Proceedings of the 21st ICSMGE, Vienna, Austria, 14 – 19 June 2026. Pistrol, Adam & Schweiger (eds.)
Published by: ÖGG, Austrian Society for Geomechanics, Salzburg, Austria, ISBN 978-3-9503898-4-5
https://doi.org/10.53243/ICSMGE2026-515

Leveraging AI to Analyze Unstructured Site Investigation Reports in South Korea

Eomzi Yang, Jin-Tae Han, Byeong-Soo Yoo
Department of Geotechnical Engineering Research, Korea Institute of Civil Engineering and Building Technology, Gyeonggi, 10223, Korea, twothumbs@kict.re.kr

ABSTRACT: The importance of national-scale databases has emerged as an issue in AI-driven geotechnical engineering field because of the limited availability of investigation data. However, the geotechnical engineering field in South Korea faces challenges in standardizing geotechnical investigation reports, which vary in format. This lack of standardization necessitates manual data extraction and database registration, resulting in inefficiencies and increased resource consumption. To address these issues, this research focuses on automating geotechnical data extraction. Previous studies employing image-based AI for form classification and data extraction have limitations, such as reliance on predefined datasets and inability to process textual data or adapt to new formats. In contrast, this study leverages natural language processing (NLP) models to enable content-based data extraction. The first phase integrated PDF reading libraries with image processing to handle documents with unstructured data, ensuring seamless extraction from both Korean and English texts in complex table structures. The second phase employed a large language model (LLM)-based geotechnical term translator to standardize technical terminology across languages. This minimizes repetitive tasks while improving consistency in data interpretation. In the third phase, word embedding models trained on large-scale geotechnical corpora, including scientific articles and glossaries, were utilized to transform textual data into numerical vectors for enhanced contextual analysis. Finally, ML-driven text classification algorithms, incorporating convolutional neural networks and machine learning, categorized the text into specific labels and their associated data. This process was followed by organizing the extracted information into structured formats, such as borehole-specific depth measurements and general details. This integrated approach, combining LLM-based translation, word embedding in NLP, and ML classification, significantly enhanced the flexibility and accuracy of geotechnical data processing. The automation framework reduced manual effort, ensured consistent database development, meaningfully supporting other countries with the growing geotechnical engineering industries to initiate construct the geotechnical database.

KEYWORDS: Borehole log report, Text classification, GloVe, Table layout

1 INTRODUCTION

Text classification has emerged as a vital technique across a broad spectrum of industries, particularly those handling large volumes of unstructured textual content. For example, in content generation and distribution platforms, social media, and e-commerce systems, text classification has been extensively used to perform sentiment analysis. This allows platforms to tailor recommendations based on user-generated content such as comments, feedback, or reviews, thereby improving personalization and user engagement. In engineering management, text classification analysis was employed to automatically identify the risk in contract documents, and predict the possible responses for the potential conflicts. In the realm of engineering and construction management, similar techniques have been adopted to automatically identify latent risks embedded in complex contract documents and to anticipate appropriate responses in scenarios involving potential disputes. Furthermore, within the residential construction sector, automated classification of complaints lodged by tenants has helped categorize various types of structural or functional defects in buildings, thereby streamlining maintenance workflows and improving service response times (Pham and Han, 2023; Jeon et al., 2024; Dikmen et al., 2025).

These applications share a common goal: reducing the extensive manual effort traditionally required to review large-scale document collections. Text classification not only accelerates the identification of relevant information from vast textual datasets but also ensures timely and accurate decision-making support in professional environments.

At its core, text classification refers to the computational process of assigning predefined labels or categories to textual inputs. The general workflow of a text classification framework can be broken down into three major components: (1) preparation of labeled or unlabeled text data, (2) conversion of text into a machine-readable format through feature engineering or vectorization, and (3) application of classification algorithms to label the transformed data. The transformation step, known as text representation, plays a crucial role in enabling machines to interpret the semantic and syntactic relationships present in human language. It facilitates downstream tasks such as similarity matching between user queries and documents, thereby enhancing the performance of information retrieval systems.

Text representation methods have evolved over time. Early approaches relied on sparse encodings such as one-hot vectors, but modern techniques favor dense representations, often referred to as distributed representations. One of the pioneering methods in this domain is latent semantic indexing (LSI), which aimed to uncover the latent semantic structure of textual content by analyzing co-occurrence statistics across documents. LSI constructs a term-document matrix and then applies singular value decomposition (SVD) to reduce dimensionality, thus mitigating common issues in keyword-based retrieval such as synonymy and polysemy (Deerwester et al., 1990).

