टिव बायेसियन ऑप्टिमाइज़ेशन (MOBO), क्रमशः। हमारे प्रशिक्षित जीपीआर मॉडल में आर<sup>2</sup> मान 0.97 से की सीमा में हैं 1.00, चर और डेटासेट की भविष्यवाणी के बावजूद। हमने "SHAP" तकनीक का उपयोग करके मॉडलों की व्याख्या की है और सबसे महत्वपूर्ण इनपुट विशेषताएं पाई हैं, जो इस प्रकार हैं: इनलेट मोल CO का अंश (Y(CO, in)) और शुद्ध इनलेट प्रवाह दर (F in(nl/min))। जीपीआर और एमओबीओ अन्य मशीन लर्निंग मॉडल की तुलना में प्रदर्शन उत्कृष्ट था। कम्प्यूटेशनल उपर्युक्त एमएल/डीएल मॉडल को प्रशिक्षित करने का समय प्रत्येक के लिए लगभग 5-10 मिनट है दौड़ना। मशीन द्वारा उपयोग किया जाने वाला कॉन्फ़िगरेशन इस प्रकार है: 16 जीबी रैम, 4 जीबी जीपीयू कार्ड।

थीसिस के दूसरे भाग का उद्देश्य उपलब्ध साहित्यिक सामग्री से महत्वपूर्ण जानकारी निकालना है- प्राकृतिक भाषा प्रक्रिया (एनएलपी) टूल का उपयोग करके ज्ञान निकालने वाली भाषा का निर्माण करें तर्कसंगत निर्णय लेने में मदद करने वाले मॉडल। हाइड्रोजन उत्पादन में, हमने एक गहरा विकास किया लर्निंग मॉडल, एक्सटेंड-साइबर्ट (एक्स-साइबर्ट), जो प्रक्रिया और उत्प्रेरक को निकालता है साहित्य से जानकारी। इसमें वाक्यों के स्क्रीन-जैसे वर्गीकरण की दो परतें हैं और इसके बाद परीक्षण डेटासेट के लिए 0.890 और 0.997 के सटीकता मानों के साथ एक एनईआर कार्य किया जाता है, क्रमशः। उसी साहित्य डेटा के साथ, H2 -BERT, भी विकसित किया गया है जो निकालता है यादृच्छिक सार के लिए 0.823 की सटीकता के साथ प्रश्नोत्तर के कार्य का उपयोग करके उत्प्रेरक जानकारी। ऐसा ज्ञान हाइड्रोजन उत्पादन को बढ़ाने के लिए महत्वपूर्ण होगा क्योंकि यह नवीकरणीय है ईंधन जो टिकाऊ पर्यावरण की ओर बढ़ने में मदद करता है।

इसी तरह, अपशिष्ट प्लास्टिक रीसाइक्लिंग प्रौद्योगिकियां भी बनाने में सहायक कारक हैं टिकाऊ और परिपत्र प्रणाली प्रणाली। अपशिष्ट प्लास्टिक पुनर्चक्रण के बारे में विशाल मात्रा में ज्ञान साहित्य में संग्रहीत है। हम एक एनएलपी मॉडल विकसित करते हैं, जिसका नाम है, रीसायकल-बीईआरटी, यह स्क्रीनिंग और प्रश्नोत्तरी के लिए क्लास-बीईआरटी सहित परिष्कृत बीईआरटी मॉडल का उपयोग करता है मॉड्यूल, और अभिकारक, कार्यप्रणाली, पुनर्नवीनीकरण उत्पाद और उत्प्रेरक को निर्दिष्ट करने वाले चार मॉडल कुशल पुनर्चक्रण सूचना निष्कर्षण के लिए विवरण। वर्गीकरण मॉडल ने एक हासिल किया परीक्षण डेटासेट पर 0.974 की सटीकता। प्रश्नोत्तर कार्य के लिए, कैटलिस्ट-बीईआरटी, मेथड-बीईआरटी, रिएक्टेंट-बीईआरटी और उत्पाद-बीईआरटी के लिए एफ1-स्कोर मान 0.7646, 0.8014, 0.8221 थे। और 0.8512, क्रमशः। रीसायकल-बीईआरटी ने अच्छा प्रदर्शन किया और विश्वसनीय जानकारी निकाली साहित्य से।

