Natural Resources Research, vol.32, No. 4, August 2023 (© 2023)
https://doi.org/10.1007/s11053-023-10216-1

Original Paper

Applications of Natural Language Processing to Geoscience Text Data and Prospectivity Modeling

Christopher J. M. Lawley,[1,6] Michael G. Gadd,[2] Mohammad Parsa,[1] Graham W. Lederer,[3] Garth E. Graham,[4] and Arianne Ford[5]

Received 1 February 2023 ; accepted 8 May 2023
Published online: 2 June 2023

Geological maps are powerful models for visualizing the complex distribution of rock types through space and time. However, the descriptive information that forms the basis for a preferred map interpretation is typically stored in geological map databases as unstructured text data that are difficult to use in practice. Herein we apply natural language processing (NLP) to geoscientific text data from Canada, the U.S., and Australia to address that knowledge gap. First, rock descriptions, geological ages, lithostratigraphic and lithodemic information, and other long-form text data are translated to numerical vectors, i.e., a word embedding, using a geoscience language model. Network analysis of word associations, nearest neighbors, and principal component analysis are then used to extract meaningful semantic relationships between rock types. We further demonstrate using simple Naive Bayes classifiers and the area under receiver operating characteristics plots (AUC) how word vectors can be used to: (1) predict the locations of ''pegmatitic'' (AUC = 0.962) and ''alkalic'' (AUC = 0.938) rocks; (2) predict mineral potential for Mississippi-Valley-type (AUC = 0.868) and clastic-dominated (AUC = 0.809) Zn-Pb deposits; and (3) search geoscientific text data for analogues of the giant Mount Isa clastic-dominated Zn-Pb deposit using the cosine similarities between word vectors. This form of semantic search is a promising NLP approach for assessing mineral potential with limited training data. Overall, the results highlight how geoscience language models and NLP can be used to extract new knowledge from unstructured text data and reduce the mineral exploration search space for critical raw materials.

KEY WORDS: Natural language processing, Language model, Word embedding, Semantics, Prospectivity, Critical mineral.

INTRODUCTION

Bedrock geological maps are the standard tool for visualizing the complex spatial-distribution of rocks at surface and their associations through time (Giles & Bain, 1995; Laxton & Becken, 1996; Reed et al., 2005; Loudon, 2009; Sharpe, 2015). However, every geological map is an interpretation, based on a conceptual understanding of geological processes and data derived from field observations and laboratory analyses (Giles & Bain, 1995; Brodaric et al.,

[1] Geological Survey of Canada, Natural Resources Canada, 601 Booth Street, Ottawa, ON K1A 0E8, Canada.
[2] Geological Survey of Canada, Natural Resources Canada, 3303 33 Street NW, Calgary, AB T2L 2A7, Canada.
[3] U.S. Geological Survey, Geology, Energy and Minerals Science Center, 12201 Sunrise Valley Drive, Mailstop 954, Reston, VA 20192-0002, USA.
[4] U.S. Geological Survey, Geology, Geochemistry, and Geophysics Science Center, Denver, CO 80225, USA.
[5] Geoscience Australia, 101 Jerrabomberra Ave, Symonston, ACT 2609, Australia.
[6] To whom correspondence should be addressed; e-mail: christopher.lawley@nrcan-rncan.gc.ca

2004; Thorleifson, 2005; Whitmeyer et al., 2010; Brodaric, 2012; Mantovani et al., 2020). The digital geoscientific data used to make the preferred map interpretation are typically stored in geological map databases as numeric and/or text attributes linked to spatial polygons with geometrical information (Laxton & Becken, 1996; Laxton, 2017). Ideally, each numeric and text attribute is defined in a database model and corresponds to a different geoscientific concept (Raymond et al., 2012a; Wilson et al., 2015; Horton et al., 2017; Mantovani et al., 2020). Data in this form are ''structured'' and is relatively straightforward to represent on a bedrock map (e.g., rock types, geochronological information, polygon identification numbers), particularly if the vocabularies for each geoscientific concept correspond to international data standards like the North America Data Model (NADM), International Union of Geological Sciences (IUGS) geoscience markup language (GeoSciML), and the European data specification for geology (INSPIRE; Sen & Duffy, 2005; Simons et al., 2006; Raymond et al., 2012a; Laxton, 2017; Mantovani et al., 2020).

A significantly larger and growing proportion of geoscientific data are ''unstructured''. Such data include detailed rock descriptions, lithostratigraphic relationships, interpretations of geological processes, and other text attributes collected over many decades and at great financial expense by geological survey organizations (Wheeler et al., 1996; Reed et al., 2005; Laxton, 2017; Stephenson et al., 2022). These unstructured forms of geoscientific text data are essential to the map-building process because they contain the concepts and observations underpinning the preferred map representation (Brodaric et al., 2004; Pavlis et al., 2010; Mantovani et al., 2020). The availability of this type of long-form text data continues to grow as field computers allow geologists to digitally record their observations, and as geoscientific publications become linked to geospatial databases. For paper-based geological maps, this type of unstructured geoscientific information is typically reported in the map legend or as stratigraphic columns in the map margins. However, paper-based geological maps were not considered as part of the current study. Instead, we focused on the large volumes of text available in geodatabases and associated unstructured rock descriptions in lexicons of geologic units (e.g., WEBLEX). The large volumes of text available in a digital form make this type of unstructured data difficult to use and relatively ''inaccessible'' in practice. Text data are also difficult to visualize on maps, which, coupled with the complex usage of geoscientific terms by multiple authors over time and differences in map scale, present several practical challenges to the application of unstructured text in a geological mapping context. Text attributes that do not conform to standard vocabularies or that mix disparate concepts are sometimes referred to as ''semi-structured'' and are also difficult to use. For the purposes of this study, all forms of text data (i.e., structured, semi-structured, and unstructured) were combined prior to further analysis.

Natural language processing (NLP) is a subfield of artificial intelligence focused on interpreting human language by learning the meaning of words and sentences (i.e., text semantics; Bengio et al., 2000; Mikolov et al., 2013a, 2013b; Pennington et al., 2014; Devlin et al., 2019; Chowdhary, 2020). The application of NLP to geoscience text data has so far included summarizing articles (Ma et al., 2021), translating languages (Qiu et al., 2018; Consoli et al., 2020; Gomes et al., 2021), generating keywords (Qiu et al., 2018, 2019), and information discovery (Peters et al., 2014, 2018; Wang et al., 2018; Holden et al., 2019; Enkhsaikhan et al., 2021a, 2021b; Ma 2022; Wang et al., 2022). These and other geoscience NLP applications are possible because of recent open-source tools developed by the artificial intelligence community, improved access to high-performance cloud computing, and the increased availability of internet text-data for training state-of-the-art language models (e.g., Open AI's GPT-3 and -4; Floridi & Chiriatti, 2020; Dale, 2021).

Language models comprise the mathematical rules for representing words, parts of words, and/or sentences as numerical vectors (i.e., word embeddings) that are, in turn, used to solve other machine learning tasks (Hirschberg & Manning, 2015; Chowdhary, 2020). The best-performing NLP models are currently based on transformers, deep-learning models that can interpret the meaning of words based on their context (Vaswani et al., 2017; Devlin et al., 2019). Transformers and self-attention are used to capture long-range and bi-directional dependencies that are particularly important for representing words that have multiple meanings (i.e., polysemy). For example, the word ''rock'' may refer to a music genre or a solid-mass of minerals depending on the context. However, even simple language models, trained using the frequency of co-occurring words, have proven effective at encoding words with similar meaning as closely associated

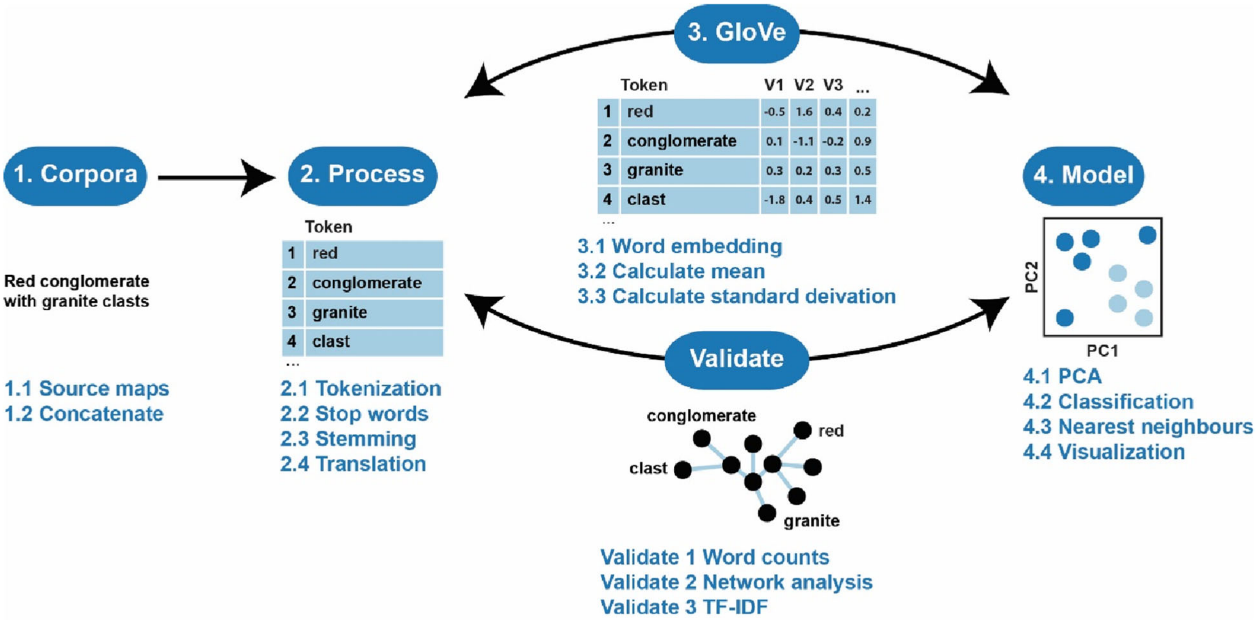

Figure 1. Text processing and modelling workflow. Geoscientific text data were sourced from four geological map databases. Multiple natural language processing methods were used to prepare this text data for further analysis (i.e., tokenization, removing stop words, stemming, and French to English translation). Word counts, word associations, network analysis, and nearest neighbors (based on cosine similarities) were used to validate the quality of the processing workflow. Tokens were then joined with the geoscience GloVe model (Lawley et al., 2022a) before calculating mean vectors for each map polygon (e.g., V1, V2, V3). The standard deviation of cosine similarities between each word and each polygon's mean vector was used to estimate map uncertainty. Principal Component Analysis (PCA) was used to reduce the dimensionality of word vectors prior to predictive modelling and visualization.

numerical vectors (Mikolov et al., 2013a, 2013b; Pennington et al., 2014; Bojanowski et al., 2017). Static word embeddings like N-Gram, Word2Vec, FastText, and Global Vectors for Word Representations (GloVE) can capture text semantics based on the statistical distribution of words and, in some domain-specific applications, can even outperform more advanced language models (Lawley et al., 2022a). Simple language models pre-trained on general internet text data can also be re-trained on smaller volume of domain-specific text, such as geoscientific publications, to improve performance for particular down-stream, geoscience tasks (Padarian & Fuentes 2019; Fuentes et al., 2020; Lawley et al., 2022a).

Herein we apply this type of geoscience language model to the task of prospectivity modeling. First, text data are combined from four geological map databases across Canada, the U.S., and Australia (Fig. 1). Second, open-source NLP tools are used to process structured, semi-structured, and unstructured text data (Fig. 1). Third, the average and standard deviation of word vectors for each map polygon (Figs. 1 and 2) are calculated using the preferred geoscience GloVe model reported in Lawley et al. (2022a; Figs. 2 and 3). Word counts, word associations, network analysis, term frequency-inverse document frequency scores, and principal component analysis are used as intrinsic evaluation methods (Fig. 1) for the text processing pipeline before applying the geoscience GloVe language model to prospectivity modeling (Figs. 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, and 14). The language model is tested on three down-stream tasks: (1) predicting the location of rare rock types, i.e., ‘‘alkalic’’ and ‘‘pegmatitic’’, based on the available rock descriptions (Fig. 10); (2) assessing the mineral potential for Mississippi Valley-type (MVT; Fig. 11) and clastic-dominated (CD; Fig. 12) Zn-Pb deposits based on the text descriptions of favorable host rocks (Lawley et al., 2022b); and (3) calculating cosine similarities between word vectors for semantic search of Mount Isa deposit analogues (Fig. 14). Each of these three down-stream applications extracts knowledge from unstructured text data and expand the applications of geological map databases for prospectivity modeling.

