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/ICSMGE2026-581

A natural language processing approach to site classification

Charles John MacRobert
Department of Civil Engineering, Stellenbosch University, South Africa, macrobert@sun.ac.za

Armand Brits
PeraGage, South Africa

ABSTRACT: This paper introduces a lightweight, offline natural language processing (NLP) method for the automated classification of geotechnical soil profile descriptions, specifically aligned with South African logging standards. Soil profiles, typically described using qualitative terms, are essential for site classification and foundation design. The proposed method uses cosine similarity to match new descriptions with a labelled training dataset, enabling the assignment of both Group and Class labels. A structured classification flowchart, developed in collaboration with domain experts, guides the categorisation of stratigraphic layers based on descriptors such as moisture condition, colour, consistency, structure, soil texture, and origin. A total of 416 stratigraphic layers were classified and used to train and test the model. The algorithm achieved 76% accuracy in Group assignment, with the highest performance observed in expansive and collapsible profiles. Class-level accuracy was lower, averaging 55%, with notable challenges in distinguishing between collapsible sub-classes due to limited descriptive variation. A confidence measure was introduced by evaluating the frequency of assigned labels among the top ten cosine similarity matches, providing a qualitative indication of prediction reliability. Statistical analysis confirmed that higher confidence scores were significantly associated with correct predictions, suggesting their utility in supporting decision-making. To facilitate practical application, a Shiny web interface was developed, allowing users to input soil descriptions and receive predicted classifications along with a qualitative confidence rating. The tool is designed for ease of use and offline functionality, making it suitable for integration into geotechnical workflows. While the current model demonstrates strong potential, future improvements should focus on expanding the training dataset, refining confidence score boundaries, and developing automated methods for determining dominant stratigraphic layers within profiles. This work contributes a novel, accessible solution for enhancing consistency and efficiency in geotechnical site classification using NLP.

KEYWORDS: Qualitive soil descriptions, Natural language processing.

1 INTRODUCTION

Geotechnical information is often qualitative, comprising of standardised descriptions of the ground profile. These descriptions are then used to categorise sites and, in some cases, obtain quantitative parameters to predict behaviour. Natural Language Processing (NLP) is a field of artificial intelligence that helps computers understand and work with human language. Today, NLP tools cannot only handle large amounts of text but can also help turn qualitative textural data into quantitative parameters. This makes it easier for researchers to explore complex topics by combining statistics with deeper, human-centred insights (Dunivin 2025).

NLP allows computers to understand human language by converting text into numbers, a process known as text vectorisation. Common techniques include TF-IDF (Term Frequency-Invers Document Frequency), which weights words based on their importance in a document, and word embeddings like Word2Vec, which capture the meaning of words in context. These methods are essential for tasks like sentiment analysis, classification, and topic modelling, especially when working with large datasets (Wendland, Zenere, and Niemann 2021).

In South Africa codes of practice required that test pits are excavated to evaluate ground conditions for housing developments. The stratigraphy is described using specific terms provided by the Guidelines for Soil and Rock Logging in South Africa published by the South African Institute of Engineers and Engineering Geologists (Brink and Bruin 2002). Each soil layer is described according to its moisture condition, colour, consistency, structure, soil texture, and origin (MCCSTO).

These descriptors provide information for assessing soil behaviour and engineering properties. For example, moisture condition reflects relative water content and varies with soil type; colour can indicate expansive or collapsible soils; consistency relates to soil strength and permeability; structure describes macro features like fissures or shattering; texture gives an indication of grain size; and origin is indeed from the geomorphic or geological setting (Jennings, Brink, and Williams 1973).

According to the Geotechnical Site Investigations for Housing Developments (GFSH-2, 2002) and SANS 10400-H (2012) standard, a provisional site classification must be determined by interpreting the soil profile and making foundation recommendations based on the identified site class. Site classes are derived from an estimation of the range of expected soil movement experienced by single-storey and double-storey type 1 masonry buildings, where the foundation width is limited to 0.6 m for single-storey buildings and 0.8 m for double-storey buildings and the load on the foundation does not cause the soil bearing pressure to exceed 50 kPa. Table 1 lists the different site classes with their respective expected soil movements.

This paper presents a lightweight, offline NLP method for the automated classification of soil profile descriptions using cosine similarity. Cosine similarity is a mathematical measure used in NLP to determine how similar two pieces of text are by comparing the angle between their vector representations. It is widely applied in tasks such as text classification, summarisation, and information retrieval, where it helps identify semantically related documents or sentences (Li and Han 2013).