विस्तार के रूप में, एक एनएलपी-आधारित ढांचा प्रस्तावित किया गया है जो रीसाइक्लिंग प्रौद्योगिकियों और कार्यप्रणाली के विकास का अध्ययन करने के लिए पॉलीथिन टेरेफ्थैलेट (पीईटी) प्लास्टिक अपशिष्ट-आधारित साहित्य के रीसाइक्लिंग की प्रक्रिया करता है। इसमें तीन दृष्टिकोण शामिल हैं इस प्रकार है: (i) टाइम सीरीज़ नॉलेज ग्राफ़ (टीएसकेजी), जो जानकारी के तीन टुकड़े जैसे कि अभिकारक, प्रौद्योगिकी का नाम और निर्मित उत्पाद निकालते हैं, और ज्ञान में दिखाए जाते हैं ग्राफ़; (ii) डायनेमिक ट्रांसफार्मर-आधारित टॉपिक मॉडलिंग (डीटीटीएम), जो कि मॉडल बनाना है ट्रांसफॉर्मर आधारित दृष्टिकोण का उपयोग करके विषयों को निकालने के लिए चरणों की श्रृंखला का पालन किया जाता है पीईटी-आधारित प्लास्टिक अपशिष्ट पुनर्चक्रण की पद्धतियाँ, (iii) जिनका उपयोग करके आगे मात्रा निर्धारित की जाती है लोकप्रियता सूचकांक गणना। इसके साथ, कोई भी व्यक्ति शीघ्रता से एक कुशल और का चयन कर सकता है उपयुक्त पुनर्चक्रण विधि। यह संयुक्त अध्ययन कुशल को चुनने की क्षमता दर्शाता है और रीसाइक्लिंग क्षेत्र में ट्रेडिंग दृष्टिकोण। उपरोक्त सभी को प्रशिक्षित करने का कम्प्यूटेशनल समय-उल्लिखित ट्रांसफार्मर-आधारित मॉडल प्रत्येक रन के लिए कम से कम 4-5 घंटे का है। विन्यास मशीन द्वारा उपयोग इस प्रकार है: 32 जीबी रैम, 32 जीबी जीपीयू कार्ड। संक्षेप में, उपरोक्त तकनीकें और बनाए गए मॉडल योगदान देने की उनकी क्षमता को प्रदर्शित करते हैं संपूर्ण तरीके से स्थायी प्रक्रियाओं और वातावरण की स्थापना।

**कीवर्ड:** डेटा-संचालित मॉडल; टेक्स्ट खनन; ध्यान लगा के पढ़ना या सीखना; प्राकृतिक भाषा प्रसंस्करण; सौर ऊर्जा; फोटो-सक्रिय उत्प्रेरक; मेथनॉल; हाइड्रोजन उत्पादन; ट्रांसफार्मर; पूर्व-साइबर्ट; टिकाऊ प्रक्रियाएं; अपशिष्ट प्लास्टिक का पुनर्चक्रण; रीसायकल-बर्ट; टीएसकेजी; डीटीटीएम, पीआई

# Contents

<b>ACKNOWLEDGEMENTS</b>	<b>i</b>
<b>ABSTRACT</b>	<b>ii</b>
<b>LIST OF FIGURES</b>	<b>xv</b>
<b>LIST OF TABLES</b>	<b>xviii</b>
<b>ABBREVIATIONS</b>	<b>xix</b>
<b>NOTATION</b>	<b>xxi</b>
<b>1 INTRODUCTION</b>	<b>1</b>
1.1 Motivation . . . . .	1
1.2 Research Contributions . . . . .	5
1.3 Outline of the Thesis . . . . .	8
<b>2 Literature Review</b>	<b>9</b>
2.1 Utilization of solar energy by photo-catalytic reactor systems . . . . .	9
2.2 Conversion of $CO_2$ into value-added chemical supportive System . . . . .	14
2.3 Text mining using natural language processing tools . . . . .	17
2.3.1 Natural Language Processing . . . . .	17
2.3.2 Architecture of transformer-based models: BERT, SciBERT . . . . .	18
2.4 Knowledge Retrieval System for Hydrogen Production . . . . .	20
2.5 Knowledge Management Systems for Plastic Recycling Waste . . . . .	23
2.5.1 Waste plastic recycling . . . . .	23
2.5.2 Language Models for Recycling . . . . .	24
2.5.3 A study on a specific waste plastic recycling . . . . .	27
2.6 Research gap based on literature review . . . . .	31