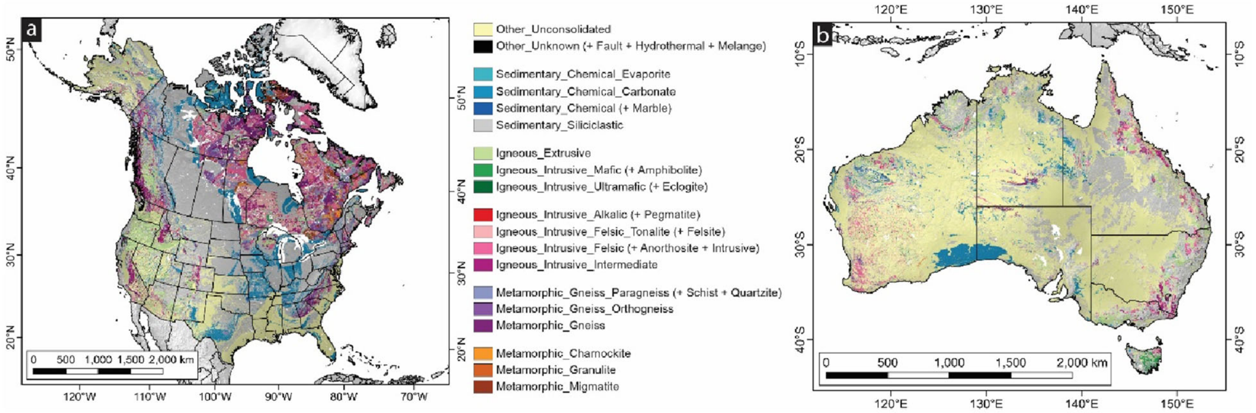
a
b
Other_Unconsolidated
Other_Unknown (+ Fault + Hydrothermal + Melange)
Sedimentary_Chemical_Evaporite
Sedimentary_Chemical_Carbonate
Sedimentary_Chemical (+ Marble)
Sedimentary_Siliciclastic
Igneous_Extrusive
Igneous_Intrusive_Mafic (+ Amphibolite)
Igneous_Intrusive_Ultramafic (+ Eclogite)
Igneous_Intrusive_Alkalic (+ Pegmatite)
Igneous_Intrusive_Felsic_Tonalite (+ Felsite)
Igneous_Intrusive_Felsic (+ Anorthosite + Intrusive)
Igneous_Intrusive_Intermediate
Metamorphic_Gneiss_Paragneiss (+ Schist + Quartzite)
Metamorphic_Gneiss_Orthogneiss
Metamorphic_Gneiss
Metamorphic_Charnockite
Metamorphic_Granulite
Metamorphic_Migmatite
0 500 1,000 1,500 2,000 km
120°W 110°W 100°W 90°W 80°W 70°W
20°N 30°N 40°N 50°N
120°E 130°E 140°E 150°E
10°S 20°S 30°S 40°S

Figure 2. (**a**) Generalized rock types for Canada, the conterminous U.S., and Alaska. (**b**) Generalized rock types for Australia. Map colors are based on the generalized rock types reported in Lawley et al. (2022b).

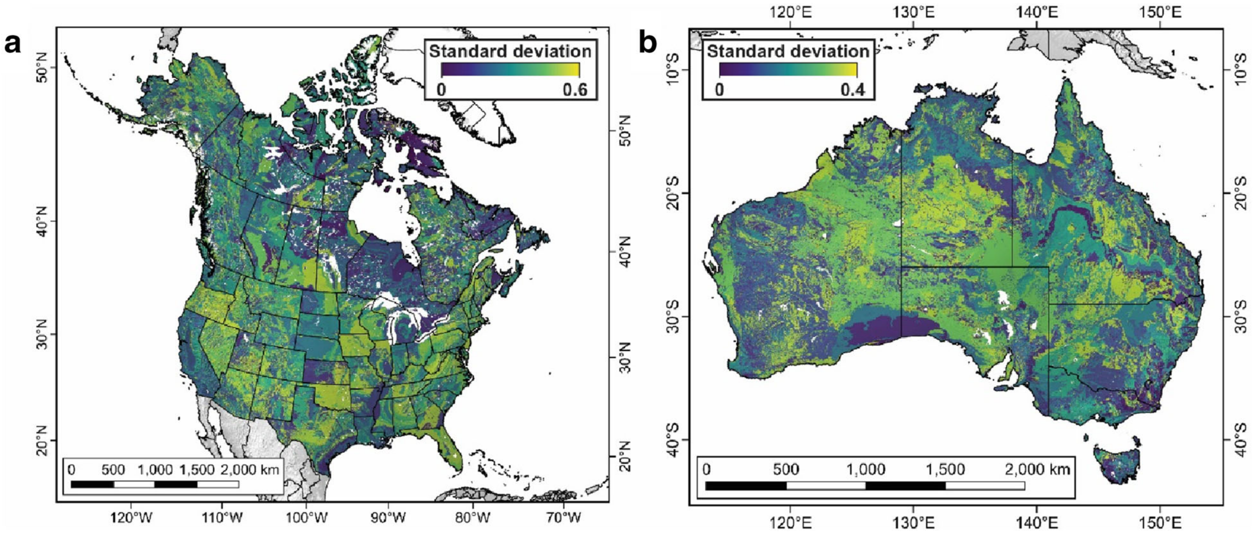

Figure 3. (**a**) Standard deviation of cosine similarities for Canada, the conterminous U.S., and Alaska. (**b**) Standard deviation of cosine similarities for Australia. Regions with higher standard deviations point to words with complex semantics and relatively high uncertainty.

TEXT DATA

Text data used in the current study were sourced from four geological map compilations (i.e., the corpora; Figs. 1 and 2). The quality, level of detail, and length of text data was variable within and among geological databases (Online Supplementary Table; Fig. 4). Overall, the conterminous U.S. and Alaska contained the longest and most detailed rock and lithostratigraphic descriptions (Fig. 4).

Data for Canada were sourced from 22 provincial and territorial geological map compilations comprising 316,579 polygons, as reported in Lawley et al. (2022b). Individual geological maps within this compilation range in scale from 1:30,000 to 1:5,000,000 (Fig. 2a). Missing data were imputed from the seamless but more generalized geological map of Canada (1:5,000,000 scale; Wheeler et al., 1996). Text data for Canada were sourced from rock types, geological periods, lithostratigraphic information, and rock descriptions. Expanded rock

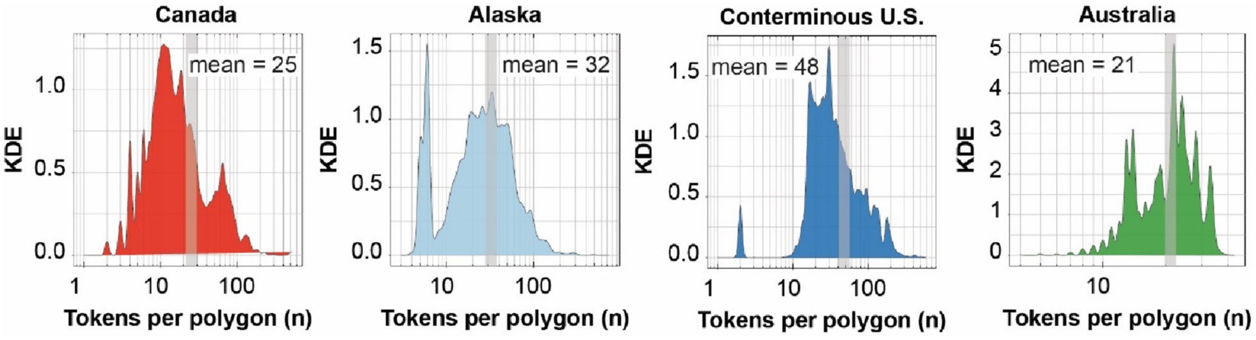

Figure 4. Kernel density estimates (KDE) for the frequency of tokens from each geological map database. Counts are based on the total number of tokens for each map polygon after processing. Overall, the conterminous U.S. and Alaska yield the longest rock descriptions; whereas Canada and Australia are associated with shorter rock descriptions.

descriptions within an online database of geological names (https://weblex.canada.ca) were concatenated with map polygons using lithostratigraphic and lithodemic names wherever possible (Lawley et al., 2022b).

Data for the conterminous U.S. were sourced from the State Geologic Map Compilation (SGMC; Horton et al., 2017). This published geodatabase comprises 313,732 polygons covering 48 states (Fig. 2a). Individual state maps used in the Horton et al. (2017) compilation have scales that range from 1:50,000 to 1:1,000,000 and are a patchwork of polygons with map boundary artifacts. Text data for the conterminous U.S. were concatenated from the unit name (i.e., UNIT_NAME), age information (i.e., AGE_MIN, AGE_MAX), major (i.e., MAJOR1, MAJOR2, MAJOR3) and minor rock types (i.e., MINOR1, MINOR2, MINOR3, MINOR4, MINOR5), generalized geology (i.e., GENERALIZE), and the long-form rock descriptions that are available through linked tables (i.e., UNITDESC).

Data for the state of Alaska were sourced from the Geologic Map of Alaska (Wilson et al., 2015). This published map compilation comprises 245,562 polygons and is based on 1:63,360 to 1:250,000 scale maps (Fig. 2a). Text data for Alaska were concatenated from the unit name (i.e., STATE_UNIT), age information (i.e., AGE_RANGE), and long-form rock descriptions available through linked tables (NSA class and unit).

Data for Australia were sourced from the national 1:1,000,000 geological map dataset (Raymond et al., 2012b). This dataset comprises 242,703 polygons and is seamless with digital attributes formatted to match GeoSciMl standards wherever possible (Fig. 2b). Text data for Australia were combined from map unit names (i.e., NAME), rock types (i.e., LITHOLOGY), and detailed rock descriptions (i.e., DESCR and GEOLHIST). It is noted that unlike the data compiled for the U.S. and Canada, the data for Australia are a national compilation and not a compilation from state- or territory-based geological map databases, which typically contain more detailed text data. Additional text data are available in the Australian Stratigraphic Units Database (ASUD). This database contains 17,500 stratigraphic names and their associated long-form rock descriptions could be used in future research similar to the Canada dataset.

Collectively, the four geological map databases comprise 1,118,576 polygons and provide complete coverage across Canada, the U.S., and Australia.

TEXT PROCESSING

All text processing was completed using the “tidyverse” (Wickham et al., 2019), “tidytext” (Silge & Robinson, 2016) and “sf” (Pebesma, 2018) packages in R (R Core Team, 2023). Virtual machines were used for the most memory intensive operations, typically using the EC2 M5 instances available through Amazon Web Services (https://aws.amazon.com), Posit Workbench, and SageMaker (Joshi, 2020). The entire text processing workflow can be simplified into three NLP tasks: (1) tokenization; (2) stop words; and (3) stemming (Fig. 1). First, the concatenated text data for each geological map database were converted to lowercase and converted to “tokens” using the “unnest_tokens” function from the “tidytext” (Silge & Robinson, 2016) and “tokenizer” packages (Lincoln et al., 2018). The tokenization process converts the concatenated text fields for each map unit into a digital table, with