While LSI provided foundational insights into semantic representation, it did not account for the sequential order of words—a key component in natural language understanding. To address this limitation, neural network-based language models were introduced. These models, such as the neural probabilistic language model (NPLM), are designed to predict the next word in a sentence given the previous n−1 words. NPLM jointly learns both a fixed-dimensional embedding for each word and a probability distribution for word prediction, thereby capturing local contextual dependencies in the text (Bengio et al., 2003).

Skip-gram method in Word2Vec was built on neural network-based optimization of word vectors, and it was introduced as a computationally efficient model by replacing the nonlinear hidden layers with a simple projection layer (Mikolov et al., 2013). The proposed skip-gram method yielded reasonable results in predicting a group of semantically relevant words

from a given word within local context window. The term ‘word embedding’ gained prominence around this time, and its definition became more concretely established as the process of mapping a word into a continuous numeric vector based on contextual similarity. Global vectors for word representation (GloVe) utilized the statistical information from a large corpus to extract the embedding vectors by matrix factorization similar to LSI, and enhanced the semantic analogy with local context by weighting the word co-occurrence based on context window (Pennington et al., 2014).

Due to the availability of large-scale textual resources, the geoscience and geological engineering communities have actively employed natural language processing (NLP) and text mining techniques to conduct structured analyses of their data. For example, text classification has been successfully utilized in constructing three-dimensional lithological maps by analyzing borehole descriptions extracted from structured databases, such as the Australian Groundwater Explorer (Fuentes et al., 2020; Lawley et al., 2023; Chu et al., 2025;).

Despite these advances, the application of text classification in geotechnical engineering—particularly for processing raw site investigation documents—remains limited. This is surprising given that the thorough review and interpretation of site investigation reports is a foundational step in most geotechnical design projects. These reports typically include multiple forms of test data, most notably borehole logs, which contain essential subsurface information such as standard penetration test (SPT) N-values. For large-scale construction projects, borehole logs often span hundreds of pages, and unfortunately, their formats vary widely depending on the client, contractor, and country. As a result, the extraction of key geotechnical parameters from these reports is still predominantly handled manually by engineers. The underlying reason is that the reports are formatted in ways that are not readily interpretable by machines—they remain unstructured and inconsistent.

This study applied text classification to automate data extraction from unstructured borehole log reports, aiming to improve geotechnical engineering workflows. Non-English terms were translated using OpenAI's Python GPT API, guided by geotechnical and geological context. Word embeddings were trained on a domain-specific corpus of SCI articles and technical dictionaries related to site investigation, enabling linkage between English knowledge sources and non-English report content. Finally, machine learning models were trained to automatically identify relevant headings based on geotechnical domain knowledge.

2 DATASET DESCRIPTIONS

To develop a high-quality corpus that adequately reflects the linguistic and technical nuances of the geotechnical engineering domain, this study systematically collected text data from multiple reliable sources. First, a substantial volume of scientific literature was gathered, focusing on peer-reviewed articles indexed in the Science Citation Index (SCI). These publications were chosen specifically to reflect formal academic writing styles commonly used in the field. In addition, geotechnical terminology dictionaries were incorporated into the dataset to ensure comprehensive coverage of standardized terms used in site investigations and design.

Metadata from SCI journal articles—including titles, digital object identifiers (DOIs), and access permissions—was retrieved using the Elsevier Developer Portal API. The initial data crawling process targeted approximately 800 articles using general keywords such as “geotechnical engineering,” while explicitly filtering out articles associated with environmental topics to narrow the focus. To further enrich the dataset with content directly relevant to borehole logging and in-situ testing, an additional 400 article records were obtained using the keywords “standard penetration” or “borehole.” Finally, a query identified approximately 5,000 more articles that contained the phrase “standard penetration” within the main body of the manuscript. As a result, nearly 6,000 metadata records were curated, each representing a publication related to geotechnical site investigation practices.

For each article, the full text was retrieved using its DOI. If the article was open access, its abstract and main content—particularly sections such as the introduction, methodology, and results—were downloaded. In contrast, only the abstract was extracted for non-open access articles. After collection, a thorough text cleaning process was conducted to remove non-alphabetic characters and formatting artifacts, such as markup tags (e.g., <ce:section id> or </ce:label>) that are commonly found in XML-formatted scientific papers. These elements do not convey semantic meaning and were therefore excluded. In addition, notations related to in-text citations (e.g., "et al." and publication years) and references to tables or figures were also removed to avoid noise, as such expressions are overly specific and do not contribute significantly to semantic context.