---

<b>3</b>	<b>A Convolutional Neural Network-based gradient boosting framework for prediction of the band gap of photo-active catalysts</b>	<b>32</b>
3.1	Data Collection, Annotation, and Feature Engineering . . . . .	34
3.1.1	Data Collection . . . . .	34
3.1.2	Data Annotation . . . . .	34
3.1.3	Feature Engineering . . . . .	34
3.2	Classification of Photo-catalysts and Prediction of Band Gap . . . . .	36
3.2.1	Classification of photo-active catalysts using CNN . . . . .	36
3.2.2	Prediction of band gap using Convolutional Neural Network-based gradient boosting . . . . .	37
3.3	Results and Discussion . . . . .	38
3.3.1	Classification . . . . .	38
3.3.2	Prediction of Band gap . . . . .	40
3.4	Conclusion and Future Work . . . . .	42
<b>4</b>	<b>A multi-objective Bayesian optimization framework for the synthesis of methanol from syngas using Gaussian process models</b>	<b>45</b>
4.1	Datasets and pre-processing . . . . .	47
4.2	Methods: GPR modeling, SHAP, Multi-objective Bayesian optimization . . . . .	48
4.2.1	Gaussian Process Regression . . . . .	49
4.2.2	Interpreting the GPR prediction model using SHAP . . . . .	52
4.2.3	Multi-objective Bayesian Optimization (MOBO) . . . . .	54
4.3	Results and Discussion . . . . .	55
4.3.1	Recycle Reactor . . . . .	55
4.3.2	Once-Through Reactor . . . . .	56
4.4	Conclusion . . . . .	64
<b>5</b>	<b>A text mining framework for screening catalysts and critical process parameters from scientific literature - a study on Hydrogen production from alcohol</b>	<b>65</b>
5.1	Data collection and pre-processing . . . . .	67
5.1.1	Data extraction . . . . .	67
5.1.2	Data annoatation . . . . .	68

5.2	Text mining tools . . . . .	72
5.2.1	LDA . . . . .	72
5.2.2	BERT & SciBERT . . . . .	73
5.3	The proposed model: Ex-SciBERT & $H_2 - BERT$ . . . . .	73
5.3.1	Transfer learning framework . . . . .	75
5.4	Result and Discussion . . . . .	76
5.4.1	Topic modeling using LDA . . . . .	76
5.4.2	Sentence classification and NER using Ex-SciBERT . . . . .	79
5.4.3	Abstract classification and Q&A using $H_2 - BERT$ . . . . .	81
5.5	Conclusion . . . . .	82
<b>6</b>	<b>Recycle-BERT: Extracting Knowledge about Plastic Waste Recycling by Natural Language Processing</b>	<b>87</b>
6.1	Corpus Preparation and Annotation . . . . .	89
6.1.1	Text corpus generation . . . . .	89
6.1.2	Text annotation . . . . .	89
6.2	The proposed LM: Recycle-BERT . . . . .	93
6.3	Results and Discussion . . . . .	96
6.3.1	Results of Recycle-BERT's training and testing . . . . .	96
6.3.2	Results of Recycle-BERT's performance in information extraction . . . . .	99
6.4	Conclusions and Future Work . . . . .	109
6.5	Data and Code Availability . . . . .	110
<b>7</b>	<b>An evolutionary study on technologies for PET waste recycling using natural language processing</b>	<b>111</b>
7.1	Data collection and Pre-processing . . . . .	113
7.2	Proposed methodologies: TSKGs, DTTM, PI . . . . .	113
7.2.1	Time Series Knowledge Graphs Framework (TSKGs) . . . . .	115
7.2.2	Dynamic Transformer-based Topic Model (DTTM) . . . . .	118
7.2.3	Retrieved Information Quantification: Popularity Index . . . . .	120
7.3	Results and Discussion . . . . .	122
7.4	Conclusive Remarks . . . . .	133

<b>8</b>	<b>Conclusion and Future directions</b>	<b>135</b>
8.1	Summary and Conclusions . . . . .	135
8.2	Future Directions . . . . .	137
	<b>BIBLIOGRAPHY</b>	<b>137</b>
	<b>APPENDIX</b>	<b>165</b>
<b>A</b>	<b>A multi-objective Bayesian optimization framework for the synthesis of methanol from syngas using interpretable Gaussian process models</b>	<b>166</b>
A.1	The description of all trained supervised machine learning algorithms . . . .	166
A.2	The results of Gaussian process regression with different sets of kernels . . .	168
A.3	The information for both datasets utilized in this work . . . . .	172
<b>B</b>	<b>Recycle-BERT: Extracting Knowledge about Plastic Waste Recycling by Natural Language Processing</b>	<b>173</b>
B.1	This is the supplemental data for the Recycle-BERT project proposal . . . .	173
B.1.1	Four individual word cloud plots . . . . .	173
B.1.2	Zipf plots for odd last two years . . . . .	173
B.1.3	Performance of four individual Q&A models for two random abstracts . . . . .	173
B.2	The scalability testing of our Recycle-BERT on random abstracts . . . . .	174
B.2.1	Performance of sub-model of Recycle-BERT module, Class-BERT, for the Classification task . . . . .	178
B.2.2	The Zipf and word cloud plots of answer entities . . . . .	179
B.2.3	The performance of complete Question and Answer module (Recycle-BERT) . . . . .	180
<b>C</b>	<b>An evolutionary study on technologies for PET waste recycling using natural language processing</b>	<b>211</b>
C.1	This is the supplemental data for the proposed NLP framework . . . . .	211
C.1.1	TSKGS: Time Series Knowledge Graphs . . . . .	211
C.1.2	DTTM: Dynamic Transformer Topic Modeling . . . . .	211
	<b>LIST OF PUBLICATIONS</b>	<b>234</b>
	<b>Curriculum Vitae</b>	<b>234</b>

