MACHINE LEARNING AND NATURAL LANGUAGE PROCESSING FOR CONSTRUCTION MATERIALS

MOHD ZAKI

DEPARTMENT OF CIVIL ENGINEERING

INDIAN INSTITUTE OF TECHNOLGY DELHI

INDIA

JANUARY 2025

MACHINE LEARNING AND NATURAL LANGUAGE PROCESSING FOR CONSTRUCTION MATERIALS

by

MOHD ZAKI
Department of Civil Engineering

Submitted

in fulfillment of the requirements of the degree

DOCTOR OF PHILOSOPHY

to the

Indian Institute of Technology Delhi

January 2025

Dedicated to my family!

CERTIFICATE

This is to certify that the thesis titled

Machine Learning and Natural Language Processing for Construction Materials

is being submitted by Mr. Mohd Zaki (2019CEZ8233), to the Indian Institute of Technology Delhi, for the award of the degree of Doctor of Philosophy. This is a bonafide record of research work and is entirely carried out by him under my supervision and guidance. 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.

Place: New Delhi
Date:23rd January 2025

(Dr. N. M. Anoop Krishnan)
Associate Professor
Department of Civil Engineering
IIT Delhi, New Delhi,
Delhi, 110016

(Dr. Jayadeva)
Professor
Department of Electrical Engineering
IIT Delhi, New Delhi,
Delhi, 110016

DECLARATION

I hereby declare that the thesis contains the work carried out in fulfilment of the requirements for the award of the degree of Doctor of Philosophy under the guidance and supervision of Dr. N. M. Anoop Krishnan, Associate Professor, Department of Civil Engineering, IIT Delhi and Dr. Jayadeva, Professor, Department of Electrical Engineering, IIT Delhi. The contents of the dissertation, including text, tables, figures, etc., have not been reproduced from other reports, theses, etc. Wherever reproduction has been made from books, journals, reports, manuals, websites, etc., the sources have been duly acknowledged and referenced appropriately.

Mohd Zaki

Date: 23/January/2025

Mohd Zaki
2019CEZ8233

ACKNOWLEDGMENT

I would like to express my deepest gratitude to my esteemed supervisors, Prof. N. M. Anoop Krishan and Prof. Jayadeva, for their invaluable expertise, profound knowledge, and dedication to academic excellence, which provided constant motivation throughout my research journey. Their insightful feedback, constructive criticism, and unwavering support enabled me to push beyond my limitations and explore new avenues of research. With their expert guidance, I performed interdisciplinary research at the intersection of civil engineering, materials science, and machine learning, honing skills that will undoubtedly serve me well in my future endeavors.

I would like to extend my sincere thanks to my research committee members: Prof. Ramanathan Ayothiraman (Chairperson), Prof. Shashank Bishnoi (Departmental expert), and Prof. Prathosh A. P. (External expert, IISc Bengaluru), for their invaluable encouragement and insightful suggestions that enhanced the quality of my research from multiple perspectives. I am also thankful to the Head of the Department and the esteemed faculty members of the Department of Civil Engineering for their continued support, along with the department's staff, whose assistance was greatly appreciated. I also wish to express my heartfelt gratitude to Prof. Mausam for his insights into the natural language processing aspect of my work, and to his students—Tanishq Gupta, Devanshi Khatsuriya, Kausik Hira, Vaibhav Mishra, and Somaditya Singh—for their collaboration in developing exceptional tools like MatSciBERT, DiSCoMaT, and LLaMat, and facilitating materials discovery. I am also thankful to Prof. Keshav Ravi Bharadwaj (IISc Bengaluru), for his guidance in creating the largest dataset of cementitious materials, a significant contribution to my research.

I am also grateful to the almighty and my family for their unwavering support, understanding, and love, without which this thesis would not have been possible.

ABSTRACT

Materials discovery for targeted applications has been the keystone for the progress of human civilizations. In civil engineering, key materials of interest include cement, concrete, and glass. Concrete, the most widely used man-made material, plays a pivotal role in the construction industry. Glasses, an infinitely recyclable class of material, has become an important material for constructing energy efficient buildings. Since glass can be manufactured by utilizing a wide range of periodic table elements and their compounds, it has applications across several fields like architecture, energy, environment, transportation, information technology, and healthcare. Cement, used in concrete can also be manufactured by using alternate materials and hence reducing its carbon footprint. However, developing improved materials for growing human needs faces two major challenges (a) Search amongst existing materials and resources, and (b) obtain properties of materials without experimentation.

In this thesis, we tackle the first challenge of information extraction for cementitious materials by developing a language model that can understand the text in the domain of cement and concrete research. This language model is than used to train specialized model for extracting entities like materials, synthesis methods, characterization methods, applications, material descriptors, properties, and phase labels for cementitious materials from text of related research papers. Further, complete details of compositions of cementitious materials are found in the table of research papers. Therefore, we used a graph neural networks-based pipeline combined with handcrafted rules to extract the oxide compositions from tables. In the case of glasses, we used rule-based systems, un-supervised machine learning, and manual extraction to create database of structure factor of glasses.