Table 1. Site classes.

Nature of founding material	Expected soil movement (mm)	Site class
Stable	Negligible	R
Expansive	<7.5	H
	7.5 to 15	H1
	15 to 30	H2
	>30	H3
Compressible and collapsible	<5	C
	5 to 10	C1
	>10	C2
Compressible	<10	S
	10 to 20	S1
	>20	S2
Fill (Uncontrolled)	Variable	PU
Fill (Compacted)	Variable	PC

2 METHOD

2.1 *Classification*

A structured method was employed to classify soil layers within the profiles obtained from test pit logs, ensuring consistency in the labelled data used to train the machine learning models. A flowchart was developed in collaboration with domain experts to guide the classification process. This flowchart follows a step-by-step sequence of logical checks, beginning with the identification of the dominant soil texture to determine the appropriate site class (Brits and MacRobert 2025).

Profile descriptions describing rock horizons were classified as stable (Class R). Profiles with Mudrock were excluded as these can swell and are not considered as class R.

For clay-dominant soils (Figure 1), the initial assessment involved determining whether the clay was potentially expansive, based on its origin. If the origin was not indicative of expansive behaviour, settlement was considered more critical than heave, and the classification followed the flowchart for Group S classes. The structure of the clay was then evaluated. Where slickensiding and shattering were observed (Netterberg 2019), the layer was assigned to Class H3. For fissured and intact layers, soil colour was used to guide classification. Fissured clays that were black, dark grey, maroon, or mottled were assigned to Class H2; otherwise, they were assigned to Class H1. Intact clays with these colours were also assigned to Class H1. If the intact clay did not exhibit these colours, its moisture condition was considered. Very moist or wet clays with low expansive potential were assumed to undergo greater settlement than heave and were therefore assigned to Group S classes.

The classification process for Group C classes was based on the typical field characteristics of collapsible soils (Figure 2). These soils were generally identified as clayey or silty sands with low moisture content, dense consistency, and a pinhole-voided structure. The origin of the sand was also considered as part of the classification criteria. Collapsible soils typically originate from various transported materials and certain residual soils, such as those derived from the granitic rocks of the Basement Complex (Schwartz 1985).

Where these conditions were not met, the classification followed the criteria for Group S classes. Subdivision within Group C classes were determined by evaluating the soil's consistency and the nature of its voids. Soils assigned to Class C were loose and weakly voided. Class C1 soils were medium-dense with weak voiding, while Class C2 soils were dense and exhibited a pinhole-voided structure.

For silts, sands, and clays that did not fall under Groups C or H respectively (i.e., Group S classes, see Figure 3), classification was based solely on the consistency of the soil. Soils described as very loose to loose, or very soft to soft were classified as S2, medium dense or firm were classified as S1, and dense to very dense, or stiff to very stiff as S. Where consistency was described as a range (for example, medium dense to dense), the lower bound of the range—medium dense—was used for classification purposes.

When classifying fill material, a distinction was made between controlled and uncontrolled fill. Controlled fill, such as compacted engineered material, was assigned to Class PC. Uncontrolled fill, including construction rubble and end-tipped soil, was assigned to Class PU.

A total of 416 labelled descriptions were classified. These were distributed as follows: R – 28, H – 29, H1 – 35, H2 – 25, H3 – 44, C – 26, C1 – 40, C2 – 28, S – 33, S1 – 35, S2 – 40, PU – 28 and PC – 24. These classifications were based on separate stratigraphic layers. A soil profile would be comprised of several stratigraphic layers, and a decision would still be required to determine which stratigraphic layer dominates behaviour for a class. An automated approach to this final decision is not available yet.