## List of Figures

2.1	A complete architecture of TransformerXiao and Zhu (2023) . . . . .	19
2.2	An example of knowledge graph for medical domain Huang <i>et al.</i> (2021) . .	30
3.1	A schematic diagram of the proposed framework: a convolutional neural network-based gradient boosting framework . . . . .	33
3.2	An illustration of data extraction and feature generation, each step is shown for data extraction to the final data set with 90 features. . . . .	35
3.3	A complete schematic diagram of the proposed methodology in this chapter.	38
3.4	The performance of 2D CNN classification model, (a) and (b) confusion matrix for train and test, (c) ROC curve plot with AUC for train and test in legend and at last (d) learning curves of metrics (loss, accuracy) per epoch are shown. . . . .	41
3.5	The overall performance of proposed framework, (a) variation of loss per estimator for task regression, (b) predicted and actual value scatter plot for photo-active band gap values. . . . .	43
3.6	The effectiveness of the suggested framework as a whole, (a) a multiple bars plot, where train and test metrics values are compared, (b) variation of performance of the model with per 10% increment in training data. . . . .	43
4.1	A schematic diagram of the proposed framework: A convolutional neural network-based gradient boosting algorithm. . . . .	47
4.2	Distribution of all descriptors represented in the form of box plots, where (a) and (b) represent recycle and once-through reactor datasets, respectively. .	48
4.3	Prediction results of recycle reactor models presented as scatter plots. Measured and predicted selectivity along with $R^2$ values (a) GPR with rational quadratic kernel shown with a standard deviation around the mean (b) combined results of selectivity prediction of all ML models. Measured and predicted conversion (c) GPR rational quadratic kernel with a standard deviation around the mean (d) combined results of all models used for conversion prediction. The other ML models used for comparison purposes are LR, LOR, RR, ELR, PCR, PLSR, SVR, and MLP. . . . .	57
4.4	$R^2$ metric of the test set for GPR performance with different kernels is shown in bar plots for selectivity, and conversion parameters are displayed in (a) and (b), respectively. . . . .	58

4.5	Prediction results of once-through reactor models presented as scatter plots. Measured and predicted selectivity along with $R^2$ values (a) GPR with a rational quadratic kernel with a standard deviation around a mean (b) combined results of selectivity prediction of all ML models. For conversion (c), GPR rational quadratic kernel with a standard deviation around the mean (d) combined results of all models used for conversion prediction. The other ML models used for comparison purposes are LR, LOR, RR, ELR, PCR, PLSR, SVR, and MLP. . . . .	59
4.6	The evaluated mean $R^2$ of Train and Test and mean RMSE of Train and Test are shown in (a), (b), (c), and (d), where (a) and (b) stand for selectivity and conversion in recycle reactor respectively. The bar plots (c) and (d) stand for selectivity and conversion in the once-through reactor, respectively. . . . .	60
4.7	The summary plots of shap result of (a) conversion (b) selectivity for recycle reactor . . . . .	61
4.8	The summary plots of SHAP result of (a) conversion (b) selectivity for once through reactor . . . . .	62
4.9	The value of objective functions w.r.t. weight, where Figure (a) and Figure (b) stand for the results of the recycle reactor and once through, respectively. . . . .	63
5.1	A schematic of the entire pipeline followed in this chapter . . . . .	66
5.2	A Schematic diagram of the following steps, where (a) signifies the steps for text extraction, (b) stands for data annotation, and (c) shows the transfer learning implementation on the BERT model. . . . .	67
5.3	Schematic of the strides of text extraction . . . . .	69
5.4	Data annotation pipeline . . . . .	69
5.5	Data annotation for classification and NER . . . . .	69
5.6	The plots represent the catalyst frequency that exists in the data by employing (a) Zipf plot, (b) histogram plot, and (c) a catalyst cloud image of all chemical entities . . . . .	71
5.7	Schematic representing the transformation of BERT to Ex-SciBERT model . . . . .	74
5.8	The architecture of $Ex-SciBERT$ based on the $BERT_{base}$ architecture with 12 layers of encoders, a description of encoder . . . . .	74
5.9	SciBERT and Ex-SciBERT . . . . .	74
5.10	The histogram plots and statistical analysis of some key parameters are shown: (a) Number of sentences, (b) Average length of words, (c) Number of words, and (d) Count of stop words present in each abstract with respective mean of the distribution . . . . .	77