Following the collection and cleaning of full-text documents, a multi-stage natural language processing (NLP) pipeline was applied to prepare the corpus for downstream analysis. This preprocessing workflow included several standard steps: sentence segmentation, tokenization, part-of-speech (PoS) tagging, lemmatization, and stop word removal. During the sentence segmentation stage, the cleaned text was divided into individual sentences using punctuation markers such as periods. Each sentence was then split into individual word tokens based on whitespace delimiters.

PoS tagging was carried out using the Natural Language Toolkit (NLTK) library, allowing the system to assign grammatical roles—such as noun, verb, or adjective—to each word. This tagging step was critical for enabling lemmatization, a process in which words are reduced to their base or dictionary forms. For instance, plural nouns were converted into their singular counterparts, and verb conjugations were normalized. The lemmatization process leveraged WordNet, a widely used lexical database for English, to ensure consistency and semantic accuracy. Subsequently, all remaining punctuation marks and common stop words—such as “a,” “the,” “is,” and “be”—were removed to reduce noise. Through this sequential NLP pipeline, a large and domain-specific text corpus was constructed, suitable for training word embedding models tailored to the site investigation domain.

In parallel with the collection of scientific texts, borehole log reports written in Korean were obtained from various geotechnical projects conducted in South Korea (Figure 1). These documents were diverse in format, representing nine different reporting templates typically used by engineering firms, public institutions, or academic research groups. The borehole reports were provided in PDF format and spanned approximately 800 pages in total. These PDFs contained a combination of text-based and image-based elements, including tabular test results, stratigraphy records, and project metadata.

The diversity and complexity of these borehole log formats underscore the importance of creating an automated system capable of handling unstructured geotechnical data. The dataset assembled in this study thus serves as a foundation not only for training NLP models but also for validating the effectiveness of

automated information extraction and classification pipelines applied to real-world engineering documentation.

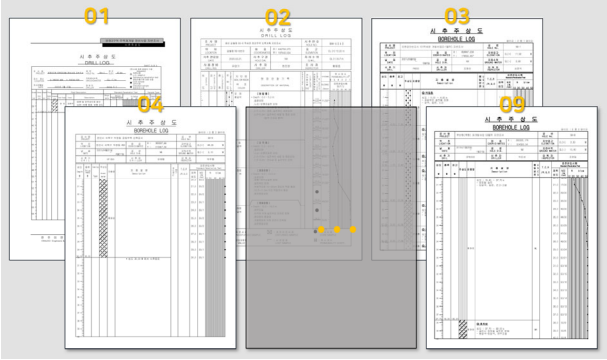

Figure 1. Borehole log reports

3 RESULTS

3.1 *SCI Geotechnical corpus analysis using GLoVe*

GloVe learns word embedding vectors by applying matrix factorization to a word co-occurrence matrix constructed from the corpus. It counts how often pairs of distinct words appear together and weights these frequencies based on their distance within a context window, which defines how far apart words must be to be considered contextually related.

From the site investigation corpus, GloVe constructed a vocabulary composed of nearly 2M words. Among them, the word “use” appeared most frequently. “model” and “soil” each occurred nearly 10M times, followed by “show” with 7.3M times.

Unlike raw word counts, co-occurrence frequencies in GloVe are non-integer values, as they are weighted by the distance between word pairs within a predefined context window. These weighted frequencies are stored in the co-occurrence matrix, which forms the basis for generating the GloVe embeddings. Table 1 presents sample co-occurrence frequencies for a word “standard” along with their top 10 semantically related terms ranked by GloVe using a context window size of 10.

Since the corpus was mainly obtained from SCI articles including site investigation contents, particularly those focused on the Standard Penetration Test (SPT), the word most frequently co-occurring with “standard” was “penetration,” rather than “deviation,” which would typically dominate in general statistical contexts.

Table 1. Co-occurrence frequency of geotechnical corpus.