The second challenge, i.e, understanding the effect of composition, processing, and testing conditions on properties of glasses, is addressed through developing machine learning models by

considering wide range of input parameters including compositions, and additional parameters like wavelength, temperature, testing loads, frequencies, depending upon property of interest.

Finally, considering the recent rise in the application of large language models (LLMs) in materials discovery related tasks, we created the first benchmarking dataset to evaluate the performance of LLMs on different aspects of materials science and compare the performance of 3 different LLMs on questions from different subdomains of materials science. Further, we also evaluate the performance of GPT-4 on extraction of material composition from tables and writing python codes related to materials discovery.

Keywords: construction materials, oxide glass, cement, concrete, materials databases, materials discovery, composition – property relationships, materials tetrahedron, explainable machine learning, SHAP analysis, language modelling, natural language processing, large language models, information extraction, computer vision, optical microscopy

सार

मानव सभ्यताओं की प्रगति के लिए लक्षित अनुप्रयोगों के लिए सामग्री की खोज एक आधारशिला रही है। सिविल इंजीनियरिंग अनुप्रयोगों के लिए कुछ रुचिकर सामग्री सीमेंट, कंक्रीट और कांच हैं। कंक्रीट सबसे अधिक उपयोग किया जाने वाला मानव निर्मित पदार्थ है, जिसका उपयोग निर्माण क्षेत्र में होता है। कांच, जो अनंत बार पुनर्नवीनीकरण किया जा सकता है, ऊर्जा कुशल इमारतों के निर्माण के लिए एक महत्वपूर्ण सामग्री बन गया है। चूंकि कांच को आवर्त सारणी के विभिन्न तत्वों और उनके यौगिकों का उपयोग करके बनाया जा सकता है, इसका उपयोग वास्तुकला, ऊर्जा, पर्यावरण, परिवहन, सूचना प्रौद्योगिकी और स्वास्थ्य सेवा जैसे कई क्षेत्रों में होता है। बढ़ती मानव आवश्यकताओं के लिए बेहतर सामग्री विकसित करना दो मुख्य चुनौतियों का सामना करता है: (a) मौजूदा सामग्री और संसाधनों के बीच खोज, और (b) प्रयोग किए बिना सामग्री के गुण प्राप्त करना।

इस शोध प्रबंध में, हम सीमेंट और कंक्रीट अनुसंधान के क्षेत्र में पाठ को समझने में सक्षम भाषा मॉडल विकसित करके सूचना निष्कर्षण की पहली चुनौती का समाधान करते हैं। इसके बाद, हमने विशेष मॉडल का उपयोग करके शोध पत्रों के पाठ से सीमेंटस सामग्री के लिए सामग्री, संश्लेषण विधियाँ, गुणों के चर, अनुप्रयोग, गुण, और चरण लेबल जैसी संस्थाओं को निकाला। इसके अलावा, सीमेंटस सामग्री की संरचना का पूरा विवरण शोध पत्रों की तालिकाओं में मिलता है। इसलिए, हमने ऑक्साइड संरचना को तालिकाओं से निकालने के लिए ग्राफ़ न्यूरल नेटवर्क-आधारित पाइपलाइन का उपयोग किया, जिसमें हाथ से बनाई गई नियम भी सम्मिलित किए गए। कांच के मामले में, हमने संरचना कारकों का डेटाबेस बनाने के लिए नियम-आधारित प्रणालियों, बिना निगरानी के मशीन लर्निंग और मैन्युअल निष्कर्षण का उपयोग किया।

दूसरी चुनौती, यानी संरचना, प्रसंस्करण और परीक्षण स्थितियों का गुणों पर प्रभाव समझने के लिए, हमने व्यापक इनपुट पैरामीटरों को ध्यान में रखते हुए मशीन लर्निंग मॉडल विकसित किए। इनमें संरचना के साथ-

साथ वेवलेंथ, तापमान, परीक्षण भार, आवृत्तियाँ जैसे अतिरिक्त पैरामीटर शामिल हैं, जो गुणों पर निर्भर करते हैं।

अंत में, सामग्री की खोज से संबंधित कार्यों में बड़े भाषा मॉडल (एलएलएम) के हालिया उपयोग को ध्यान में रखते हुए, हमने एलएलएम के प्रदर्शन का मूल्यांकन करने के लिए सामग्री विज्ञान के विभिन्न पहलुओं पर पहला बेंचमार्किंग डेटाबेस बनाया और सामग्री की खोज के विभिन्न उप-क्षेत्रों से प्रश्नों पर तीन अलग-अलग एलएलएम के प्रदर्शन की तुलना की। इसके अलावा, हमने तालिकाओं से सामग्री संरचना निकालने और सामग्री की खोज से संबंधित पायथन कोड लिखने में GPT-4 के प्रदर्शन का भी मूल्यांकन किया।

Table of Contents

List of Figures

List of Tables