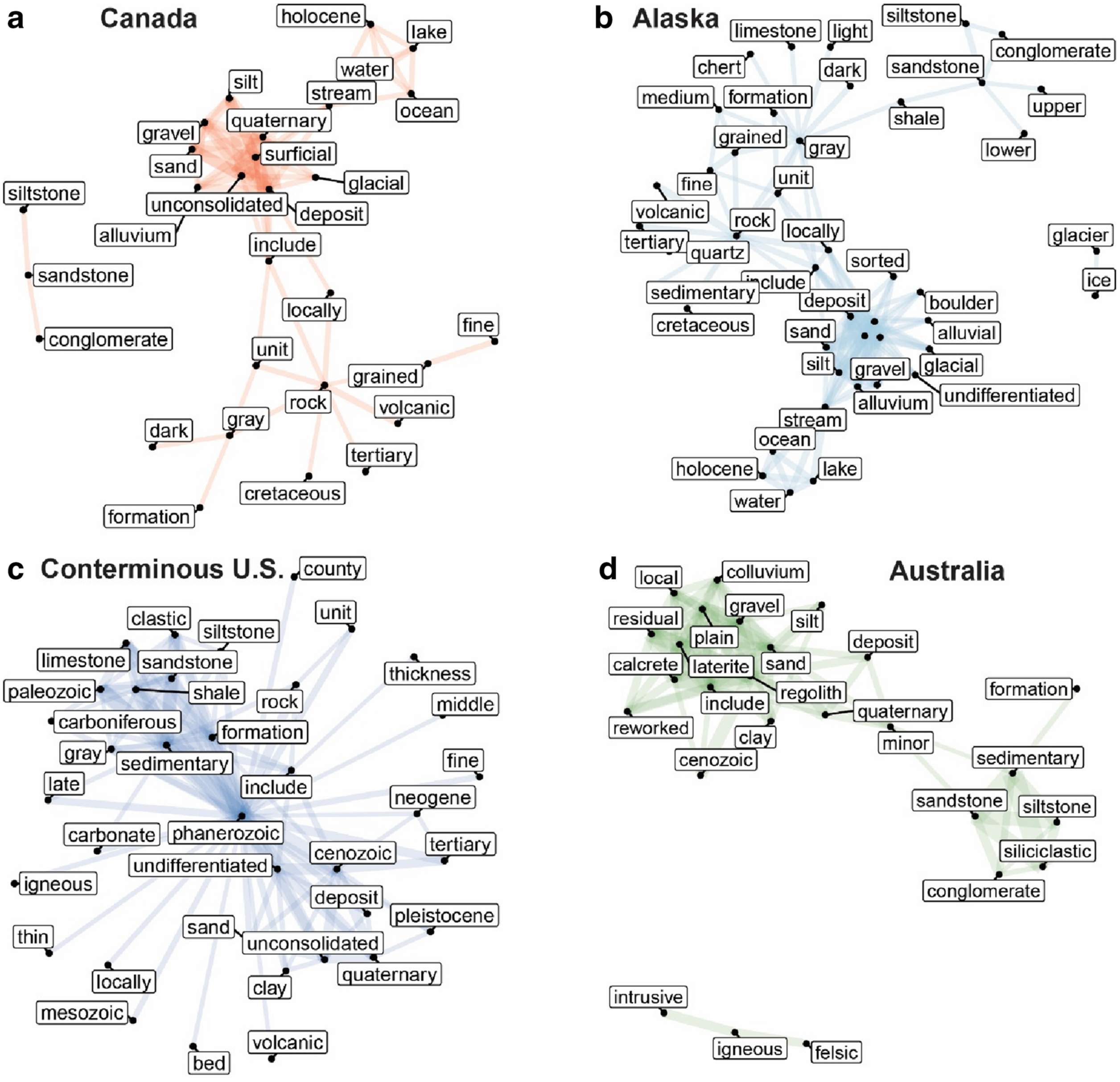

Figure 5. Network analysis of the mostly frequent token associations for Canada (**a**), Alaska, (**b**), the conterminous U.S. (**c**), and Australia (**d**). The network layout is based on the Fruchterman–Reingold algorithm to place the most common co-occurring token pairs closer together (Fruchterman and Reingold, 1991). Natural groupings of nodes represent words with similar meaning. The lines between nodes represent connections between geoscientific concepts; whereas the node thickness is proportional to counts of co-occurring words.

individual words represented as one-token-per row (Fig. 1; Step 2.1). Words separated by hyphens, other forms of punctuation, and white space were broken into separate rows by this tokenization process. Blank spaces, numbers, and punctuation were then removed from the separated tokens and excluded from further analysis. It is noted that numerical age dates, relative abundances of minerals, and any other numerical data were excluded from further analysis as part of this text processing workflow. This type of numerical information contains important information but would have been impossible to use with the geoscience GloVe model vocabulary (Lawley et al., 2022a). Tokens with fewer than 2 characters were also excluded.

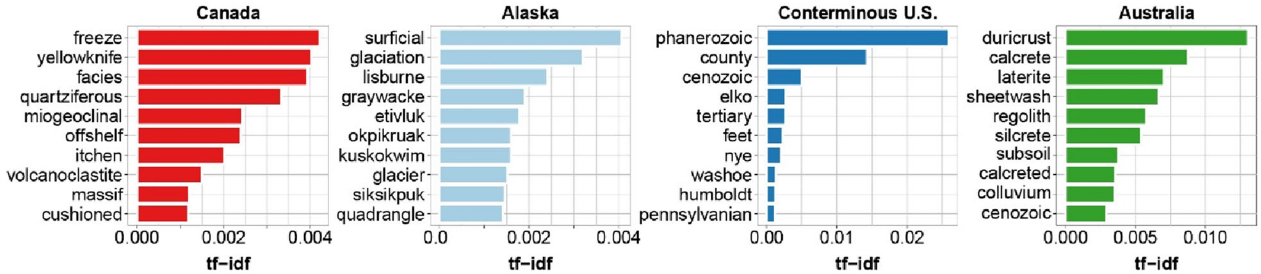

Figure 6. Term frequency-inverse document frequency (TF-IDF) scores for Canada, Alaska, conterminous U.S., and Australia. Words with high TF-IDF scores are the most characteristic for that particular database (i.e., word occurs frequently in one database but is rare overall).

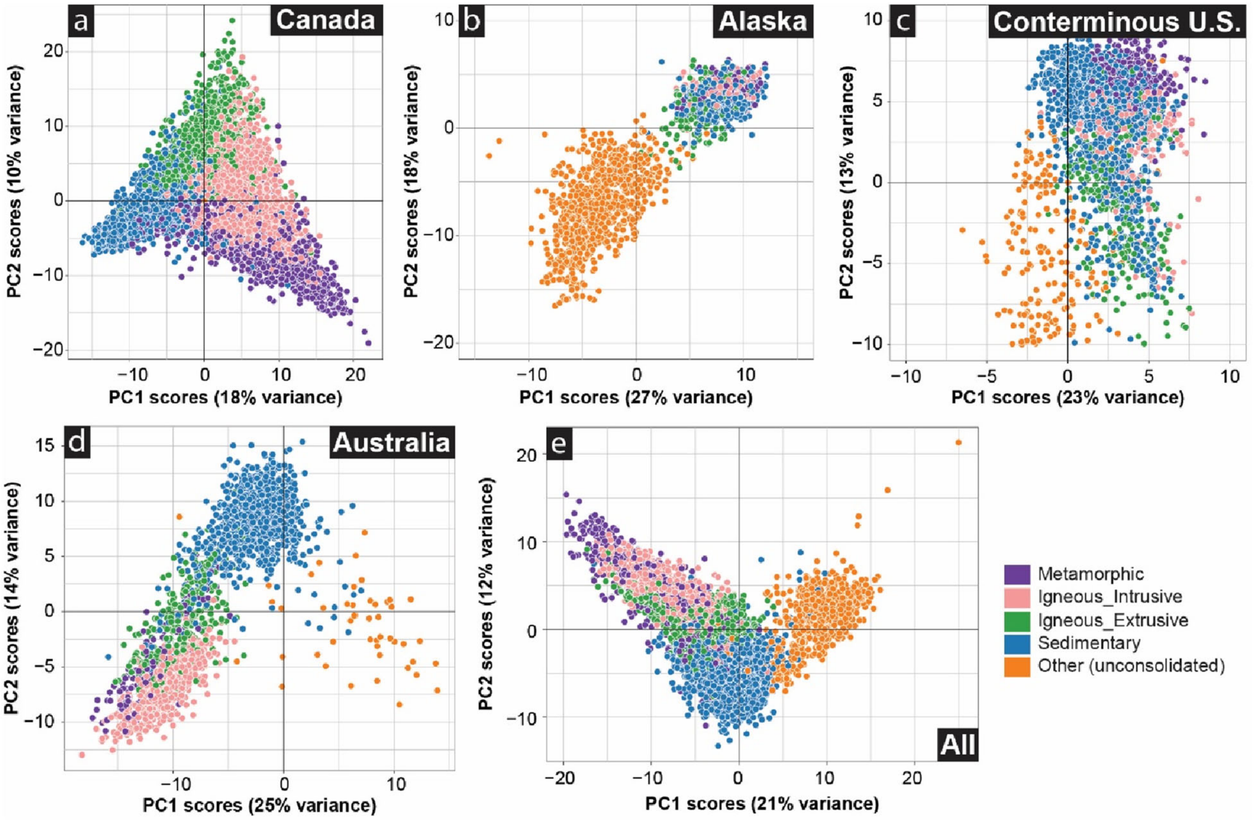

Figure 7. Principal component analysis (PCA) biplots for geological map databases from Canada (**a**), Alaska (**b**), the conterminous U.S. (**c**), and Australia (**d**). The PCA was repeated for all four geological map databases combined (**e**). Each data point corresponds to a map polygon. Colors are based on the generalized rock types described in Lawley et al. (2022b). The clustering of rock types along the first and second principal components (i.e., PC1 and PC2) suggests that linear combinations of word vectors capture meaningful differences between map unit descriptions. Random sampling was used to limit the amount of data for visualization purposes (n = 10,000).

The second step in text processing (Fig. 1; Step 2.2) removes ''stop words'', i.e., common but uninformative tokens from the corpora. Our study used the pre-made list of English stop words included in the ''tidytext'' package (e.g., ''they'', ''this'', ''that'', ''what''; n = 1149; Silge & Robinson, 2016). Country, state, provincial, and territorial names were also removed as special cases because these tokens were added during concatenation of some text fields. Igneous and sedimentary formations named after states are excluded during this step since these words have multiple meanings. Similarly, tokens that were

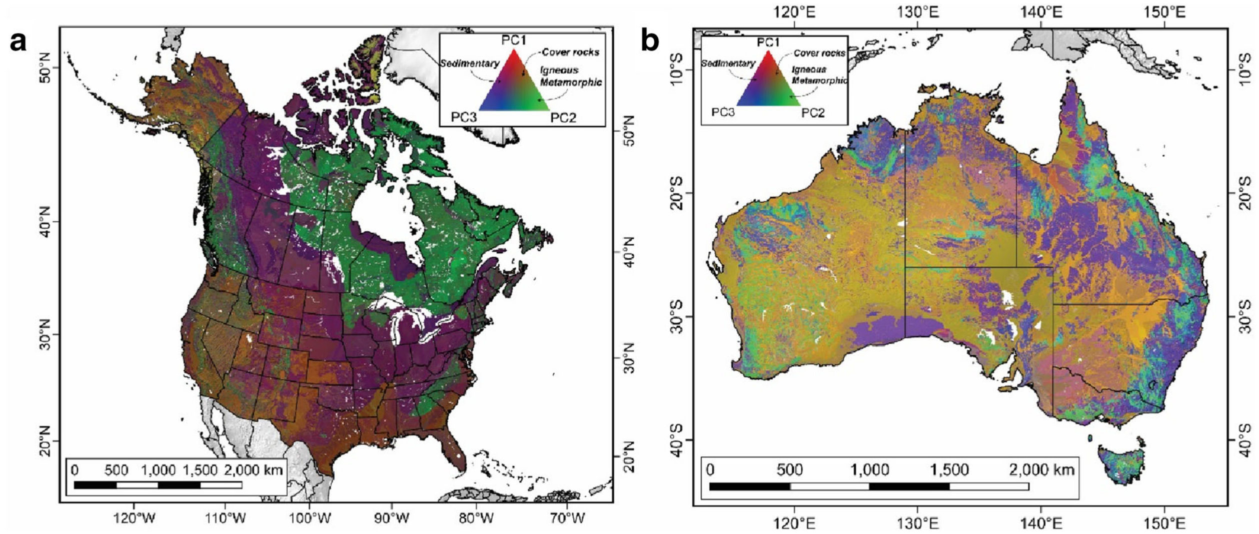

Figure 8. (**a**) Average word vectors transformed by Principal Component Analysis (PCA) for Canada, the conterminous U.S., and Alaska. (**b**) Average word vectors transformed by Principal Component Analysis (PCA) for Australia. Ternary colors are based on the three most important linear combinations of word vectors (PC1 = red; PC2 = green; and PC3 = blue). The general agreement between NLP maps and the generalized geology map suggests that rock type is the dominant source of data variance. The NLP map highlight polygons with multiple rock types, including partly covered bedrock in Alaska (**a**), the conterminous U.S. (**a**), and Australia (**b**).

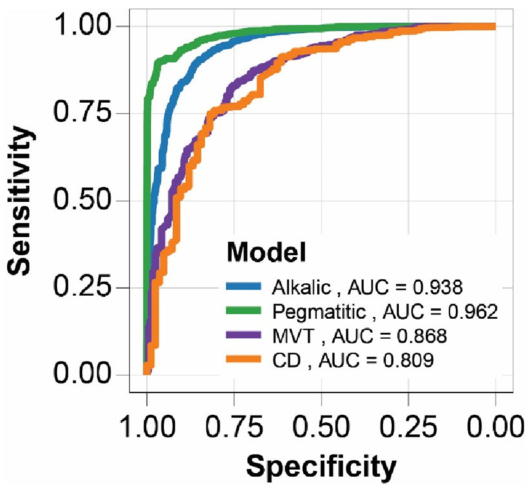

Figure 9. Receiver Operating Characteristics (ROC) plots showing classification results for pegmatitic, alkalic, Mississippi Valley-type (MVT) Zn-Pb deposits, and clastic-dominated (CD) Zn-Pb deposits. The area under the curve (AUC) is a measure of classification performance. Poor models yield an AUC of 0.5; whereas a perfect classifier will yield an AUC of 1. Overall, the results suggest that simple Naive Bayes classifiers and geoscience word embeddings can be used to predict map polygons that are likely to contain rare rock types or used as input into prospectivity modelling.

used to impute missing values were removed from further analysis (i.e., ‘‘unknown’’) because the frequency of these stop words is artificially inflated and could impact the calculation of mean vectors.