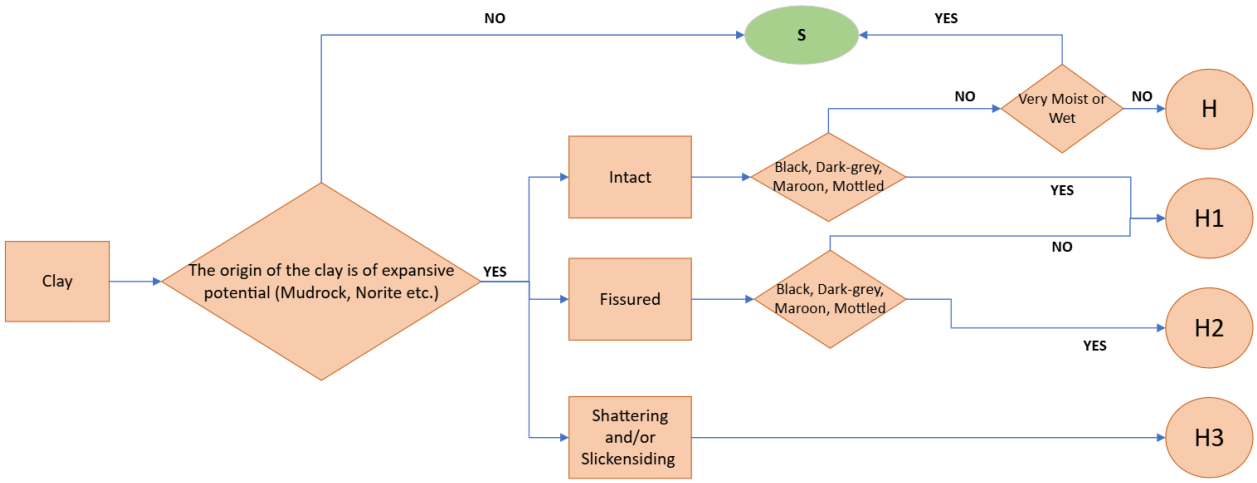

Figure 1. Classification flow-chart for class H, H1, H2 and H3

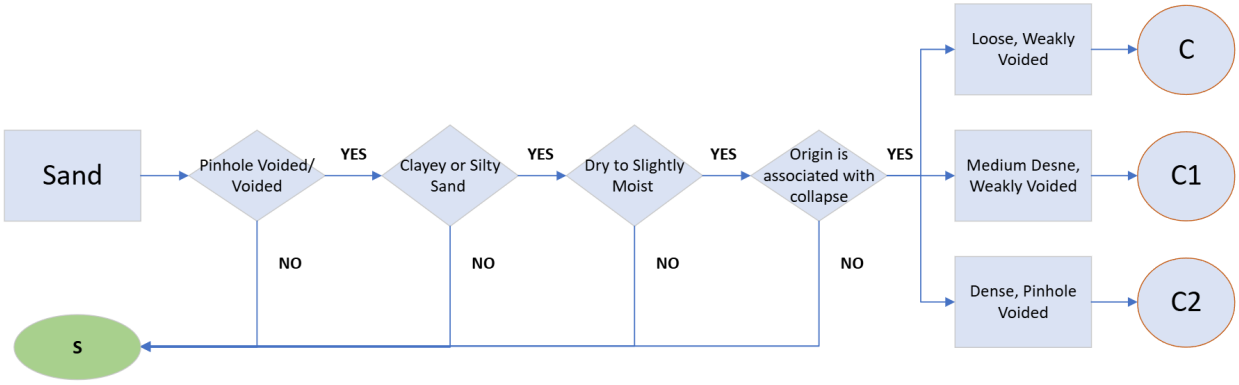

Figure 2. Classification flow-chart for Class C, C1 and C2

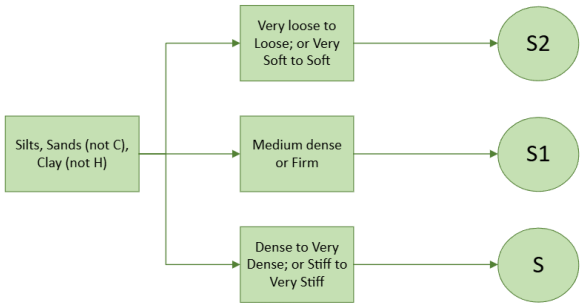

Figure 3. Classification flow-chart for Class S, S1, S2

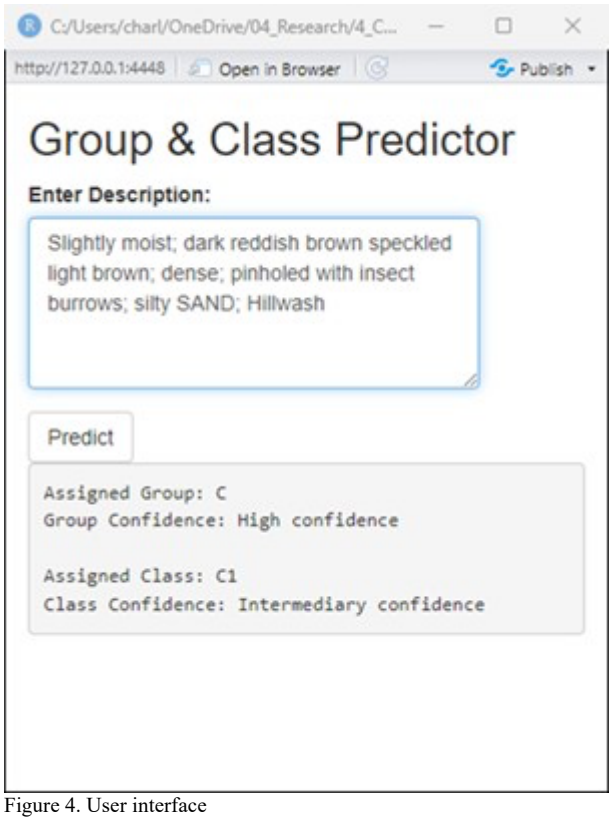

Figure 4. User interface

This total data set was separated into a training data set (80%) and testing data set (20%) with a similar distribution of classes in each. Exploratory modelling showed that performance was better for obtaining the first term in the site classification (e.g., H or C) and therefore a sub-categorisation, termed a Group, was introduced.

2.2 *Natural language processing algorithm*

Traditional natural language processing algorithms like Support Vector Machine (SVM) learning, Decisions Trees and Random Forests were initially used. However, as these rely on finding a key bag-of-words associated with a site class they become sensitive to term mismatches (i.e., if a new description does not contain a key word, it cannot be classified.

This study developed a lightweight, offline method for classifying soil profile descriptions using basic natural language processing techniques and cosine similarity. The method was implemented using base R and a minimal set of tidyverse packages, making it suitable for offline use and integration into geotechnical workflows.

Descriptions were pre-processed by converting text to lowercase, removing punctuation, and tokenising into individual words. No stemming or lemmatisation was applied, preserving the engineering terminology. A vocabulary was built from the training data, and each description was vectorised using term frequency (TF) counts over this vocabulary.

Cosine similarity was calculated between each test description and all training descriptions. The training description with the highest similarity score determined the assigned Group and Class. The classification groups used were R, H, C, S, and P, with further categorisation into site Class.

To assess classification consistency, the top ten assigned labels were captured, and the frequency of the assigned Group and Class within this subset was recorded. This provided a confidence measure for each assignment.

