

MACHINE LEARNING AND NATURAL LANGUAGE PROCESSING FOR CONSTRUCTION MATERIALS

MOHD ZAKI



**DEPARTMENT OF CIVIL ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY DELHI
INDIA**

JANUARY 2025

© Indian Institute of Technology Delhi, 2025
All rights reserved

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.

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 then 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, unsupervised 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 मैनुअल 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

DEDICATION	
CERTIFICATE	i
DECLARATION	iii
ACKNOWLEDGEMENTS	v
ABSTRACT	vii
ABSTRACT (Hindi)	ix
TABLE OF CONTENTS	xi
LIST OF FIGURES	xv
LIST OF TABLES	xix
1 Introduction	1-6
1.1 Background and motivation	1
1.2 Challenges in research of cementitious materials	2
1.3 Challenges to develop glasses for real-world applications	3
1.4 Organization of thesis.....	4
2 Literature Review.....	7-20
2.1 Introduction	7
2.2 Information extraction for cementitious materials.....	7
2.3 Information extraction for glasses.....	10
2.4 Machine learning composition-property relationship of glasses	13
2.5 Large language models for materials science	16
2.6 Research gaps	18
2.7 Research Objectives	19
3 Methodology	21-30
3.1 Datasets preparation for glasses	21

3.2	Regression machine learning model development for glasses	22
3.2.1	Neural networks	22
3.2.2	XGBoost.....	23
3.3	Explaining predictions of ML models through SHAP.....	24
3.4	Natural language processing	26
3.4.1	Word vectors	26
3.4.2	Named entity recognition.....	27
3.4.3	Language models.....	28
3.4.4	Information extraction from tables.....	29
4	Information extraction from construction materials research literature.....	31-68
4.1	Introduction	31
4.2	Research contributions	32
4.3	Composition extraction of cementitious materials	33
4.3.1	Pretraining dataset details.....	34
4.3.2	Cementitious materials NER dataset preparation	35
4.3.3	Mineralogical tables dataset preparation.....	36
4.3.4	Language model pretraining.....	37
4.3.5	Finetuning models for downstream tasks.....	38
4.3.6	Language model pretraining and evaluation	39
4.4	Extracting cementitious materials compositions from tables	41
4.4.1	Mineralogical composition extraction.....	41
4.4.2	Extracting oxide composition from tables	43
4.4.3	Applications of extracted compositions	44
4.4.4	Conclusion.....	48
4.5	Information extraction from clinker optical micrographs.....	49
4.5.1	Introduction	49
4.5.2	Methodology	51
4.5.3	Model training	52
4.5.4	Experimental details to obtain clinker images	56
4.5.5	Results	56
4.5.6	Application of Cementron model.....	66

4.5.7 Conclusion.....	67
5 Information extraction from glass literature	69-84
5.1 Introduction	69
5.2 Methodology	70
5.2.1 Data collection.....	70
5.2.2 Metadata analysis	71
5.2.3 Structure factor data extraction	72
5.3 Results and discussion.....	73
5.3.1 Usage of the corpus	73
5.3.2 Metadata analysis of the corpus	74
5.3.3 Structure factor repository.....	81
5.4 Conclusion.....	84
6 Explainable machine learning for understanding the effect of composition, processing, and testing conditions on properties of glasses	85-140
6.1 Introduction	85
6.2 Research contributions	88
6.3 Explaining the effect of oxide compositions on optical properties	88
6.3.1 Dataset preparation.....	88
6.3.2 Model training	89
6.3.3 Results	91
6.3.4 Discussion	100
6.3.5 Conclusion.....	105
6.4 Understanding compositional and testing condition effects on refractive index, density, dielectric constant, and loss tangent inorganic melts and glasses.....	106
6.4.1 Dataset preparation.....	106
6.4.2 Machine learning, hyperparameter optimization, and explainability	107
6.4.3 Results and discussion.....	108
6.4.4 Composition–testing parameters–property relationship	115

6.4.5	Delineating combined effect of composition and testing parameters	118
6.4.6	Conclusion.....	122
6.5	Understanding compositional, processing and testing condition effects on hardness of oxide glasses	123
6.5.1	Introduction	123
6.5.2	Methodology	125
6.5.3	Results and discussion.....	130
6.5.4	Conclusion.....	138
7	Large language models for materials discovery.....	141-166
7.1	Introduction	141
7.2	Methodology	142
7.2.1	Dataset preparation.....	142
7.2.2	Answering questions using LLMs	144
7.3	Results	146
7.3.1	Dataset visualization	146
7.3.2	Performance evaluation.....	150
7.4	Discussion	155
7.4.1	Error analysis.....	155
7.4.2	Comparison with human performance	160
7.4.3	Compositions extraction from materials tables.....	161
7.4.4	Code writing	164
7.5	Conclusion.....	165
8	Conclusions and Recommendations for future work.....	167-172
8.1	Concluding Remarks	167
8.2	Recommendations for future work	170
	References	173-197

