

# Application of Natural Language Processing in Detecting New Critical Mineral Deposits: BC Carbonatite Case Study

*Afshin Amini, James R Barlow, Karl E Flower, Travis MacLean and Nicole D Barlow  
Purple Rock Inc.*

## Introduction

Critical minerals are minerals essential to the economy and whose supply may be disrupted. They present a generational opportunity for Canada's workers, economy, and net zero future in areas such as exploration, extraction, processing, downstream product manufacturing and recycling. One deposit type for a number of critical minerals is carbonatite, which is a rare type of igneous rock that (generally) contains 50% or greater by volume of carbonate minerals. Carbonatites can occur in extrusive, oceanic, or subduction zone environments, but they are predominantly found as mantle-derived plume intrusions in continental extension zones, and typically appear within a larger alkaline igneous complex. They have been found on every continent with ages spanning from 3 Bya to present day. Carbonatites are a major source of rare earth element deposits and can hold ore-grade levels of Cu, Fe, and fluorite, among others (Simandl and Paradis, 2018). Although they are rare, with only approximately 600 occurrences known worldwide, they are geographically widespread, with known occurrences in British Columbia. It is suspected that there are more undiscovered deposits in the province as well.

The BC Geological Survey (BCGS) houses a database of more than 40,000 mineral assessment reports (ARIS, the Assessment Report Indexing System; BC Geological Survey, 2023a), which describes mineral exploration activities throughout the province. The BCGS also has a supporting document database (Property File; BC Geological Survey, 2023b) with more than 70,000 archival geoscience documents. There are reports in this database that might not directly relate to carbonatite deposits but may contain information that can point to new carbonatite deposits. It is time consuming and costly for a knowledge specialist to read thousands of documents; therefore, other methods that can speed up/automate this process while maintaining high accuracy are needed. In this work, we describe the development of a natural language processing (NLP) pipeline to automate the process of reviewing the content of 110,000 geological reports and identifying candidates that point to new possible carbonatite deposits in British Columbia.

## Methodology

This project uses state-of-the-art NLP techniques to process and analyze the unstructured text ('natural language') of reports. Natural language processing is a branch of artificial intelligence (AI) that enables computers to comprehend, generate, and manipulate human language. NLP combines computational linguistics—rule-based modeling of human language—with statistical, machine learning, and deep learning models to process human language in the form of text or voice data and to 'understand' its full meaning. In natural language processing, word embedding is a representation of a word in vector space; this representation is a real-valued vector that encodes the meaning of the word in such a way that words that are closer to each other in the

vector space are expected to be similar in meaning. Word embedding methods can be divided in two categories: non-contextual word embedding models such as Global Vectors (GloVe; Pennington, 2014), Word2Vec and N-Gram and contextual word embeddings like BERT and GPT. Although the latter category of models is generally considered to be more advanced, models like GloVe that are trained using the frequency of co-occurring words have proven effective at encoding words with similar meaning as closely associated numerical vectors. These models capture text semantics based on the statistical distribution of words and, in some domain-specific applications such as geosciences, they can even outperform more advanced language models (Lawley et al., 2022).

For the purposes of this project, we used the geoscience GloVe model (Lawley et al., 2022) to analyze geoscientific reports. geoscience GloVe is a GloVe model that is retrained on a smaller subset of public geoscientific documents sourced from the Natural Resources Canada (NRCAN) publication database (GEOSCAN), Canadian provincial geological survey publication databases (i.e., Ontario, Alberta and British Columbia), and open-source peer-reviewed publications and therefore it is specifically tuned for analyzing geoscientific texts. Both GloVe and geoscience GloVe models are based on the assumption that words occurring together are more closely related.

Text processing and analysis is performed in three steps: First, open-source NLP tools are used to process unstructured text data (geological reports). Text processing was completed using spaCy, an open-source software library for advanced natural language processing, written for the Python programming language (Honnibal and Montani, 2017). Our text processing workflow is carried out using three NLP tasks: (1) tokenization, (2) removing stop words, and (3) stemming. Next, the geoscience GloVe model is used to calculate a representative vector for each report. Each of the ~400k words in the geoscience GloVe model is associated with a 300-dimensional numerical vector. In this analysis, individual words within the processed text were joined with their corresponding vector before calculating a representative vector for each report.