3 MODEL PERFORMANCE

3.1 *Group assignment*

The developed algorithm was able to correctly assign 76% of the Groups within the test data set. Performance was best for Group C (collapsible) where 86% of classifications were correct. Performance was similarly good for Group H (expansive) with 83% correct and Group P (fill) with 80% correct. Performance was worst for Group R (stable) with 60% correct and Group S (compressible) with 64% correct.

As Group R did not contain any sub-categorises it had had fewer labelled descriptions in the training data. Having more Group R descriptions could improve the performance. There is more ambiguity to assigning stratigraphic descriptions in Group S than there is in Groups C or H due as the expansive clays and collapsible sands having more clearly identifiable aspects than compressible sands and silts. Consequently, there is less of a pattern to descriptions within the Group S.

3.2 *Class assignment*

Performance in assigning a Class was poorer than the performance in assigning a Group. On average the correct Class was assigned to 55% of the test data. Performance was best within the expansive soils (H, H1, H2 and H3) where correct Classes were assigned in 63% if the cases. Performance was similarly good for Fill (PU and PC) with 60% correct, for Stable (R) with 60% correct, and compressible (S, S1, S2) with 55% correct. Performance was poor for assigning Group C classes (C, C1 and C2) where correct assignments were only made in 36% of the test cases. For this group the only significant difference in mapping descriptions to classes was the consistency with other terms being similar. Thus, there was little difference for a pattern to be ascertained. However, having more descriptions should improve performance.

3.3 *Confidence measure*

To evaluate the consistency of the classification, the top ten assigned labels with the highest cosine similarities were recorded, and the frequency of the assigned Group and Class within this subset determined. For example, if C1 was assigned and C1 appeared 6 times in the top ten assigned classes this would give a confidence score of 6.0 for a C1 prediction. The average confidence score for correct and wrong predictions was evaluated to determine if this score would be a useful aide to decision making.