List of Figures

Figure 3.1	An example of how words are converted into vector representations.	27
Figure 3.2	Materials knowledge graph (a) Connecting two unconnected entities, (b) exploring entities related to characterization method “XRD”[131].	28
Figure 3.3	Demonstrating the importance of context for information extraction application.	29
Figure 3.4	The architecture of DiSCoMaT pipeline for IE from tables [36].	30
Figure 4.1	Information extraction methodology from different parts of a research paper.	33
Figure 4.2	Annotating named entities in doccano[191].	36
Figure 4.3	Visualizing (a) Loss curve, and (b) Perplexity of different checkpoints of CemSciBERT on train and validation (val) corpus.	39
Figure 4.4	Comparing the performance of NER models on CemNER dataset.	40
Figure 4.5	Example of table provided as input to GPT-4 for mineralogical composition extraction.	42
Figure 4.6	Frequency of the mineralogical phases extracted from tables.	43
Figure 4.7	Visualizing the frequency of different oxides found in the extracted oxide compositions.	44
Figure 4.8	Visualizing the frequency of different materials reported in research literature.	45
Figure 4.9	Visualizing the trend of different SCM compositions being reported in research papers since 1997.	46
Figure 4.10	Visualizing the CO ₂ emissions based on cement compositions reported in papers from 1997 to 2023.	48
Figure 4.11	The methodology adopted to develop segmentation models from optical micrographs of cement clinker. The major steps include data collection, annotations, model training, and evaluation.	52
Figure 4.12	Different images of clinker microstructure showing challenges in the segmentation. The particles with angular edges are alite and the rounded particles are belite. The alite phase is shown in (a, b, c) dark blue color with different reflectance, (d) brown, (e) blue with poor contrast, (f) blue. The belite particles are shown in (a,b)olive green, (c) brown, (d) blue, (e) mixed reflectance of blue and olive green, (f) gray.	58
Figure 4.13	Visualizing steps and output for MoW approach: (a) actual image, (b) annotated mask, (c) sampling windows on the actual image, (d) predicted pixel labels.	59

Figure 4.14	Visualizing predictions of Cementron trained on RGB images for identifying individual particles: (a) image having both alite and belite, (b) image having only alite.	62
Figure 4.15	Visualizing predictions of Cementron while identifying belite particles with finger-like structures.	63
Figure 4.16	Visualizing predictions and performance of Cementron model on microscopy image of clinker C-1 (a) actual image, (b) annotated mask for alite, (c) predicted mask for alite, (d) annotated mask for belite, (e) predicted mask for belite.	65
Figure 4.17	Clinker microstructure with identified particles (a) centers are marked with a “*” symbol, and the number alongside the symbol indicates the assigned particle's id, (b) particle size distribution curve obtained using particle size and area.	67
Figure 5.1	Extracting research papers reporting structure factors by using caption cluster plot and LDA plots.	74
Figure 5.3	Visualizing the number of research papers published in glass science from 1962 to 2019.	75
Figure 5.4	Most active countries in glass research.	76
Figure 5.5	Number of glass-related publications for the top 5 most prolific countries in 1956 – 2019.	77
Figure 5.6	Top twenty active institutions in the field of glass research.	78
Figure 5.7	Visualizing the most prominent keywords and properties. The bar graph shows the top 20 keywords, and the pie chart shows the number of papers investigating different properties of glasses.	79
Figure 5.8	Research journal and the respective number of published papers.	80
Figure 5.9	Comparing (a) existing and (b) extracted structure factor plots of pure V ₂ O ₅ glass obtained from neutron and XRD measurements [265].	83
Figure 5.10	Visualizing the compositions in the structure factor dataset. The bar chart shows the frequency of each chemical component. The pie chart in the inset shows the number of single and multi-component glasses.	84
Figure 6.1	Selected key applications with the corresponding properties of glasses.	86
Figure 6.2	Visualizing refractive index at 587.6 nm (nd) (a) number of oxides and their frequency, (b) frequency of glasses with the number of components, (c) distribution of property values.	92
Figure 6.3	Visualizing Abbe number (V_d) dataset (a) number of oxides and their frequency, (b) frequency of glasses with the number of components, (c) distribution of property values.	93