Stemming, the third NLP step (Fig. 1; Step 2.3), removes suffixes and prefixes to reduce the vocabulary size of the corpora and to focus text analysis on root words. In a geoscience context, stemming is a useful tool for addressing all of the different terms that are used by geoscientists to describe the same concept. For example, siliceous, silicified, and silicic are all used to describe a silica-rich rock in different contexts. Standard stemming algorithms, such as the Porter stemmer, attempts to intelligently remove suffixes to a word stem using a series of pre-defined rules (Bouchet-Valat, 2020). However, the output of the standard stemming algorithms are not guaranteed to return a real word (e.g., ‘‘silica’’ becomes ‘‘silic’’). Testing completed as part of the current study identified that a large number of important geoscientific concepts would have been reduced to meaningless word stems using these automated stemming methods. The geoscience GloVe model vocabulary contains mostly English words and thus meaningless word stems would have been excluded from further analysis. To include as many important geoscientific words as possible, some of the most common prefixes (‘‘macro’’, ‘‘micro’’, ‘‘mega’’, and ‘‘meta’’) were replaced manually for each token using regular expression (‘‘regex’’) operations rather than more advanced stemming algorithms. Similarly, plural terms were identified and replaced using the Harman (1991) method, as described in Hvitfeldt

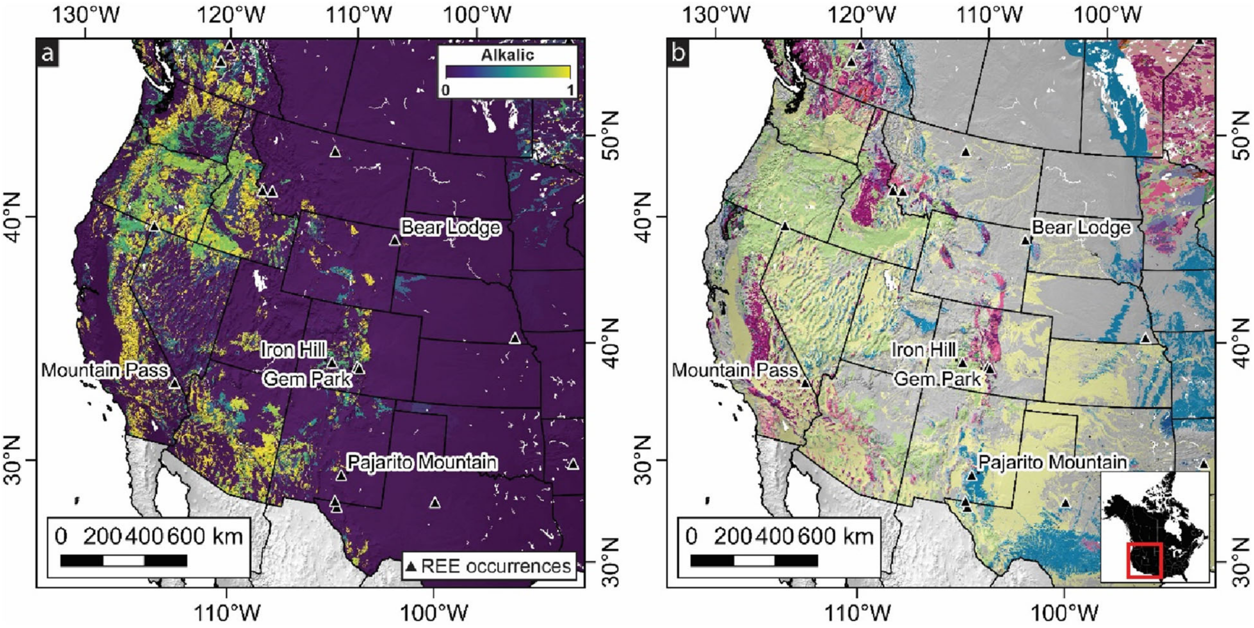

Figure 10. (**a**) Classification model results for ''alkalic'' rocks in the southern conterminous U.S. Rare earth element deposits and mineral occurrences are shown for reference and were not used for model training. (**b**) Generalized geological map of the southern conterminous U.S. Colors are the same as the legend for Fig. 2.

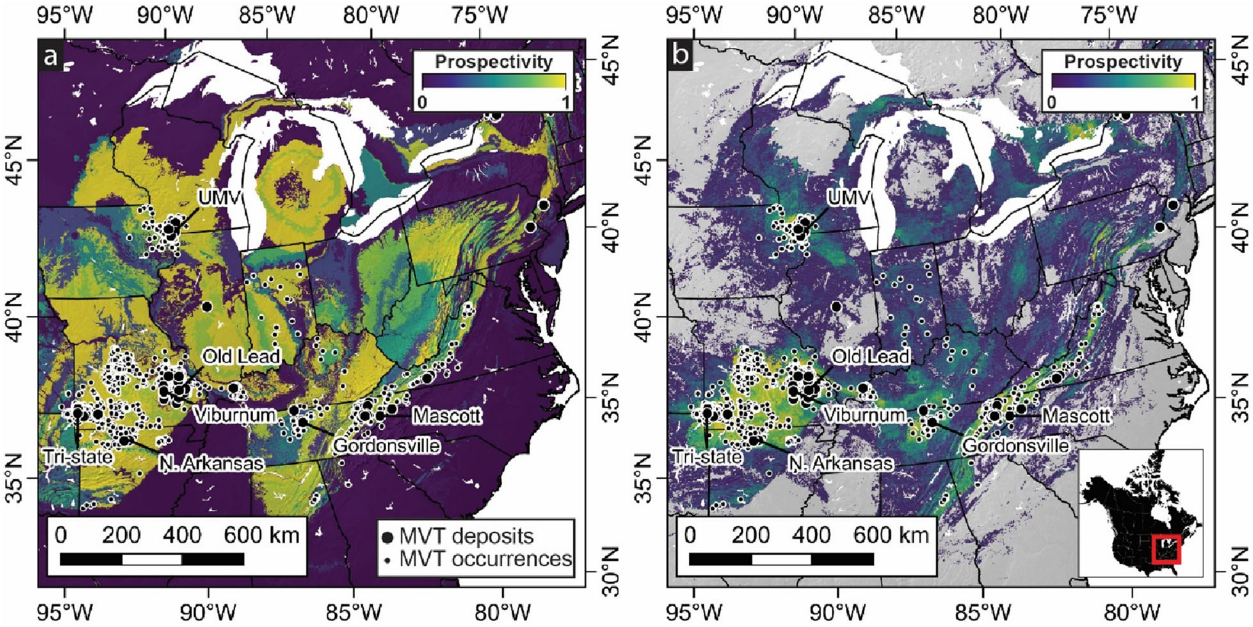

Figure 11. (**a**) Prospectivity models based on NLP in the U.S. and Canada that have high potential for Mississippi Valley-type (MVT) Zn-Pb deposits. The MVT deposits and mineral occurrences used for training are shown for reference (UMV = Upper Mississippi Valley district). (**b**) Previously published prospectivity model for MVT deposits based on geology and geophysics (Lawley et al., 2022b). Prospectivity values are filtered to the top 10% for visualization purposes.

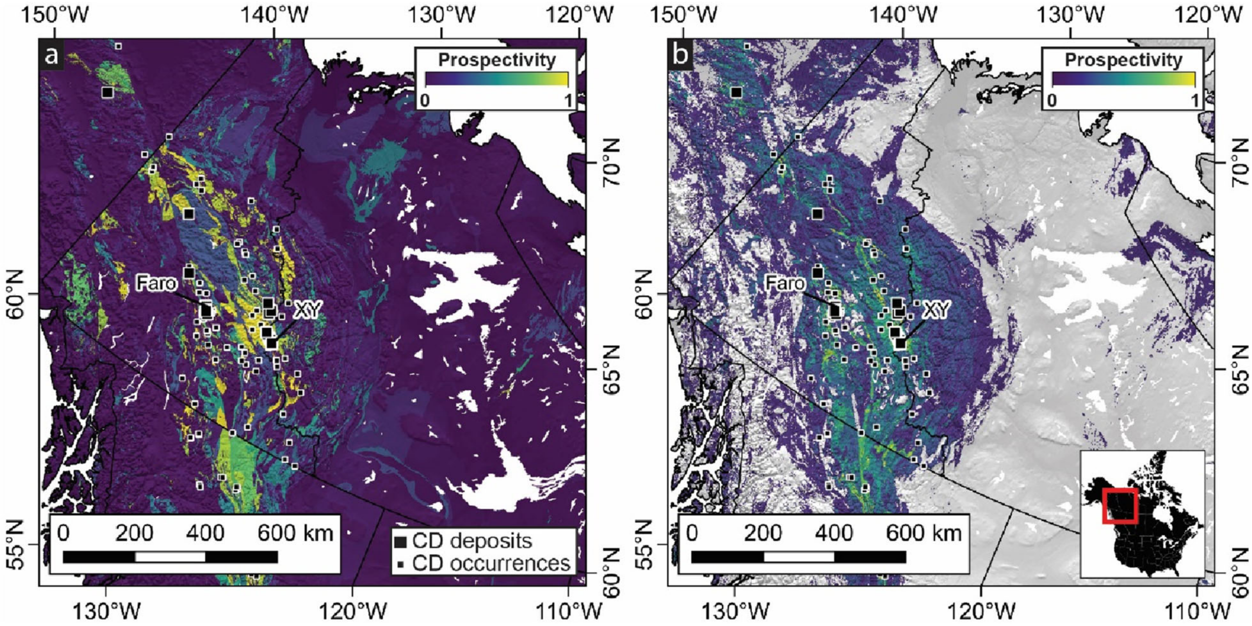

Figure. 12. (a) Prospectivity models based on NLP in Alaska and Canada that have high potential for clastic-dominated (CD) Zn-Pb deposits. The CD deposits and mineral occurrences used for training are shown for reference. (b) Previously published prospectivity model for CD deposits based on geology and geophysics (Lawley et al., 2022b). Prospectivity values are filtered to the top 10% for visualization purposes.

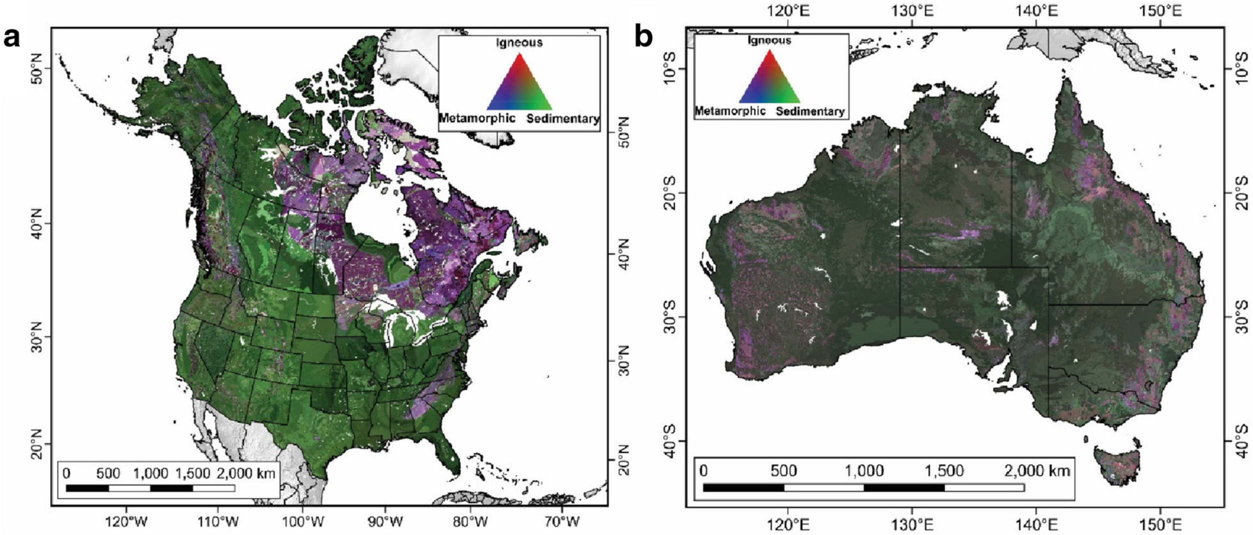

Figure 13. (**a**) Natural language processing results for Canada, the conterminous U.S., and Alaska. (**b**) Natural language processing results for Australia. Ternary colors are based on the cosine similarity between word vectors and igneous (red), sedimentary (green), and metamorphic (blue) geoscientific concepts. Rock descriptions that are more similar to igneous and metamorphic concepts highlight the Canadian shield (**a**); whereas most of the rock descriptions from the conterminous U.S., Alaska, and Australia share more similarities with sedimentary rocks.

and Silge (2021). This simple heuristic: (1) removes ''s'' as a last letter for each word; (2) replaces ''ies'' with ''y''; and (3) replaces ''es'' with ''e''. Although these manual stemming methods are relatively simple, the resulting words stems are more similar to the geoscience GloVe model vocabulary (Lawley et al., 2022a).