Finally, the Euclidean distance and cosine similarities of each report with a list of keywords (defined by knowledge specialists) is calculated to identify the reports that are semantically close to the list of keywords. The main challenge with this task is the lack of a labeled dataset. To overcome this challenge, several document analysis methods were designed and tested on assessment reports. Next, top results from each method were then reviewed by a team of geologists to determine the accuracy of each method. Finally, the best performing method was selected to find candidate document from the Property File database and was further analyzed by our geologists.

## **Analysis and Results**

A total of 16 methods (measures of similarity) were designed and tested to find the most suitable method for identifying reports with the potential to point to new carbonatite deposits. The top results from each method were selected and then reviewed by our knowledge specialists to identify whether a document contains information related to carbonatite deposits. All identified reports were then categorized into three groups of 'Yes', 'No' or 'Maybe', indicating whether a report is pointing to a carbonatite deposit (Yes) or not (No) or if there is a possibility, but further investigation is required (Maybe).

Using this categorization, a dataset of 150 reports was constructed, indicating the usefulness of each report. This dataset was used to further analyze the performance of each method (Figures 1 & 2) and the best performing method was identified and applied on a second dataset of 70,000 Property File documents, which resulted in 90 more candidate reports for our geologists' review. Overall, using this pipeline, 109,000 reports were analyzed and a subset of 241 reports were identified as target reports for the geologists' review. Out of these reports, 100 reports were pointing to already discovered carbonatite deposits and 68 were pointing to new target areas for possible carbonatite deposits.

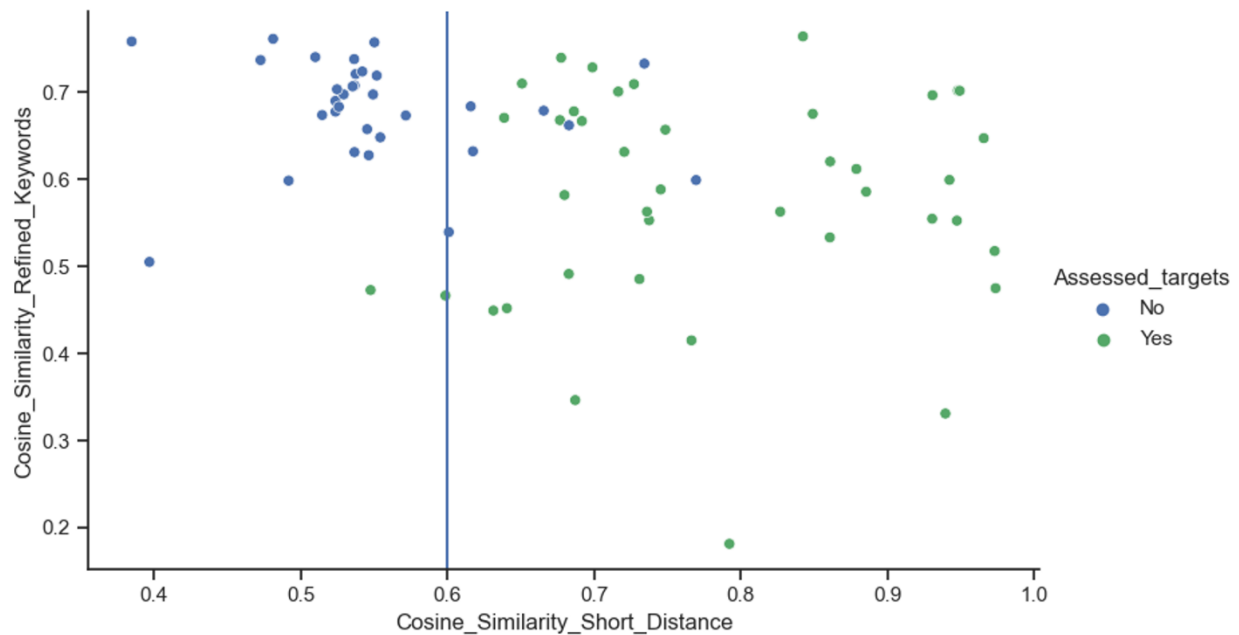


Figure 1- Result of classifying documents based on their similarity with the word 'carbonatite'. Only positive and negative groups are shown for simplicity. Vertical blue line represents the empirically derived cut-off.