Given word	Relevant word	Co-occurrence frequency
standard	penetration	6566.28
	spt	6415.63
	test	5551.88
	deviation	4370.41
	astm	1819.34
	use	1446.52
	value	1430.32
	mean	1401.37
	method	1268.04
	soil	854.16

3.2 *Borehole log report corpus analysis using Kiwi*

To analyze the linguistic characteristics of borehole log reports, approximately 800 pages of documents—written in nine different formats—were collected and processed. Text content was extracted and tokenized using Kiwi, a morphological analyzer for Korean. After tokenization, around 600 unique words were identified, representing the working vocabulary used across the borehole logs. This relatively limited lexical diversity reflects the repetitive and domain-specific nature of these reports. In contrast, the SCI-level geotechnical engineering corpus, analyzed using the GloVe embedding model, contained a significantly broader vocabulary, comprising nearly 20,000 distinct terms. This disparity highlights the constrained linguistic structure of borehole logs compared to academic literature, which tends to exhibit greater terminological variety and contextual richness.

The results showed the most frequently found words as follows: 'weathered', 'joint plane', 'depth', 'fissure', 'rock', 'elevation', 'roughness', 'permeability', 'sand', 'moist', 'dark gray', 'brown', and 'observation'. It suggested that description about the weathering is common in borehole log reports in South Korea. This is geologically reasonable because weathered rock is typically found at shallow depths (shallower than ~5 m) in South Korea.

On the other hand, we found that information in borehole log reports is treated as unstructured data due to the differing writing orientations in Korean reports compared to typical English ones. Korean is traditionally familiar with vertical writing from top to bottom, a pattern commonly found in some Asian cultures (Figure 2). In contrast, vertical writing in English documents, when it exists, is often oriented from bottom to top. As a result, standard PDF parsing libraries failed to correctly interpret the structure and semantics of the text in Korean borehole logs. To address this issue, we developed an image-processing-based algorithm that preserves the original semantic structure. Through this process, the textual components within each table cell were correctly merged and treated as a single meaningful entry (Yoo et al., 2024).

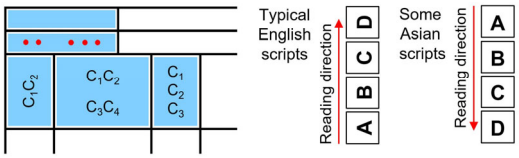

Figure 2. Differences in writing directions

3.3 *Data processing using text classification*

GloVe (Global Vectors for Word Representation) is a word embedding model that learns fixed-dimension vector representations for words based on their co-occurrence frequencies in a large corpus. Unlike prediction-based models like Word2Vec, GloVe constructs a co-occurrence matrix that captures how often each word appears near others in a given context window. It then optimizes word vectors so that the relationship between words is reflected in their vector similarities. As a result, words that appear in similar contexts are mapped to nearby points in the vector space, preserving semantic relationships.

The 200-dimensional word embeddings generated by the GloVe model, trained on a geotechnical domain-specific corpus, were projected into a two-dimensional space using principal component analysis (PCA) for visualization (Figure 3). The result reveals that semantically related terms are positioned in close proximity, forming distinct clusters that reflect their contextual associations in geotechnical engineering.

For example, structural geology terms such as orientation, plane, joint, and fault are grouped together, indicating their

frequent co-occurrence in descriptions of geological structures. Similarly, in-situ testing terms including n-value, count, blow, and penetration form a coherent cluster associated with standard penetration test (SPT) terminology. Another cluster comprises hydrogeological and foundation-related terms. These clustering patterns demonstrate that the GloVe model effectively captured domain-specific semantic relationships from co-occurrence statistics.

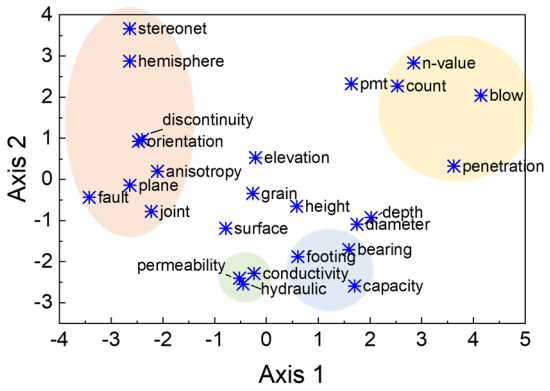

Figure 3. Visualization of word embedding in geotechnical corpus

Based on the optimized word embedding, we applied a K-nearest neighbors (KNN)–based text classification method to evaluate whether headings in borehole log reports can be effectively distinguished from one another and from other text elements. In this step, prompt-based translation using the GPT API was employed to map the raw text extracted from Korean borehole log reports to equivalent terms in an English SCI corpus.

The classification targeted SPT N-value, testing depth, description, layer elevation, boring diameter, groundwater level, ground surface elevation, borehole coordinates, client, borehole number, and project name. Each category was assigned a sequential label ranging from 0 to 10 (Figure 4). After training the KNN model with 1,000 heading samples, 400 test samples were fed into the trained model. The heading classification accuracy, using single text data in a unit cell, was 0.7233.