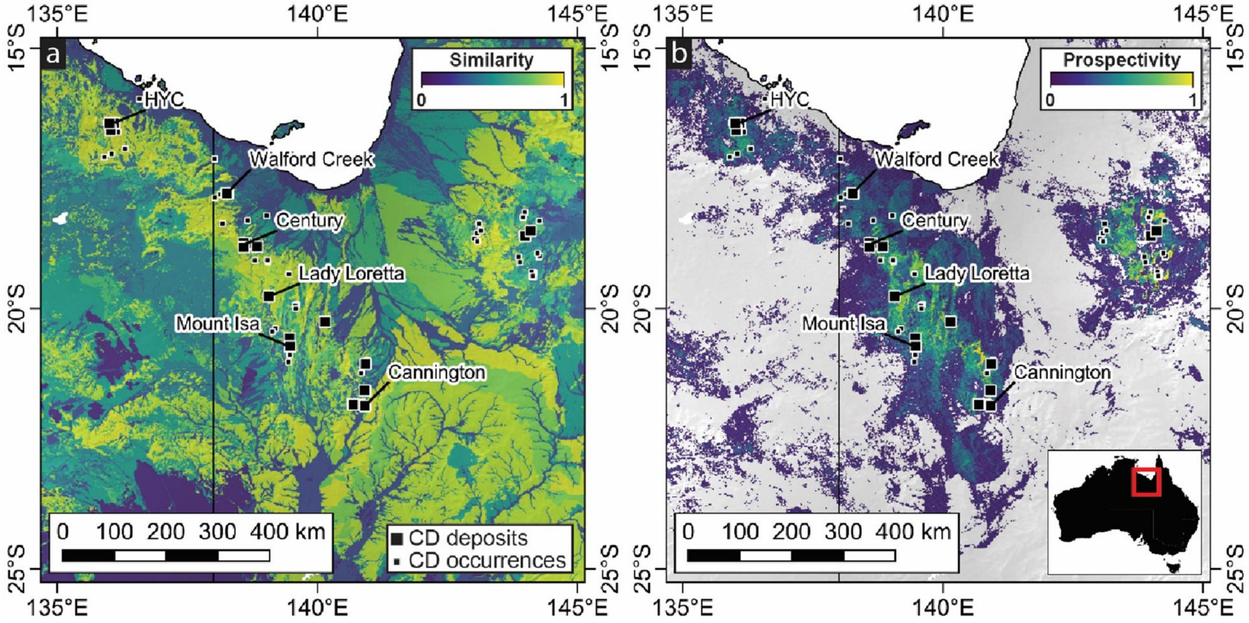

Figure 14. (**a**) Semantic search results for Mount Isa Zn-Pb deposit analogues in Australia. This form of nearest neighbors analysis and semantic search can be used to estimate mineral potential for application with limited training data. (**b**) Previously published prospectivity model for CD deposits based on geology and geophysics (Lawley et al., 2022b). Prospectivity values are filtered to the top 10% for visualization purposes. The known CD deposits and mineral occurrences are shown for reference.

Data from Canada also required an extra translation step prior to analysis (Fig. 1; Step 2.4). Machine translation for the Québec geological map database used the application programming interface to DeepL (https://www.deepl.com). All translated tokens were then manually checked for errors since technical terms remain a significant challenge for machine translation. Overall, the final, processed, and combined text dataset comprises 36,222,640 tokens from the conterminous U.S. (15,038,657 tokens), Canada (8,089,963 tokens), Alaska (7,965,923 tokens), and Australia (5,128,097 tokens). Hereafter "tokens" are referred to as "words" for simplicity.

GEOSCIENCE LANGUAGE MODEL

The original GloVe model was pre-trained on a large matrix of co-occurring words (i.e., six billion tokens) taken from the Wikipedia 2014 and Gigawords datasets (Parker et al., 2011). This pre-trained language model was then re-trained 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, British Columbia), and open-source peer-reviewed publications as described in Lawley et al. (2022a). Both GloVe models are based on the assumption that words occurring together are more closely related (Pennington et al., 2014). Countries and their capital cities represent a famous example of this word proximity relationship (Mikolov et al., 2013a, 2013b), although the same relationship applies to geoscientific text (e.g., Paleozoic and Cambrian; igneous and granite; biotite and schist; fluvial and sandstone; Padarian & Fuentes, 2019; Fuentes et al., 2020; Lawley et al., 2022a). Each of the 400 k words in the geoscience GloVe model is associated with a 300-dimensional numerical vector (Fig. 1; Step 3). Individual words within the processed text data were then joined with their corresponding vector before calculating an average word embedding for each polygon (Fig. 1; Step 3). Average word embeddings are not impacted by the length of text data for each map polygon, allowing short and long rock descriptions to be considered together for the purpose of this study (Fig. 4; Mitchell & Lapata, 2010; Wieting et al., 2016; Adi et al., 2017; Shen et al., 2018). The output of this text processing pipeline is a data table containing map polygons as rows and 300-dimensional vectors as separate columns (Fig. 1; Step 3). Data in this form can then be used as input for the classification tasks in modeling discussed below (Fig. 1; Step 4). The standard deviation of cosine similarities (discussed below) was also calculated to measure the variability of word vectors contributing to each map polygon (Fig. 3).

VALIDATION MODELS

Word Counts

Word counts and the frequencies of co-occurring words (i.e., ''word pairs'') were calculated in R using the ''tidyverse'' (Wickham et al., 2019), ''tidytext'', and ''widyr'' packages (Silge & Robinson, 2016). These descriptive statistics provide simple evaluation metrics for documenting the similarities and differences of the four geological map databases. The most frequently used words occur in more map polygons and are thus expected to have a significant impact on the performance of down-stream tasks. Rare words may also be important if they are associated with numerical vectors that are distinct from other text data within the same polygon (discussed below).

Network Analysis

Network analysis was completed using the ''tidytext'' (Silge & Robinson, 2016), ''igraph'' (Csardi & Nepusz, 2006), and ''ggraph'' (Pedersen, 2021) packages in R (R Core Team, 2023). The networks are based on the co-occurrence of words for each map polygon, with each node representing a word and each node-edge representing a connection between words (Fig. 5). The shape of the network is controlled by the layout function, which, in this case, used the Fruchterman-Reingold algorithm (Fruchterman & Reingold, 1991). Other graph algorithms and layouts have the potential to highlight different aspects of the text data (Csardi & Nepusz, 2006). The layout function is thus a subjective choice, and, in this case, was selected to minimize the number of overlapping nodes edges for visualization purposes (Fruchterman & Reingold, 1991). Network analysis based on this method can highlight natural groupings of nodes that define geoscientific concepts (Fig. 5). The connections between node clusters define the connectivity between those geoscientific concepts (Fig. 5). The same method and algorithm was used by Morrison et al. (2017) to define the connectivity of mineralogical systems. Ma (2022) provides a recent review of graph theory applications in geoscience. Network analysis is used with domain knowledge as a form of intrinsic evaluation of the NLP workflow (Figs. 1 and 5).

Term Frequency–Inverse Document Frequency

The term frequency-inverse document frequency (TF-IDF) statistic was calculated using the ''tidytext'' package (Silge & Robinson, 2016). The statistic is based on the relative frequency of words for each of the four geological map databases (i.e., term frequency; TF) multiplied by a factor that decreases the weight of the most commonly used words (inverse document frequency; IDF; Hvitfeldt & Silge, 2021):

$$TF - IDF = TF(t, d) \times IDF(t)$$

where t is the number of times a term (t) occurs in document (d).

Frequently used words that appear in most geological map databases yield low TF-IDF scores; whereas words that appear frequently in some geological map databases but not others yield high TF-IDF scores (Fig. 6). The TF-IDF scores can be used to evaluate the importance, or relevancy, of words within the compilation of geological map databases. Calculated TF-IDF scores were also used as an intrinsic evaluation method for the text processing pipeline (e.g., translation errors, stemming issues, missed stop words).

UNSUPERVISED MACHINE LEARNING METHODS

Principal Component Analysis (PCA) and scaling (Fig. 1; Step 4) were calculated using the ''prcomp'' function in R (R Core Team, 2023) to extract the most important linear combinations of word vectors for predictive modeling and visualization purposes. Combining PCA results, or other forms of unsupervised machine learning, with domain knowledge is essential for attaching geoscientific significance to the largest sources of data variance and was used here as an intrinsic evaluation method (Lawley et al., 2022a). First, principal com-

Table 1. Predictive model results

Results	Model			
	Alkalic[a]	Pegmatitic[a]	Mississippi Valley-type (Zn-Pb) deposits[b]	Clastic-dominated (Zn-Pb) deposits[b]
All positives (n)	54,078	26,133	1180	418
All negatives (n)	1,064,498	1,092,443	1,117,396	1,118,158
Training positives (n)	43,263	20,907	944	335
Training negatives (n)	851,599	873,955	893,917	894,527
Test positives (n)	10,815	5226	236	83
Test negatives (n)	212,899	218,488	223,479	223,631
Training AUC	0.938	0.966	0.838	0.841
Test AUC	0.938	0.962	0.868	0.809

[a]True positives are based on previously published vocabularies (Lawley et al., 2022a, 2022b)
[b]True positives include known locations of deposits and mineral occurrences (Lawley et al., 2022a, 2022b)

ponents were calculated separately for each of the four geological map databases to evaluate the text processing pipeline (e.g., detect outliers; Fig. 7a, b, c, and d). Second, PCA scores were calculated for text data from all four geological map databases combined (Fig. 7e). The combined PCA scores were then joined back to the manual rock classification and geometrical attributes of each map polygon for visualization purposes (Fig. 8).

SUPERVISED MACHINE LEARNING METHODS

Classification

All predictive modeling was completed using the ''h2o'' package in R (R Core Team, 2023), which is the interface to the H2O artificial intelligence platform (www.h2o.org). For our investigative study, classification models were trained for two different tasks: (1) prediction of rare rock types (Figs. 9 and 10); and (2) assessing the mineral potential for basin-hosted Zn-Pb deposits (Figs. 11 and 12). The first classification task sought to predict areas with ''pegmatitic'' and ''alkalic'' intrusions based on the available rock descriptions. The training data for this application were defined using the presence or absence of rock types using custom search terms, as described in Lawley et al. (2022b). For ''pegmatitic'' rocks the original list of search terms prior to French to English translation included: ''megacryst'', ''pegmatite'', ''pegmatitic'', and ''pegmatitique''. The original list of search terms for ''alkalic'' rocks was more comprehensive and includes: ''alkali'', ''alkalic'', ''basanite'', ''essexite'', ''foid'', ''hawaiite'', ''larvikite'', ''latite'', ''monzonite'', ''neph'', ''nepheline'', ''néphéline'', ''nordmarkite'', ''phonolite'', ''pulaskite'', ''quartz-monzonite'', ''quartz-syenite'', ''shonkinite'', ''syenite'', ''syénite'', ''syenitic'', ''syénitique'', ''syenodiorite'', ''syenodioritic'', ''tinguaite'', ''trachy'', and ''trachyte''. In some cases, geochemistry may have been used to identify ''alkalic'' rocks, although this information is not contained within the geological map databases. The prediction of ''pegmatitic'' and ''alkalic'' rocks is relatively simple because it tests the ability of word embeddings to capture rock information that is known to exist in the training data (Table 1). Rock descriptions that do not contain these words were treated as negative for the purposes of predictive modeling (Table 1).

The second classification task was to predict the mineral potential for MVT and CD Zn-Pb deposits based on the characteristics of their host rocks. Carbonate (e.g., limestone, dolostone) and siliciclastic (e.g., carbonaceous shale and its metamorphosed equivalents) rocks represent the most favorable host rocks for MVT and CD deposits, respectively. Both deposit types are also more prospective during certain geological periods, lithostratigraphic information that also exists within the unstructured text data (Leach et al., 2001; Lyons et al., 2006; Huston et al., 2016, 2022). Training data for this application are based on the locations of known mineral occurrences and deposits for all three countries (Lawley et al., 2022b). The locations of these deposits were not explicitly included in the original text data and must be inferred from the favorable rock descriptions. As a result, predicting mineral potential represents a significantly harder machine learning task.