5.11	The results of LDA are represented as (a) a bar plot of the number of documents v/s topic. (b) the figure shows a variation of the Coherence score with topics for different values of $\alpha$ and $\beta$ (scatter). The line plots variation for $\alpha = 0.61$ and $\beta = 0.31$ (c) Bokeh plot visualizing the cluster of documents in a 2D space (2D from the 9D topic simplex used by LDA's Dirichlet distribution) using the t-SNE (t-distributed stochastic neighbor embedding) algorithm. (d) WordCloud of top 50 most dominant words in most represented topics, Topic 4 and Topic 5 . . . . .	79
5.12	The variation of loss with per epoch for tasks, (a) Classification (b) Name entity recognition, while transfer learning the SciBERT model . . . . .	80
5.13	This Schematic represents the working our Ex-SciBERT models (1) Classification (2) Named entity recognition . . . . .	82
5.14	The performance evaluation of the classification task by the $H_2 - BERT$ model (a) The variation of loss per epoch (b) Confusion matrix for the test dataset . . . . .	85
6.1	A schematic diagram of the entire pipeline of this chapter's work is as follows: (a) Corpus extraction from Elsevier database, (b) Corpus annotation with the help of custom Python scripts (1) and (2) for Classification and Q&A, respectively, and (c) Transfer learning adaptation for fine-tuning BERT model. . . . .	88
6.2	(a) Text extraction from Elsevier database and (b) annotation steps for classification and Question and Answer tasks. . . . .	90
6.3	Top 30 Journals with increasing frequency of relevant abstracts on plastic recycling obtained from Elsevier database in last 22 years only . . . . .	91
6.4	An increasing trend in the number of relevant articles published in the last 22 years. . . . .	92
6.5	Extension of BERT to Recycle-BERT by deploying transfer learning . . . . .	94
6.6	Working of Recycle-BERT for a specific catalyst-based question . . . . .	94
6.7	The working of transfer learning and Recycle-BERT model . . . . .	94
6.8	A complete functioning of Q&A BERT model architecture to provide (a) starting and (b) ending alphabets for prediction of answer. . . . .	96
6.9	The evaluation of classification task, (a) variation of loss per epoch and (b) test dataset confusion matrix and similarly for Q&A models, have learning curves for (c) Catalyst-BERT, (d) Method-BERT, (e) Reactant-BERT and (f) Product-BERT for respective Q & A task. . . . .	97
6.10	Analysis of prediction of two random abstracts by Class-BERT to (a) non-relevant and (b) relevant class based on the contextual meaning of words present. . . . .	100

6.11	Analysis of two random relevant abstracts after passing through Class-BERT by individual Q&A models ((1) Catalyst-BERT, (2) Method-BERT, (3) Reactant-BERT, and (4) Product-BERT) one after another for predicting the catalysts used, reactants used, products formed, and methods used in the given abstracts, shown in (a), and (b), respectively. . . . .	101
6.12	Zipf plots representing the list of most useful catalysts (a, b), and methods (c, d) obtained using Catalyst-BERT, and Method-BERT, respectively for literature published in 2020 and 2022. . . . .	103
6.13	Zipf plots representing the list of most useful reactants (a, b), and products (c, d) obtained using Reactant-BERT, and Product-BERT, respectively for literature published in 2020 and 2022. . . . .	104
6.14	Each combined plot of word cloud and bar plot for literature published in (a) 2019, (b) 2020, (c) 2021, and (d) 2022. Each word cloud shows all the most probable catalyst-related entities and the bar plot represents the top 20 catalyst entities with their frequency of appearance. . . . .	105
6.15	Each combined plot of word cloud and bar plot for literature published in (a) 2019, (b) 2020, (c) 2021, and (d) 2022. Each word cloud shows all the most probable methodology-related entities, and the bar plot represents the top 20 methodology entities with their frequency of appearance. . . . .	106
6.16	Each combined plot of word cloud and bar plot for literature published in (a) 2019, (b) 2020, (c) 2021, and (d) 2022. Each word cloud shows all the most probable reactant-related entities, and the bar plot represents the top 20 reactant entities with their frequency of appearance. . . . .	107
6.17	Each combined plot of word cloud and bar plot for literature published in (a) 2019, (b) 2020, (c) 2021, and (d) 2022. Each word cloud shows all the most probable product-related entities, and the bar plot represents the top 20 product entities with their frequency of appearance. . . . .	108
7.1	A schematic diagram for each step of this proposed framework is as follows: Firstly, corpus extraction from the Elsevier database is shown. Second, three approaches are shown for extracting information related to technologies involved in PET waste recycling: (1) proposing time series Knowledge graphs (TSKGs), (2) the utilization of a pre-trained transformer-based model for forming similar document clusters, and (3) estimating the popularity index values for each cluster. . . . .	112
7.2	Top 30 Journals with increasing frequency of relevant abstracts on plastic recycling obtained from Elsevier database in last 22 years only . . . . .	114
7.3	An increasing trend in the number of relevant articles published in the last 22 years. . . . .	115
7.4	A working summary schematic flow diagram of the proposed framework in (a) time series knowledge graphs (TSKGs), (b) five steps in a dynamic transformer-based topic modeling (DTTM), and followed by calculation of popularity index for all extracted knowledge about PET plastic recycling. . . . .	116