Major misclassifications were observed between classes 3 and 6 (green-shaded area), and between classes 4, 7, and 9 (orange-shaded area). Classes 3 and 6 correspond to layer elevation and ground surface elevation, respectively, while classes 4, 7, and 9 correspond to boring diameter, borehole coordinates, and borehole number, respectively. These results suggest that misclassification occurred due to semantic similarities between terms, such as “elevation” and “borehole”, which makes the observed confusion understandable.

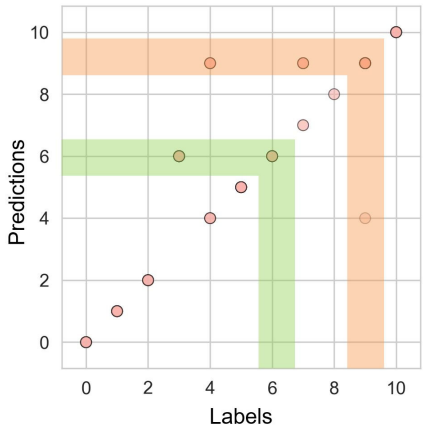

Figure 4. Classification results for headings in borehole log reports

4 CONCLUSIONS

While the formats of these reports vary, the headings and the positional relationships between headings and their associated data are generally consistent. Once the headings are detected, the dependent data can be located in predefined directions relative to the headings. Therefore, in processing tabular data, accurate detection of headings is crucial, and we aimed to achieve this through word embedding–based text classification. The current classification accuracy was approximately 0.72; however, analysis of misclassification cases indicated the need for additional methods to handle semantic similarities between terms with closely related meanings.

5 ACKNOWLEDGEMENTS

This research was supported by the KICT Research Program (Project No. 20260113-001, Database Construction for Ground Liquefaction Assessment Based on AI Technology) funded by the Ministry of Science and ICT, and partially supported by the Seoul Metropolitan Government (Project No. 20260074-001, Development of a Geotechnical Characteristics Analysis Mapping Model).

6 REFERENCES

Pham, H.T.T.L., and Han, S. 2023. Natural Language Processing with Multitask Classification for Semantic Prediction of Risk-Handling Actions in Construction Contracts, *Journal of Computing in Civil Engineering* 37(6), 04023027.

Dikmen, I., Eken, G., Erol, H., and Birgonul, M.T. 2025. Automated construction contract analysis for risk and responsibility assessment using natural language processing and machine learning, *Computers in Industry* 166, 104251.

Jeon, K., Lee, G., Yang, S., Kim, Y., and S. Suh. 2024. Dynamic building defect categorization through enhanced unsupervised text classification with domain-specific corpus embedding methods, *Automation in Construction* 157, 105182.

Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., and Harshman, R. 1990. Indexing by latent semantic analysis, *Journal of the American Society for Information Science* 41, 391–407.

Bengio, Y., Ducharme, R., Vincent, P., and Jauvin, C. 2003. A Neural Probabilistic Language Model, *J. Mach. Learn. Res.* 3, 1137–1155.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. 2013. Efficient estimation of word representations in vector space, *1st International Conference on Learning Representations, ICLR 2013,* Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings, 1–12.

Pennington, J., Socher, R., and Manning, C. 2014. Glove: Global Vectors for Word Representation, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP),* Doha, Qatar, 2014, pp. 1532–1543.

Fuentes, I., Padarian, J., Iwanaga, T., and Vervoort, R.W. 2020. 3D lithological mapping of borehole descriptions using word embeddings, *Computers & Geosciences* 141, 104516.

Chu, D., Wan, B., Ni, H., Li, H., Tan, Z., Dai, Y., Wan, Z., Tang, T., and Zhou, S. 2025. GeoSMIE: An event extraction framework for Document-Level spatial morphological information extraction, *Expert Systems with Applications* 268, 126378.

Lawley, C.J.M., Gadd, M.G., Parsa, M., Lederer, G.W., Graham, G.E., and Ford, A. 2023. Applications of Natural Language Processing to Geoscience Text Data and Prospectivity Modeling, *Natural Resources Research* 32, 1503–1527.

Yoo, B.-S., Han, J.-T., and Yang, E. 2024. Automated Geotechnical Information Extraction from Construction Boring Logs Using Keyword Groups, *KSCE Journal of Civil Engineering* 28, 4887–4896.