Training and test data for both down-stream tasks were generated using an 80:20 split, making sure to preserve the class distribution for each set. The training data were then split again into five cross-validation sets using stratified sampling (i.e., evenly distributing deposits and mineral occurrences between sets). Up- and down-sampling were used on the training data to address the relatively few deposits and mineral occurrences that are available for training using the "caret" package (Kuhn, 2008). For the up-sampled training data, this required replacement so that positive and negative classes had the same frequency. Models were trained using the first 50 principal components, which represent more than 90% of the data variance. For comparison, the first ten principal components represent 69% of the data variance. All three training sets (i.e., original class distribution, up-sampled, and down-sampled) were modeled separately using the Naive Bayes algorithm in the "h2o" package. Simple Naive Bayes classifiers were selected as the preferred modeling method because: (1) the method did not require parameter tuning; (2) the method scaled well for larger numbers of predictors; and (3) the method was less susceptible to overfitting, which can negatively impact the generalization of model results to unknown areas. The area under the curve (AUC) for the receiver operating characteristic (ROC) plot was used to evaluate model performance in all cases (Fig. 9). Models with higher AUC suggest better classification performance. A perfect classifier would yield an $AUC = 1$; whereas an $AUC > 0.8$ and $AUC > 0.9$ can be interpreted as good and excellent classification performance, respectively, in the context of prospectivity modeling (Nykänen et al., 2015; Airola et al., 2019; Zuo & Wang, 2020; Chudasama et al., 2022a, b).

Nearest Neighbors

Words with similar meaning correspond to closely associated numerical vectors (Mikolov et al., 2013a, 2013b; Pennington et al., 2014). The association between two numerical vectors can be evaluated using cosine similarity as implemented in the "text2vec" package (Selivanov & Wang, 2016):

$$cos(\theta) = \frac{A \cdot B}{||A||||B||}$$

where A and B are vectors.

Nearest neighbor analysis based on cosine similarity can be used to: (1) rank the most closely associated words to a test word as an intrinsic evaluation method (Lawley et al., 2022a); (2) measure the semantic variability between each word and the average word embedding for each map polygon (Fig. 3); (3) search for map polygons that most closely match a concept included within the geoscience GloVe model (Fig. 13); and (4) search for the closest matching map polygons based on its text data and average numerical vector (Fig. 14). The most closely matching word vectors will yield cosine similarities closer to one. Nearest neighbors analysis is a form of semantic search that has the potential to greatly improve knowledge discovery from geological map databases.

RESULTS AND INTERPRETATION

Descriptive Text Statistics

The most frequently used words and word pairs are presented in Tables 2 and 3, respectively. Rock type, geochronology, and stratigraphic terms tend to be the most commonly used words and word pairs for each of the four geological map databases. The frequency of these words is due, at least in part, to the structure of the underlying databases, as rock types and geological periods are stored as semi-structured text attributes prior to concatenating with the unstructured and long-form rock descriptions. As a result, virtually every map polygon is associated with one or more rock types and ages (Fig. 2). The underlying data model for each geological map database is also the most likely explanation for the same word appearing in multiple top word pairs (e.g., Phanerozoic, sedimentary, sandstone, unconsolidated, undifferentiated; Table 2). Geological databases and text data that are based on international data standards (i.e., NADM, GeoSciML, INSPIRE) within structured and semi-structured attributes are more likely to be impacted by this frequency effect. The analysis of top words and top word pairs is a form of intrinsic evaluation to check that any text errors (e.g., spelling mistakes, stop words, translation) are limited to rare words.

Network analysis provides some additional insight into text data and processing methods (Fig. 5). For example, "siltstone", "sandstone", and "con-

Table 2. Top word pairs for each geological map database

Canada			Conterminous U.S			Alaska			Australia		
Word 1	Word 2	Counts (n)	Word 1	Word 2	Counts (n)	Word 1	Word 2	Counts (n)	Word 1	Word 2	Counts (n)
Mafic	Volcanic	57,263	Phanerozoic	Sedimentary	170,343	Deposit	Quaternary	93,192	Sand	Regolith	68,980
Sedimentary	Grey	48,966	Phanerozoic	Formation	136,525	Deposit	Surficial	90,919	Sedimentary	Siliciclastic	61,707
Sedimentary	Shale	44,783	Phanerozoic	Cenozoic	134,611	Surficial	Quaternary	90,781	Sandstone	Sedimentary	60,178
Sedimentary	Limestone	44,420	Phanerozoic	Undifferentiated	129,671	Unconsolidated	Quaternary	89,351	Include	Regolith	59,871
Sedimentary	Sandstone	44,168	Sedimentary	Formation	119,456	Deposit	Unconsolidated	89,348	Gravel	Regolith	59,539
Minor	Sedimentary	43,638	Phanerozoic	Paleozoic	114,955	Unconsolidated	Surficial	89,311	Gravel	Sand	57,198
Sedimentary	Siltstone	43,446	Phanerozoic	Sandstone	105,296	Deposit	Sand	46,726	Calcrete	Regolith	56,232
Sedimentary	Formation	43,073	Sedimentary	Sandstone	104,085	Quaternary	Sand	46,261	Residual	Regolith	54,036
Gneiss	Metamorphic	35,270	Paleozoic	Sedimentary	102,444	Surficial	Sand	46,221	Sandstone	Siliciclastic	52,872
Conglomerate	Sedimentary	34,491	Sedimentary	Shale	96,283	Unconsolidated	Sand	45,037	Residual	Calcrete	51,995
Sedimentary	Fine	33,425	Phanerozoic	Shale	96,232	Deposit	Silt	42,169	Laterite	Regolith	51,946
Sandstone	Siltstone	33,034	Phanerozoic	Include	93,708	Quaternary	Silt	42,168	Plain	Regolith	51,931
Shale	Sandstone	32,086	Phanerozoic	Limestone	91,661	Surficial	Silt	42,121	Residual	Include	51,892
Volcanic	Sedimentary	31,842	Sedimentary	Limestone	91,013	Deposit	Gravel	41,978	Include	Calcrete	51,892
Shale	Siltstone	31,316	Sedimentary	Clastic	90,388	Quaternary	Gravel	41,565	Residual	Laterite	51,892
Sedimentary	Bedded	31,100	Phanerozoic	Clastic	88,137	Surficial	Gravel	41,464	Include	Laterite	51,892
Grey	Limestone	30,548	Undifferentiated	Cenozoic	86,054	Unconsolidated	Silt	40,968	Calcrete	Laterite	51,892
Sedimentary	Grained	29,700	Phanerozoic	Unconsolidated	80,529	Unconsolidated	Gravel	40,079	Sand	Plain	51,027
Granitoid	Intrusive	28,998	Undifferentiated	Unconsolidated	79,672	Sand	Silt	36,986	Gravel	Plain	50,794
Grey	Shale	28,849	Paleozoic	Formation	77,918	Sand	Gravel	35,867	Sandstone	Siltstone	48,930

Table 3. Top word counts for each geological map database

Canada		Conterminous U.S		Alaska		Australia	
Word	Counts (n)	Word	Counts (n)	Word	Counts (n)	Word	Counts (n)
Sedimentary	273,853	Phanerozoic	562,882	Deposit	251,789	Sand	125,854
Volcanic	211,603	Formation	336,503	Rock	146,737	Sandstone	106,288
Limestone	140,023	Limestone	320,275	Surficial	112,465	Regolith	102,734
Sandstone	125,788	Shale	277,116	Unconsolidated	98,345	Sedimentary	91,279
Grey	125,517	Cenozoic	265,313	Quaternary	97,546	Ferruginous	88,285
Formation	119,770	Sandstone	256,059	Gray	92,878	Minor	82,727
Shale	118,308	Paleozoic	221,601	Unit	90,693	Deposit	76,792
Gabbro	98,171	Sedimentary	210,588	Formation	82,450	Igneous	76,100
Mafic	95,792	Rock	202,363	Sand	71,958	Intrusive	75,103
Siltstone	95,499	Gray	180,312	Locally	69,245	Felsic	69,976
Intrusive	83,032	Unit	170,476	Include	67,770	Quartz	69,532
Biotite	74,680	County	155,074	Sandstone	65,510	Plain	66,447
Gneiss	74,204	Quaternary	154,764	Volcanic	65,012	Include	64,914
Conglomerate	73,178	Include	143,126	Silt	64,851	Calcrete	64,722
Minor	70,940	Undifferentiated	141,472	Lake	64,690	Rock	64,246
Basalt	70,321	Tertiary	141,248	Chert	61,880	Siliciclastic	63,037
Fine	65,618	Sand	134,178	Limestone	57,729	Colluvium	62,653
Locally	64,177	Feet	121,552	Stream	56,268	Siltstone	61,852
Metamorphic	61,446	Cretaceous	117,840	Shale	54,662	Gravel	60,461
Bedded	61,202	Siltstone	115,284	Quartz	54,435	Mafic	54,392

glomerate'' plot close together and define a natural grouping on networks for each of the four geological databases (Fig. 5). The close association among these words is expected given that these rock types represent a continuum of increasing sediment grain sizes. Similarly, the close proximity between ''fine'' and ''volcanic'' in the Canada and Alaska datasets correctly identifies semantic relationships between this rock type and its texture. The frequency of words describing unconsolidated sediments further suggests that a large proportion of text included within geological map databases is used to describe cover rocks. As expected, this word frequency effect is most obvious for regions with extensive sedimentary cover, such as Australia and Alaska (Table 3). Network analysis is a form of intrinsic evaluation to check for semantic relationships based on word co-occurrences prior to joining with the geoscience GloVe model (Fig. 1).

Evaluating the differences between geological map databases is best handled by TF-IDF scores (Fig. 6). The calculated TF-IDF scores are a measure of ''relevance'' that are widely used by search engines and for keyword generation (Silge & Robinson, 2016). The most relevant words from the Australia geological map databases are exclusively associated with descriptions of deeply weathered rocks and soils (e.g., ''duricrust'', ''calcrete'', and ''laterite''). These weathering products are mostly absent from North America because of its cooler climate and different geological history. In contrast, the highest TF-IDF scores for Canada, conterminous U.S., and Alaska correspond to regional formation names, places, and only a few geoscientific terms (Fig. 6). The analysis of words with high TF-IDF is a form of intrinsic evaluation to look for the most important vocabulary differences between map compilations.

Unsupervised Machine Learning Results

Principal Component Analysis (PCA) results are presented in Fig. 7. Points are color-coded to the generalized rock-type classification system described in Lawley et al. (2022b). Rock types were further collapsed to five classes for visualization purposes (i.e., metamorphic, igneous extrusive, igneous intrusive, sedimentary, and other). The PCA method does not require sample labelling prior to analysis. As a result, the natural groupings of generalized rock types along the first (PC1) and second (PC2) principal components demonstrate that lithology is the primary source of data variance for each of the geological map compilations (Fig. 7a, b, c, and d). The clustering of generalized rock types makes intuitive sense, but is important to demonstrate because it suggests that the PCA-transformed word vectors preserve the essential elements of the unstructured rock descriptions even after information loss. Because the correct lithological relationships are also observed on the combined PCA biplot (Fig. 7e), we suggest that any differences in geoscientific terminology between Canada, the U.S., and Australia are less important than the semantic differences between the generalized rock types.

Ternary maps of the PCA results are presented in Fig. 8 based on the combined PCA results. Each ternary color represents one of the three most important linear combinations of word vectors after PCA (i.e., PC1, PC2, and PC3). These maps highlight the large number of map polygons containing descriptions of multiple rock types. Map units with two or more rock types are common during regional mapping and are difficult to visualize on conventional geological maps (Fig. 2). Grouping similar rocks together using classification hierarchies is the usual solution to this problem (Fig. 2). However, here we demonstrate how the ternary NLP maps capture the basic elements of the generalized geological maps, such as the predominantly igneous and metamorphic rocks comprising the Canadian Shield (Fig. 8a), partly covered basement rocks in North America (Fig. 8a), complex mixtures of sedimentary and igneous rocks exposed along the Australian shoreline (Fig. 8b) and unconsolidated sediments covering most of the Australia interior (Fig. 8b). The PCA-transformed word embeddings are used to train the classification models below.