7.5	A schematic diagram to showcase all necessary steps taken while making the time series knowledge graphs (TSKGs). . . . .	117
7.6	A diagram depicting every stage required in the Dynamic Transformer-based Topic Model (DTTM). . . . .	121
7.7	The output of TSKGs is shown in the form of a KG for publishing years 2000 to 2010. . . . .	123
7.8	The output of TSKGs is shown in the form of a KG for publishing years of 2022. . . . .	125
7.9	A multiple bar plot to present accumulative percentage change over the years for each recycling class . . . . .	127
7.10	(a) The number of clusters is decided using the Elbow method, where loss vs. number of clusters is shown and (b) all five clusters are shown with five colors. . . . .	131
7.11	The performance of DTTM is represented as follows: (a), (b), (c), (d), and (e) are five-word clouds with top hundred words based on importance metrics, and (f) shows a statistical analysis for all five clusters. . . . .	132
7.12	The quantitative study for all DTTM's extracted recycling classes (a) the Popularity Index (PI) and (b) accumulative percentage change for each class year-wise. . . . .	134
B.1	The word cloud plots for each question's answer extracted entities are shown as (a) Catalyst-based answer entities, (b) Reactant-based answer entities, (c) Product-based answer entities, and (d) Method-based answer entities. . . .	174
B.2	Zipf plots representing the list of most useful catalysts (a, b) and reactants (c, d) obtained using Catalyst-BERT, Reactant-BERT, Product-BERT, and Method-BERT, respectively for literature published in 2019 and 2021. . . .	175
B.3	Zipf plots representing the list of most useful products (a, b) and methods (c, d) obtained using Catalyst-BERT, Reactant-BERT, Product-BERT, and Method-BERT, respectively for literature published in 2019 and 2021. . . .	176
B.4	Results of two random relevant abstracts after passing through Class-BERT by individual Q&A models ((1) Catalyst-BERT, (2) Method-BERT, (3) Reactant-BERT, and (4) Product-BERT) one after another for predicting the catalysts used, reactants used, products formed, and methods used in the given abstracts, shown in (a) and (b), respectively. . . . .	177
B.5	The performance of the Class-BERT task of the Recycle-BERT on random mixture of relevant and non-relevant abstracts. . . . .	178
B.6	Top 30 journals with increasing frequency of relevant abstracts on plastic recycling obtained by using abstracts extracted from Elsevier database . .	201
B.7	A growth of published articles in the period of 2000-2022. . . . .	202

B.8	The word cloud and Zipf plot pair for catalyst-based answer entities for each year. The top pair (a,b) stands for the year 2019, and (c,d) for 2020 . . . . .	203
B.9	The word cloud and Zipf plot pair for catalyst-based answer entities for each year. The top pair (a,b) for 2021, and (c,d) last pair for 2022. . . . .	204
B.10	The word cloud and Zipf plot pair for reactant-based answer entities for each year. The top pair (a,b) stands for the year 2019, and (c,d) for 2020 . . . . .	205
B.11	The word cloud and Zipf plot pair for reactant-based answer entities for each year. The top pair stands for (a,b) for 2021, and (c,d) the last pair for 2022.	206
B.12	The word cloud and Zipf plot pair for method-based answer entities for each year. The top pair (a,b) stands for the year 2019, (c,d) for 2020. . . . .	207
B.13	The word cloud and Zipf plot pair for method-based answer entities for each year. The top pair (a,d) for 2021, and (c,d) last pair for 2022. . . . .	208
B.14	The word cloud and Zipf plot pair for product-based answer entities for each year. The top pair (a,b) stands for the year 2019, (c,d) for 2020. . . . .	209
B.15	The word cloud and Zipf plot pair for product-based answer entities for each year. The top pair (a,b) for 2021, and (c,d) last pair for 2022. . . . .	210
C.1	The output of TSKGs is shown in the form of a KG for publishing years 2011 to 2015 . . . . .	212
C.2	The output of TSKGs is shown in the form of a KG for publishing years 2016 to 2018 . . . . .	213
C.3	The output of TSKGs is shown in the form of a KG for publishing years 2019 to 2020 . . . . .	214
C.4	The output of TSKGs is shown in the form of a KG for publishing year 2021	215
C.5	The word clouds for cluster 0 by using the DTTM approach: for the period of (a) 2000–2004, (b) 2005–2008, (c) 2009–2012, and (d) 2013–2014. . . . .	224
C.6	The word clouds for cluster 0 by using the DTTM approach: for the period of (a) 2015–2016, (b) 2017–2018, (c) 2019–2020, and (d) 2021–2022. . . . .	225
C.7	The word clouds for cluster 1 by using the DTTM approach: for the period of (a) 2000–2004, (b) 2005–2008, (c) 2009–2012, and (d) 2013–2014. . . . .	226
C.8	The word clouds for cluster 1 by using the DTTM approach: for the period of (a) 2015–2016, (b) 2017–2018, (c) 2019–2020, and (d) 2021–2022. . . . .	227
C.9	The word clouds for cluster 2 by using the DTTM approach: for the period of (a) 2000–2004, (b) 2005–2008, (c) 2009–2012, and (d) 2013–2014. . . . .	228
C.10	The word clouds for cluster 2 by using the DTTM approach: for the period of (a) 2015–2016, (b) 2017–2018, (c) 2019–2020, and (d) 2021–2022. . . . .	229
C.11	The word clouds for cluster 2 by using the DTTM approach: for the period of (a) 2000–2004, (b) 2005–2008, (c) 2009–2012, and (d) 2013–2014. . . . .	230