Table 2 shows that the average confidence scores were significantly different between correct and wrong predictions for both Group and Class assignments. This suggested that if a confidence score above 6.0 was achieved there is a high level of confidence that the correct group and class have been assigned (The average confidence score for correct groups was 6.5 and correct classes was 4.2). On the other hand, if this score was below or equal to 3.0 the prediction was with a low level of confidence (The average confidence score for wrong group assignments was 4.4 and correct classes was 3.2). Confidence scores between these, carry an intermediary level of confidence. These boundaries will need to be revaluated for a larger training set.

Table 2. Confidence measure.

Parameter	Value
Average confidence score for correct group assignments	6.5
Standard deviation	2.4
Average confidence score for wrong group assignments	4.4
Standard deviation	2.4
p-value: two-tailed Student's t-test (equal variance) on confidence scores for correct and wrong groups	0.002
Average confidence score for correct class assignments	4.2
Standard deviation	2.0
Average confidence score for wrong class assignments	3.0
Standard deviation	1.6
p-value: two-tailed Student's t-test (equal variance) on confidence scores for correct and wrong groups	0.007

4 DEVELOPED APPLICATION

A Shiny web application was developed to provide an interactive interface for the classification of soil profile descriptions. Users enter the soil layer description, and the app returns a predicted Group and Class label, accompanied by a qualitative confidence rating (High, Intermediary, or Low). The interface is simple and user-friendly, consisting of a text input area and a button to initiate prediction. Results are displayed directly within the app. A screenshot of the interface is provided in Figure 4, illustrating the layout and functionality of the tool.

5 CONCLUSIONS

This study presents a practical, offline natural language processing approach for the automated classification of geotechnical soil profile descriptions, tailored to South African logging standards. By combining domain-informed classification logic with cosine similarity, the method effectively translates qualitative stratigraphic data into structured site classes. Group-level classification achieved promising accuracy, particularly for clay and sand-dominant profiles, while Site Class-level performance highlighted areas for improvement, especially within the collapsible sand group. The introduction of a confidence score offers a valuable decision-support tool, helping users gauge the reliability of predictions.

The development of a Shiny web application further enhances accessibility, enabling geotechnical practitioners to interactively classify soil descriptions with qualitative confidence feedback. The tool's simplicity and offline functionality make it well-suited for integration into existing workflows. While the current model shows strong potential, future work should focus on expanding the training dataset, refining confidence boundaries, and developing automated methods for determining dominant stratigraphic layers within profiles to improve usability. This would enable an application to be deployed for industry use. This could also include a feedback mechanism to provide an additional metric to improve predictions on an ongoing basis.

6 ACKNOWLEDGEMENTS

Professor Peter Day (Stellenbosch University, Jones and Wagner) and Mr. Tony A'Bear (Bear Geo Consultants) are acknowledged for their guidance in developing the classification flowcharts.

7 REFERENCES

Brink, A.B.A., and R.M.H. Bruin. 2002. *Guidelines for soil and rock logging in South Africa* (AEG-SA, SAICE, SAIEG).

Brits, L.A., and C. J. MacRobert. 2025. "Machine-learning approach to site classification." In *2nd Southern Conference African Geotechnical*, edited by S.W. Jacobsz, 55–60. Durban, South Africa: SAICE Geotechnical Division.

Dunivin, Z.O. 2025. 'Scaling hermeneutics: a guide to qualitative coding with LLMs for reflexive content analysis', *EPJ Data Science*, 14: 28.

Jennings, J.E., A.B.A. Brink, and A.A.B. Williams. 1973. 'Revised guide to soil profiling for Civil Engineering Purposes in SA', *The Civil Engineer in South Africa*, 5: 3–12.

Li, B., and L. Han. 2013. "Distance Weighted Cosine Similarity Measure for Text Classification." In, 611–18. Berlin, Heidelberg: Springer Berlin Heidelberg.

Netterberg, F. . 2019. "Identification of potentially expansive clay soils from soil structure." In *17th African Regional Conference on Soil Mechanics and Geotechnical Engineering*, edited by S.W. Jacobsz. Cape Town: SAICE.

Schwartz, K. 1985. 'Collapsible soils', *Civil Engineering*, 27: 379–93.

Wendland, A., M. Zenere, and J. Niemann. 2021. "Introduction to Text Classification: Impact of Stemming and Comparing TF-IDF and Count Vectorization as Feature Extraction Technique." In, 289–300. Cham: Springer International Publishing.