Figure 6.4	Predicted values of (a) Abbe number (V_d) and (b) refractive index at 587.6 nm (n_d) of oxide glasses using optimized ML models with respect to the experimental values for the test dataset. The heat map shows the number of data points per pixel.	95
Figure 6.5	Ternary plot for (a) Abbe number (V_d) and (b) refractive index (n_d) of [BaO, B ₂ O ₃ , SiO ₂] crown glass and (c) Abbe number (V_d) and (d) refractive index of (n_d)[PbO, K ₂ O, SiO ₂]flint glass. Square markers represent properties of experimental compositions.	96
Figure 6.6	Summary plot for (a) average impact on model output, (b) SHAP values for V_d . The color of points in (b) represents the magnitude of the weight % of oxide in the glasses.	99
Figure 6.7	Summary plot for (a) average impact on model output, (b) SHAP values for n_d . The color of points in (b) represents the magnitude of weight % of the oxide in the glasses.	100
Figure 6.8	Riverflow plots for (a) V_d and (b) n_d	102
Figure 6.9	Visualising n_F - n_c with respect to n_d and V_d for selected compositions given at this link	103
Figure 6.10	Abbe diagram generated using the ML models. Points represent the experimental data. Colored regions (created using ML models) indicate the range of properties achievable using respective compositions.	104
Figure 6.11	Visualizing data of refractive index (a) Chemical components and their frequency, (b) Number of multicomponent glasses, and (c) Histogram of property values.	109
Figure 6.12	Visualizing data of density (a) Chemical components and their frequency, (b) Number of multicomponent glasses, and (c) Histogram of property values.	110
Figure 6.13	Visualizing data of dielectric constant (a) Chemical components and their frequency, (b) Number of multicomponent glasses, and (c) Histogram of property values.	111
Figure 6.14	Visualizing data of loss tangent (a) Chemical components and their frequency, (b) Number of multicomponent glasses, and (c) Histogram of property values.	113
Figure 6.15	Visualizing predictions of trained ML models for (a)refractive index, (b) density, (c) dielectric constant, and (d) loss tangent.	114
Figure 6.16	SHAP riverflow and scatter plots for (a) refractive index, (b) density, (c) dielectric constant, and (d) loss tangent.	118
Figure 6.17	SHAP interaction plots for (a) refractive index, (b) density, (c) dielectric constant, and (d) loss tangent.	120

Figure 6.18	Pipeline to prepare dataset and predict hardness of oxide glasses using ML. The first step shows the retrieval of compositions for which hardness is available at multiple loads, the second step shows obtaining the full text of research papers where the data is reported. The information from text and figures are extracted using spacy and WebPlotDigitizer, respectively. The database is then used to train ML models.	129
Figure 6.19	Frequency of top fifteen words tagged as (a) nouns, (b) verbs, (c) adjectives, and (d) proper nouns.	132
Figure 6.20	Extracting annealing temperature from the dependency parse tree created after part of speech tagging.	134
Figure 6.21	Extracting load using named entity recognition.	134
Figure 6.22	Frequency of (a) the number of glass components, and (b) each of the components. (c) Histogram of hardness values.	134
Figure 6.23	Comparing the performance of neural networks for predicting hardness with (a) only composition, (b) composition and load, (c) composition and annealing temperature, and (d) composition, load and annealing temperature as inputs.	135
Figure 6.24	Visualizing model predictions using SHAP (a) bar plot (b) violin plot, and (c) river flow plot.	137
Figure 7.1	Sample questions from each category: (a) multiple choice question (MCQ), (b) matching type question (MATCH), (c) numerical question with multiple choices (MCQN), and (d) numerical question (NUM). The correct answers are bolded and underlined.	144
Figure 7.2	The number of questions in each materials science sub-domain. The bar chart shows the distribution of questions in different sub-domains. The pie chart shows the number of questions classified according to question structure.	147
Figure 7.3	Word-cloud for different topics in MaScQA.	149
Figure 7.4	Frequency of top – 10 words in each subdomain present in MaScQA.	150
Figure 7.5	Performance of GPT-4-CoT on questions classified from materials science domain perspective.	154
Figure 7.6	Types of the errors made by GPT-4-CoT on 100 questions classified based on the structure.	157
Figure 7.7	Types of the error made by GPT-4-CoT on 100 questions classified according to domain expertise required to solve them.	158
Figure 7.8	Visualizing some of the questions where GPT-4-CoT made conceptual errors in the solution. The correct answers to each question are marked in bold and underlined.	160
Figure 7.9	(a) system message (b) Table as prompt to extract materials compositions from tables using GPT-4 API.	162
Figure 7.10	(a) An example prompt provided to the GPT-4 model for generating the complete output (b) Response of GPT-4.	164

List of Tables

Table 2.1	List of question answering datasets in mathematics, basic sciences, and materials science domain.	17
Table 4.1	Performance of rule-based system and CemSciBERT table classification model.	41
Table 4.2	Clinker compounds and their respective chemical CO ₂ emissions.	47
Table 4.3	XGBoost model’s performance on classifying pixels as Alite and Belite for microstructure the shown in Figure 4.13.	59
Table 4.4	Performance of Cementron models in identifying alite and belite particles from clinker microstructure images from the test dataset.	61
Table 4.5	Mineralogical composition (in wt %) of clinkers experimentally used for testing the Cementron models.	64
Table 4.6	Cementron’s performance on classifying pixels as Alite and Belite for microstructure of clinker C-1 (shown in Figure 4.16).	65
Table 7.1	Performance (% accuracy) of different evaluation styles using LLaMA and GPT models on various question types. The number in parenthesis represents the total number of questions under respective categories.	152
Table 7.2	p-values obtained from statistical testing of LLMs performance using paired t-test.	153
Table 7.3	Comparing the performance of GPT-4-CoT with human performance.	161
Table 7.4	Comparing the performance of GPT-4 and DiSCoMaT on the composition extraction task.	164
Table 7.5	Comparing the performance of different LLMs on code generation task.	165