Supervised Machine Learning Results

Predictive modeling results are reported in Table 1 and presented in Figs. 9, 10, 11, and 12. Overall, classification models that predict whether map polygons are "pegmatitic" (AUC = 0.962) and "alkalic" (AUC = 0.938) rock types yield the best performance (Fig. 9). Up- and down-sampled versions of these models, which address the relatively minor class imbalance of the training sets, slightly improve the predictive performance for the "pegmatitic" (i.e., up-sampled AUC = 0.967 and down-sampled AUC = 0.966) and "alkalic" (i.e., up-sampled AUC = 0.940 and down-sampled AUC = 0.941) model test sets. The good agreement between test and training set performance for "pegmatitic" (train AUC = 0.966; up-sampled train AUC = 0.972; down-sampled train AUC = 0.970) and "alkalic" (train AUC = 0.938; up-sampled train AUC = 0.939; down-sampled train AUC = 0.939) models further suggests that over-training was not an issue for this down-stream task. The excellent performance for these predictive models was somewhat expected as the text data were already known to contain terms corresponding to "alkalic" and "pegmatitic" rocks. Nevertheless, this type of data-driven approach represents a significant improvement over the manual creation of custom vocabularies for applications where sufficient training data are available (Lawley et al., 2022b). Classification results can also be visualized to highlight map polygons that are more likely to contain pegmatitic and alkalic rocks that are prospective for critical raw materials using the available rock descriptions (Fig. 10). Model predictions are based on the available text descriptions, rather than actual presence or absence of "alkalic" and "pegmatitic" rocks. Map polygons with missing descriptions for these rock types will be wrongly assigned low probabilities. Similarly, it is possible that these rock types may be included in rock descriptions even when they are known to be absent. The model performance described above does not test for either of those scenarios (i.e., missing and incorrect rock descriptions). As a result, text-based models should be combined with external sources of training data or combined with other datasets to improve classification performance (discussed below).

Test sets for the MVT and CD models yield an AUC of 0.868 and 0.809, respectively (Fig. 9). Up- and down-sampled versions of these models, which address the extreme class imbalance for these training sets, slightly improve the predictive performance for the MVT (i.e., up-sampled AUC = 0.874 and down-sampled AUC = 0.870) and CD (i.e., up-sampled AUC = 0.802 and down-sampled AUC = 0.819) model test sets. The similar AUC for the test and training sets for the MVT (AUC = 0.838; up-sampled AUC = 0.843; down-sampled AUC = 0.859) and CD (AUC = 0.841; up-sampled AUC = 0.842; down-sampled AUC = 0.843) models further suggests that model overfitting was relatively minor. The relatively few true positives examples, and the absence of true negatives examples, available for training represents a major challenge for this classification task. The lower classification performance for MVT and CD models is also expected because deposits and mineral occurrences were not explicitly mentioned in the text data used to train the predictive models. Instead, the mineral potential for the MVT and CD models is completely based on the favorability of the host rock descriptions (i.e., rock type, geological periods, and other characteristics). Nevertheless, the AUC for the text-based MVT and CD models are promising because they are comparable to previously published prospectivity modeling results for other mineral systems (e.g., AUC = 0.69–88; Nykänen et al., 2015; Chudasama et al., 2022a, 2022b) and represent an improvement over the grid-based and geophysics-only models for the same deposit and mineral occurrence training data (MVT = 0.640 and CD = 0.826; Lawley et al., 2022b). Examples of how these text-based models correctly predict the most favorable host rocks for MVT and CD deposits in the U.S. midcontinent and northern Canada are presented in Figs. 11 and 12.

Finally, nearest neighbor analysis is an alternative form of supervised machine learning that is based on the cosine similarities between two or more word vectors. Here we apply cosine similarities to search for the closest analogues of the giant Mount Isa Zn-Pb deposit based on the description of its host rocks (Fig. 14a). The available descriptions for Mount Isa include age (i.e., Stratherian) and lithological information (e.g., siltstone, shale, dolomite, sandstone, conglomerate; Online Supplementary Table), which was combined into a mean vector before predicting the closest neighbors across Canada, the U.S., and Australia. The cosine similarity results show map polygons with the closest matching description using all of the available text data, revealing favorable host rocks for this deposit type that are consistent with previously published prospectivity models (Fig. 14b; Lawley et al., 2022b).

DISCUSSION

Natural Language Processing for Knowledge Discovery

National geological maps represent compilations of smaller surveys collected over many decades and at great financial expense (Ramdeen, 2015). Geological maps also provide a high value to society, with an estimated 4:1 to 100:1 benefit-to-cost ratios (Berg et al., 2019). The observations and technical knowledge encoded in these databases over many decades represents a massive human effort, with contributions from hundreds of geoscientists, and are probably the most important contribution of geological survey organizations to society (Howard et al., 2009; Lebel 2020; Culshaw et al., 2021). Most geological map databases, and geoscience more generally, contain an untapped wealth of information in the form of unstructured and semi-structured text data. Very few, if any, research studies have provided descriptions of the text data contained within these important sources of geological information or their application to prospectivity modeling (Mantovani et al., 2020).

Simple text statistics demonstrate remarkable similarity in the terminology and use of geoscientific language across all three countries, at least for the most used words (Fig. 5; Tables 2 and 3). This result is promising since the further adoption of GeoSciML and other data standards will likely accelerate the interoperability and accessibility of text data within geological map databases in the future (Sen & Duffy, 2005; Reitsma et al., 2009; Lombardo et al., 2018; Mantovani et al., 2020). Here, we demonstrate that the co-occurrence of these commonly used words embeds semantic information that can be extracted and used to visualize the connections between disparate geoscientific concepts (Fig. 5). For example, network analysis of the most used words correctly differentiates sedimentary rocks (e.g., conglomerate, sandstone, siltstone) from nodes representing unconsolidated sediments (e.g., gravel, sand, and silt). This result is important and somewhat surprising because it correctly identifies meaningful semantic differences between otherwise very similar words (e.g., sand versus sandstone). Training language models on domain-specific text is particularly important for this type of knowledge discovery in a geoscience context (Lawley et al., 2022a). Other examples of semantic information encoded by the simple co-occurrence of words include: (1) geochronological relationships (e.g., Cenozoic and Quaternary versus Paleozoic and Carboniferous); (2) stratigraphic terms (e.g., sedimentary and formation); (3) igneous rock types and textures (e.g., fine and volcanic; Fig. 5); and (4) associations between non-geoscientific terms (e.g., lake, ocean, water, stream). The critical assessment of word nodes in a geoscientific context is an important validation method of the text processing method (Fig. 1).

Because the geoscience GloVe model was trained on a much larger corpora, even deeper levels of semantic information are encoded in the word vectors (Lawley et al., 2022a). Analogy tests are a classic method for exploring text semantics and can be calculated by simple vector arithmetic (e.g., addition and subtraction), such as the famous examples of countries and their capital cities (Ottawa – Canada + Canberra = Australia; Mikolov et al., 2013a, 2013b). The geoscience equivalent to this form of analogy test could include ''igneous is to granite as metamorphic is to gneiss'' (i.e., igneous – granite + gneiss = metamorphic), which tests for the conceptual association between rock types and their process of formation. Whether these types of analogies can be correctly answered by a language model, depends on the proximity of word vectors in the embedding space (Lawley et al., 2022a).

Nearest neighbor analysis based on cosine similarities and PCA represent two methods for visualizing the proximity relationships between word vectors directly. The results presented in Figs. 8 and 13 represent examples of semantic geological maps (Brodaric & Gahegan, 2001; Lombardo et al., 2018; Mantovani et al., 2020). However, unlike previously proposed semantic ontologies that need to be manually created and updated (i.e., top-down), the clustering of word vectors and relationships between rock types are emergent from the input text data (i.e., bottom-up). The agreement between these NLP maps and the generalized geology for Canada, the U.S., and Australia is remarkable, as it demonstrates, for the first time, how the semantic relationships between rock types can be extracted from the unstructured text data within geological map databases (Fig. 8). Moreover, unlike traditional geological maps that reflect a preferred interpretation of the dominant lithology (Fig. 2), the new NLP maps reflect all the available text data for each map polygon (Fig. 8). Map polygons containing words with complex semantics, measured here as the standard deviation of cosine similarities between each word and the mean vector of each polygon (Fig. 3), are interpreted to have higher uncertainty. Quantifying this type of geological map uncertainty has previously represented a major knowledge gap for geological survey organizations (Brodaric et al., 2004). In theory, these NLP maps can be easily updated with new geological surveys or combined with other sources of text data (e.g., publications, drill data, field notes) to improve on these results. Lithostratigraphic and lithodemic databases, such as WEBLEX in Canada and the Australian Stratigraphic Units Database (ASUD), represent important external sources of text data for discovering semantic relationships between map polygons (Holden et al., 2019; Enkhsaikhan et al., 2021a, 2021b). With appropriate training data, the NLP maps suggest that geoscience language models could readily be expanded to generate predictive bedrock geology maps across all three countries using unsupervised (e.g., clustering) and/or supervised methods (e.g., classification).

However, it is also clear from the cosine similarities that most rock descriptions represent mixtures of geoscientific concepts, with relatively few map polygons corresponding to pure ''igneous'', ''metamorphic'', or ''sedimentary'' end-members (Fig. 13). Map descriptions containing mixtures of unconsolidated sediments are particularly difficult to interpret, with some plotting as outliers on PCA biplots (Fig. 7) and/or corresponding to map polygons with large standard deviations. (Fig. 3). In Australia, non-geoscientific words or words with multiple meanings outside of a geoscientific context (e.g., ''pipeclay'', ''dunk'', ''astrea'') appear to be the largest source of uncertainty for map polygons containing unconsolidated sediments (Fig. 3). The inclusion of words with complex semantics are the most likely explanation for unconsolidated sediments in Australia that plot as muted ternary colors because the calculated word vectors for these map polygons are dissimilar to the concepts of ''igneous'', ''metamorphic'', and ''sedimentary'' (Fig. 13). In Canada and the U.S., TF-IDF scores (e.g., ''freeze'', ''yellowknife'', ''glaciation'', ''county'', and formation names; Fig. 6) and tokens that yield dissimilar cosine similarities (e.g., ''boxer'', ''butcher'', ''baptism'') point to other words that likely contribute to the uncertainty of the NLP results (Fig. 3). Most of these rare words have multiple meanings depending on their context.

Unfortunately, despite re-training on geoscientific text, simple language models like GloVe are not able to unravel these multiple meanings using the sentences or paragraphs before and after each word. Instead, to address polysemy, words with multiple meanings could be added to the list of English stop words to exclude non-geoscientific words from further analysis. Continued progress on automated methods for identifying domain-specific stop words have the potential to greatly improve this type of analysis in the future (Ayral & Yavuz, 2011; Alshanik et al., 2020). The list of stop words should also be expanded to include words that do not appear in English dictionaries to limit the impact of any spelling errors and/or acronyms remaining in the language model (Lawley et al., 2022a). Word order and/or word importance could also be addressed using more advanced pooling methods (e.g., max pooling; hierarchical pooling, neural networks; Mitchell & Lapata, 2010; Shen et al., 2018) rather than the simple average vector, potentially improving the performance of down-stream tasks for map polygons with high uncertainty (Fig. 3). Alternatively, contextual language models may be better suited for dealing with complex text semantics for mapping applications (Vaswani et al., 2017; Devlin et al., 2019; Floridi & Chiriatti, 2020; Li et al., 2021; Ma et al., 2021). Large language models based on transformers represent the current state-of-the-art because they are able to generate multiple vectors for each word depending on its context (Devlin et al., 2019; Dale, 2021; Li et al., 2021; Ma et al., 2021; Lawley et al., 2022a). Unfortunately, text data in this study were concatenated from multiple fields, removing meaningful context in most cases (Fig. 1). Rock descriptions also tend to use short sentences with mostly scientific terms that are unlike the general internet text that more advanced language models are trained on (Devlin et al., 2019). Ideally, the training data for language models should closely match the text data being modeled.

Ternary NLP maps also highlight map boundary artifacts that reflect differences in the quality, level of detail (Figs. 8 and 13), and other aspects of rock descriptions across political boundaries. The southern Canada-U.S. border and Alaska-Yukon border provide clear examples of this effect with a marked difference in the proportion of unconsolidated sediments on both the generalized geology (Fig. 2a) and NLP ternary map (Fig. 13a). These NLP results reflect, in part, how unconsolidated sediments are treated as part of the bedrock mapping process. Unconsolidated glacial sediments are rarely described in the Canadian source datasets even where present. The ternary NLP maps smooth out the boundary artifacts in the U.S. and Canada to some extent (Fig. 13a) but do not completely address differences in the underlying text data. Moreover, map boundary artifacts also likely reflect differences in the quantity and quality of text data between jurisdictions, as demonstrated by the standard deviations of word vectors (Fig. 3). These text differences are not related to the processing methodology, since the same boundary effects are not observed for the seamless national Australia geological map (Figs. 8b and 13b).