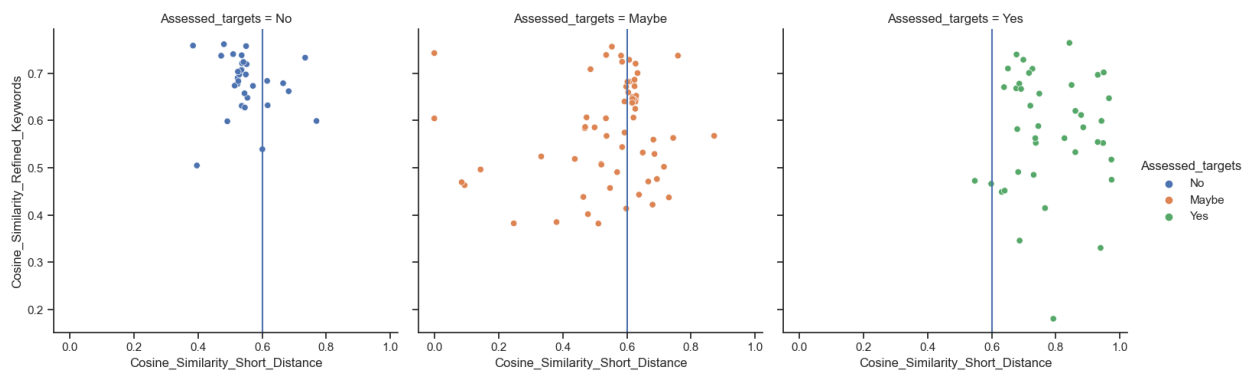


Figure 2- Result of classifying documents based on their similarity with the word 'carbonatite' separated by each category. Vertical blue line represents the empirically derived cut-off.

## Conclusions

In this research project, we investigated the application of natural language processing in analyzing geoscientific reports to identify new carbonatite target areas in British Columbia. This study is a promising showcase of the potential of application of natural language processing techniques in the context of information extraction from geoscientific documents. The techniques used can be extended to other deposit types, topics and targets in any jurisdiction or project with a collection of historical documents. This is done by increasing the speed of document processing, which greatly reduces the number of reports that need to be checked by a knowledge specialist and therefore reducing costs.

## References

- BC Geological Survey (2023a): Assessment Report Indexing System (ARIS); BC Ministry of Energy, Mines and Low Carbon Innovation, BC Geological Survey, URL <<https://apps.nrs.gov.bc.ca/pub/aris/>> [last accessed January 2024].
- BC Geological Survey (2023b): Property File digital document database; BC Ministry of Energy, Mines and Low Carbon Innovation, BC Geological Survey, URL <<http://propertyfile.gov.bc.ca/>> [last accessed January 2024].
- Honnibal, M. and Montani, I. (2017): spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing.
- Lawley, C.J.M, Raimondo, S., Chen, T., Brin, L., Zakharov, A., Kur, D., Hui, J., Newton, G., Burgoyne, S.L. and Marquis, G. (2022): Geoscience language models and their intrinsic evaluation; Applied Computing and Geosciences, Vol. 14 <<https://doi.org/10.1016/j.acags.2022.100084>>.
- Pennington, J., Socher, R. and Manning, C.D. (2014): [GloVe: Global Vectors for Word Representation](#).
- Simandl, G.J. and Paradis, S. (2018): Carbonatites: related ore deposits, resources, footprint, and exploration methods; Applied Earth Science, vol. 127, no. 4, p. 123–152.