---

C.12	The word clouds for cluster 2 by using the DTTM approach: for the period of (a) 2015–2016, (b) 2017–2018, (c) 2019–2020, and (d) 2021–2022. . . . .	231
C.13	The word clouds for cluster 2 by using the DTTM approach: for the period of (a) 2000–2004, (b) 2005–2008, (c) 2009–2012, and (d) 2013–2014. . . . .	232
C.14	The word clouds for cluster 2 by using the DTTM approach: for the period of (a) 2015–2016, (b) 2017–2018, (c) 2019–2020, and (d) 2021–2022. . . . .	233

## List of Tables

3.1	The hyperparameters values utilized to train the 2D-CNN models' weights for a binary classification task. . . . .	37
3.2	The hyperparameters values utilized to train the proposed model's weights for predicting the band gap of photo-active catalysts. . . . .	39
3.3	The evaluation of ML/DL algorithms with metrics precision, recall, F1-score, and accuracy for the test data set for the classification task. . . . .	40
3.4	The evaluation of ML/DL algorithms with metrics $R^2$ , MSE, RMSE, and MAE for the test data set, for the prediction task. . . . .	42
4.1	Description of different kernels used in the GPR modeling . . . . .	50
5.1	This table collated all necessary information of all nine topics extracted by LDA . . . . .	78
5.2	The summary of annotated data points utilization in train and test datasets for Ex-SciBERT transfer learning . . . . .	81
5.3	Performance of the Ex-SciBERT . . . . .	83
5.4	The classification and NER prediction results of Ex-SciBERT. The model first classifies the sentence and then highlights the sentence's catalyst and process parameters. The green color stands for catalyst material or sentence, and the blue color stands for process parameter or sentence. The other two colors, pink and yellow, are for neutral sentences and both catalyst and process content sentences, respectively . . . . .	84
5.5	The performance metric values for the task of Classification of the $H_2$ -BERT . . . . .	84
5.6	Performance of the $H_2$ -BERT for the task of Q&A. . . . .	85
5.7	The Q&A prediction on some random abstracts, Green, light blue, and red colors used for highlighting catalyst entities are present in the abstract, which is predicted correctly, similarly, and incorrectly, respectively. . . . .	86
6.1	The optimized hyperparameter values for Classification and Question and Answering tasks . . . . .	98
6.2	Summary of performance metrics obtained and the number of annotated data points used in training and testing of Class-BERT for the classification task. . . . .	98

6.3	F1-score and the number of data points utilized in training and testing datasets for four developed Q&A models of Recycle-BERT and their respective base models (without fine-tuning). . . . .	99
7.1	The symbolic representation for reactant form of PET waste plastic utilized in years 2000–2010 . . . . .	122
7.2	The symbolic representation of product form names after recycling the PET waste plastic utilized in years 2000–2010 . . . . .	124
7.3	The symbolic representation for reactant form of PET waste plastic utilized in 2022 . . . . .	126
7.4	The symbolic representation of product form names after recycling the PET waste plastic utilized in 2022 . . . . .	127
7.5	The description of formed clusters after implementing the DTTM approach.	130
A.1	The $R^2$ values train and test for the recycle reactor for all trained ML models by using conversion as a target variable . . . . .	166
A.2	The $R^2$ values of train and test for the recycle reactor for all trained ML models by using selectivity as a target variable . . . . .	167
A.3	The $R^2$ values of train and test for the once-through reactor for all trained ML models by using conversion as a target variable . . . . .	167
A.4	The $R^2$ values of train and test for the once-through reactor for all trained ML models by using selectivity as a target variable . . . . .	167
A.5	The $R^2$ values of train and test for GPR model with different kernels and the conversion as target variable using the recycle reactor data set . . . . .	168
A.6	The $R^2$ values of train and test for GPR model with different kernels and the selectivity as target variable using the recycle reactor data set . . . . .	169
A.7	The $R^2$ values of train and test for GPR model with different kernels and the conversion as target variable using the once-through reactor data set . . . . .	170
A.8	The $R^2$ values of train and test for GPR model with different kernels and the selectivity as target variable using the once-through reactor data set . . . . .	171
B.1	The some relevant abstracts are tested by Recycle-BERT module, all four question’s answered . . . . .	180
C.1	The symbolic representation for reactant form of PET waste plastic utilized in years 2011–2015 . . . . .	216
C.2	The symbolic representation of product form names after recycling the PET waste plastic utilized in years 2011–2015 . . . . .	217
C.3	The symbolic representation for reactant form of PET waste plastic utilized in years 2016–2018 . . . . .	218

