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

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 Intelligence, 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-कैटेलिटिक 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 रिग्रेशन (जीपीआर) रिग्रेसर और मल्टी-ऑब्जेक्टिव ऑब्जेक्टिव बायेसियन ऑप्टिमाइज़ेशन (MOBO), क्रमशः। हमारे प्रशिक्षित जीपीआर मॉडल में आर2 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, क्रमश। 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 पॉलीथीन टेरेफ्थेलेट (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 ट्रेंडिंग 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

List of Figures

List of Tables

ABBREVIATIONS

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

TF-IDF	Term Frequency-Inverse Document Frequency
Class-BERT	Classification-BERT
LM	Language Model
PET	Polyethylene teraphthalate
KG	Knowledge Graph
TSKGs	Time Series Knowledge Graphs
DTTM	Dynamic Transformer-based Topic Modeling
PI	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
$d(.,.)$	Euclidean distance
$alpha$	scale mixture
$\Gamma(.)$	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
ϕ	multinomial distributions
$theta$	probability of topic k occurring in document d
β	Dirichlet Distributions for topic-words
$PI(time, w, citation)$	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