---

C.4	The symbolic representation of product form names after recycling the PET waste plastic utilized in years 2016–2018 . . . . .	219
C.5	The symbolic representation for reactant form of PET waste plastic utilized in 2019-2020 . . . . .	220
C.6	The symbolic representation for reactant form of PET waste plastic utilized in years 2019 to 2020 . . . . .	221
C.7	The symbolic representation for reactant form of PET waste plastic utilized in 2021 . . . . .	222
C.8	The symbolic representation of product form names after recycling the PET waste plastic utilized in 2021 . . . . .	223

## ABBREVIATIONS

<b>CNN</b>	Convolutional Neural Network
<b>XGBoost</b>	Extreme Gradient Boosting
<b>VGG</b>	Visual Geometry Group
<b>ResNet</b>	Residual Neural Network
<b>MSE</b>	Mean Square Error
<b>RMSE</b>	Root Mean Square Error
<b>MAE</b>	Mean Absolute Error
$R^2$	R Square
<b>GPR</b>	Gaussian Process Regression
<b>BO</b>	Bayesian Optimization
<b>MOBO</b>	Multi-Objective Bayesian Optimization
<b>SHAP</b>	SHapley Additive exPlanations
<b>LR</b>	Linear Regression
<b>LOR</b>	LassO Regression
<b>RR</b>	Ridge Regression
<b>ELR</b>	ELastic net Regression
<b>PCR</b>	Principal Component Regression
<b>PLSR</b>	Partial least Square Regression
<b>SVR</b>	Support Vector Regression
<b>MLP</b>	Multi-Layer Perceptron
<b>NLP</b>	Natural Language Processing
<b>BERT</b>	Bidirectional Encoder Representations from Transformers
<b>LDA</b>	Latent Dirichlet Allocation
<b>NER</b>	Named Entity Recognition
<b>Q&amp;A</b>	Question and Answering
$H_2 - BERT$	Hydrogen-BERT
<b>Ex-SciBERT</b>	Extend-SciBERT

---

<b>TF-IDF</b>	Term Frequency-Inverse Document Frequency
<b>Class-BERT</b>	Classification-BERT
<b>LM</b>	Language Model
<b>PET</b>	Polyethylene teraphthalate
<b>KG</b>	Knowledge Graph
<b>TSKGs</b>	Time Series Knowledge Graphs
<b>DTTM</b>	Dynamic Transformer-based Topic Modeling
<b>PI</b>	Popularity Index

## NOTATION

$GP$	Gaussian Process
$\mu(x)$	Mean
$\eta$	Error
$k(x, x')$	Covariance matrix
$N(0, \sigma^2)$	Normal distribution
$l$	length scale
$d(\cdot, \cdot)$	Euclidean distance
$\alpha$	scale mixture
$\Gamma(\cdot)$	Bessel function
$\phi_j(f_x)$	Estimated shapley value
$\psi(x)$	Gaussian distribution's cumulative
$\phi(z)$	density function
$obj_{cov}$	Objective function for conversion
$obj_{sel}$	Objective function for selectivity
$obj_{weight}$	weighted objective
$w$	weight
$GP_1(x)$	GPR model for conversion
$GP_2(x)$	GPR model for selectivity
$K$	Number of topics
$D$	Corpus
$M$	Number of documents
$N_d$	number of words in a document $d$
$\phi$	multinomial distributions
$\theta$	probability of topic $k$ occurring in document $d$
$\beta$	Dirichlet Distributions for topic-words
$PI(\text{time}, w, \text{citation})$	Popularity index function
$f(w_i)$	Frequency of $i^{th}$ word
$imp(w_i)$	Importance of $i^{th}$ word
$Citation_j$	Citation number of $j^{th}$ article