Natural Language Processing for Prospectivity Modeling

The discussion above focused on natural clusters and proximity relationships between word vectors to extract knowledge from unstructured text data. However, word embeddings, coupled with supervised machine learning methods, can also be used for multiple down-stream tasks and to assess mineral potential more directly. For example, our Naive Bayes classification models (Figs. 9, 10, 11, and 12) can correctly predict the locations of ‘‘pegmatitic’’ and ‘‘alkalic’’ rocks from unstructured text data (Fig. 10), marking a significant improvement over approaches that required manual searching (Lawley et al., 2022b), or previously published research that used simple regex operations (Pollock et al., 2012). The excellent classification performance of these models (i.e., pegmatitic AUC = 0.962 and alkalic AUC = 0.938) is due to the relatively large number of training data available (Table 1). Moreover, the models were trained using words that were already known to exist in the text data and will be inaccurate if rocks are misidentified or are missing from the rock descriptions (Lawley et al., 2022b). Critically, the classification model results can be used to search for partly covered igneous intrusions based on all the available rock descriptions. In the future, we expect that text data from boreholes could also be included (Fuentes et al., 2020), which would likely improve classification performance for parts of Canada, the U.S. (Fig. 10b), and Australia that are buried by unconsolidated sediments.

Expanding this approach to include mineralized pegmatite locations from other databases (Burke & Khan, 2006; Woolley & Kjarsgaard, 2008; McCauley & Bradley, 2014), or to include other types of supporting data (e.g., geophysics, geochemistry, remote sensing; Eberle et al., 2012; Kesler et al., 2012), are some of the required elements for a more rigorous assessment of mineral potential. These other datasets are essential for separating ‘‘permissive’’ and ‘‘barren’’ rocks within a mineral system framework (Wyborn et al., 1994; London, 2005; McCuaig et al., 2010; Huston et al., 2016; González-Álvarez et al., 2021). Continued progress on the accurate prediction of these rare rock types, which often occur as small igneous intrusions that are poorly exposed, is important because they are major sources of lithium, rare earth elements, tantalum, niobium, beryllium and other critical raw materials (Kesler et al., 2012).

Supervised machine learning methods and text data can also be used more directly to identify prospective regions for sediment-hosted Zn-Pb mineral systems (Figs. 11 and 12). The good classification performance reported herein (MVT AUC = 0.868; CD AUC = 0.809) using text data and more simple machine learning methods (Figs. 11 and 12), provides yet another demonstration of the importance of geological information for assessing mineral potential in areas with exposed bedrock. The quality and availability of text used to train these simple models will likely grow as the methods for digitally acquiring field observations improve and as geological survey organizations continue to link geological map databases with external publications (Schetselaar, 1995; Brodaric et al., 2004; McCaffrey et al., 2005; Pavlis et al., 2010). This text-based approach contrasts with that of Lawley et al. (2022b), who relied on up to 26 different datasets to predict the potential for finding new MVT and CD deposits. Their data-driven prospectivity models produced excellent classification performance (AUC > 0.98) and are less impacted by surface sampling bias because they included seismic, magnetic, and gravity data that can penetrate unconsolidated sediments to map relevant geological features at depth. Evaluating mineral potential in covered areas (e.g., much of Alaska and Australia), or areas with sparse field observations (e.g., Canada's north), depends more heavily on geophysical datasets to map the preferred source rocks, pathways, and traps (e.g., carbonaceous and calcareous sedimentary rocks; Online Supplementary Table) of these mineral systems. Prospectivity models that include mappable proxies for the drivers, sources, pathways, and traps of mineral systems support mineral exploration by significantly reducing the search space (Figs. 11b and 12b). Ternary NLP maps can be used to identify where these partly covered rocks occur, which can be used to guide new geophysical surveys (Fig. 8).

Prospectivity models are also negatively impacted by the limited number of mineral deposits and occurrences available for training (Table 1). The CD models presented herein are based on 418 training points (Table 1). Generally, the extreme class imbalance for the training data are a major issue for most classification methods because they tend to overestimate the majority class. An alternative to classification methods is to locate the nearest neighbors of favorable host rocks using the cosine similarity of word vectors (i.e., semantic search; Fig. 14a). This form of nearest neighbor analysis only requires one example for ‘‘training'' and has been previously applied to geochemical surveys (Chen et al., 2019) and to improve fuzzy prospectivity modeling methods (Parsa et al., 2017).

Semantic search is entirely dependent on the training example used and the level of detail available in the text data, but, in theory, can be extended to find the nearest neighbors for any rock type or geoscientific concept (Fig. 13). Alternatively, cosine similarities can be as an intrinsic evaluation method with domain knowledge to make sure that nearest neighbors make geoscientific sense. For example, the nearest neighbors of ‘‘pegmatite'' are meaningful, and include closely associated words that were not considered in the original training data (i.e., aplite, granite, gneiss, biotite, granitic, muscovite, beryl, dike). Semantic search is used extensively in modern search engines and greatly expands the capabilities of geological bedrock map databases, even for cases with limited training data.

CONCLUSIONS

The volume of unstructured, geoscientific text data are rapidly expanding as digital technologies to record field observations continue to improve and as publication databases are increasingly made available for free on the internet. Geological map databases are also growing in sophistication by improving linkages with these external sources of text data, and by including long-form descriptions of rocks and their geological histories. These forms of unstructured text data are critical for the interpretation and application of geological map databases, but are difficult to represent on geological maps and much of the conceptual knowledge embedded in field observations is rarely used in practice. Herein, we address that knowledge gap using open-source NLP tools to extract meaningful semantic relationships between rock types from geological map databases across Canada, the U.S., and Australia. First, rock descriptions and associated text data were processed and tokenized before translating each map polygon to an average word vector using a geoscience GloVe model. The calculated vectoral representations of the original text data were then used to: (1) predict igneous rock-types (i.e., ‘‘pegmatitic'' and ‘‘alkalic'') that are important host rocks for critical minerals; (2) assess the mineral potential for CD and MVT Zn-Pb deposits based on the descriptions of favorable rock types; and (3) apply nearest neighbor analysis to search for analogues of the giant Mount Isa Zn-Pb deposit. Each of these applications have the potential to support mineral exploration for critical raw materials by targeting the most prospective rock types. The results further demonstrate how NLP can be used to extract knowledge from previously ''inaccessible'' text data, expanding the potential applications of unstructured text data within geological map databases. Directly utilizing the expertise encapsulated within geologic documents and maps coupled with machine learning techniques has the potential to transform how geoscientific analysis can be conducted.

ACKNOWLEDGMENTS

This work was completed as part of the Critical Mineral Mapping Initiative (www.criticalminerals.org). This work represents NRCan contribution number (20220496). Special thanks to Richard Scott for discussions on cosine similarities and distances. Catherine E. Brown (Geoscience Australia) is thanked for providing information on Australia lithostratigraphy. Geoscience Australia acknowledges the traditional custodians of the country where this work was undertaken. Published with the permission of the CEO of Geoscience Australia. Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government. The authors would like to thank Julie Bordeau, Joshua Rosera, an anonymous reviewer, and John Carranza for their constructive feedback.

FUNDING

Open Access funding provided by Natural Resources Canada. This study was funded by the Critical Minerals Geoscience and Data Initiative. The authors have no relevant financial or non-financial interests to disclose. All data generated or analyzed during this study are included in Lawley et al. (2022a, 2022b). The code and language model used to generate these results are available from https://github.com/NRCan/geoscience_language_models

SUPPLEMENTARY INFORMATION

The online version contains supplementary material available at https://doi.org/10.1007/s11053-023-10216-1.

REFERENCES

Adi, Y., Kermany, E., Belinkov, Y., Lavi, O., & Goldberg, Y. (2017). Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. arXiv. https://doi.org/10.48550/arXiv.1608.04207.

Airola, A., Pohjankukka, J., Torppa, J., Middleton, M., Nykänen, V., Heikkonen, J., & Pahikkala, T. (2019). The spatial leave-pair-out cross-validation method for reliable AUC estimation of spatial classifiers. *Data Mining and Knowledge Discovery, 33*(3), 730–747.

Alshanik, F., Apon, A., Herzog, A., Safro, I., & Sybrandt, J. (2020). Accelerating text mining using domain-specific stop word lists. In *2020 IEEE international conference on big data (big data)* (pp. 2639–2648). https://doi.org/10.1109/BigData50022.2020.9378226.

Ayral, H., & Yavuz, S. (2011). An automated domain specific stop word generation method for natural language text classification. In *2011 International symposium on innovations in intelligent systems and applications* (pp. 500–503). https://doi.org/10.1109/INISTA.2011.5946149.

Bengio, Y., Ducharme, R., & Vincent, P. (2000). A neural probabilistic language model. *Advances in Neural Information Processing Systems, 13*, 1–7.

Berg, R. C., MacCormack, K. E., & Russell, H. A. J. (2019). Chapter 4: Benefit-cost analysis for building 3D maps and models. In K. E. MacCormack, R. C. Berg, H. Kessler, H. A. J. Russell, & L. H. Thorleifson (Eds.), *2019 Synopsis of current three-dimensional geological mapping and modelling in geological survey organizations* (Vol. 112, pp. 19–23). Alberta Geological Survey, Alberta Energy Regulator, Edmonton, AB, Canada. https://ags.aer.ca/document/SPE/SPE_112.pdf#page=25. Accessed 19 December 2022.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. arXiv. https://doi.org/10.48550/arXiv.1607.04606.

Bouchet-Valat, M. (2020). SnowballC: Snowball stemmers based on the C ''libstemmer'' UTF-8 Library. https://CRAN.R-project.org/package=SnowballC.

Brodaric, B., & Gahegan, M. (2001). Learning geoscience categories in situ: Implications for geographic knowledge representation. In *Proceedings of the 9th ACM international symposium on advances in geographic information systems* (pp. 130–135). Association for Computing Machinery. https://doi.org/10.1145/512161.512190.

Brodaric, B. (2012). Characterizing and representing inference histories in geologic mapping. *International Journal of Geographical Information Science, 26*(2), 265–281.

Brodaric, B., Gahegan, M., & Harrap, R. (2004). The art and science of mapping: Computing geological categories from field data. *Computers & Geosciences, 30*(7), 719–740.

Burke, K., & Khan, S. (2006). Geoinformatic approach to global nepheline syenite and carbonatite distribution: Testing a Wilson cycle model. *Geosphere, 2*(1), 53–60.

Chen, J., Yousefi, M., Zhao, Y., Zhang, C., Zhang, S., Mao, Z., et al. (2019). Modelling ore-forming processes through a cosine similarity measure: Improved targeting of porphyry copper deposits in the Manzhouli belt, China. *Ore Geology Reviews, 107*, 108–118.

Chowdhary, K. R. (2020). Natural language processing. In K. R. Chowdhary (Ed.), *Fundamentals of artificial intelligence* (pp. 603–649). New Delhi: Springer. https://doi.org/10.1007/978-81-322-3972-7_19.

Chudasama, B., Torppa, J., Nykänen, V., & Kinnunen, J. (2022a). Target-scale prospectivity modeling for gold mineralization within the Rajapalot Au-Co project area in northern Fennoscandian Shield, Finland. Part 2: Application of self-organizing maps and artificial neural networks for exploration targeting. *Ore Geology Reviews, 147*, 104936.

Chudasama, B., Torppa, J., Nykänen, V., Kinnunen, J., Lerssi, J., & Salmirinne, H. (2022b). Target-scale prospectivity modeling for gold mineralization within the Rajapalot Au-Co project area in northern Fennoscandian Shield, Finland. Part 1: Application of knowledge-driven- and machine learning-based-hybrid- expert systems for exploration targeting and addressing model-based uncertainties. *Ore Geology Reviews, 147*, 104937.

Consoli, B., Santos, J., Gomes, D., Cordeiro, F., Vieira, R., & Moreira, V. (2020). Embeddings for named entity recognition in geoscience Portuguese literature. In *Proceedings of The 12th language resources and evaluation conference* (pp. 4625–4630). European Language Resources Association. https://aclanthology.org/2020.lrec-1.568.

Csardi, G., & Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal, Complex Systems, 1695*(5), 1–9.

Culshaw, M., Jackson, I., Peach, D., van der Meulen, M. J., Berg, R., & Thorleifson, H. (2021). Geological survey data and the move from 2-D to 4-D. In *Applied multidimensional geological modeling* (pp. 13–33). Wiley. https://doi.org/10.1002/9781119163091.ch2.

Dale, R. (2021). GPT-3: What's it good for? *Natural Language Engineering, 27*(1), 113–118.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805 *[cs]*. http://arxiv.org/abs/1810.04805.

Eberle, D. G., Daudi, E. X. F., Muiuane, E. A., Nyabeze, P., & Pontavida, A. M. (2012). Crisp clustering of airborne geophysical data from the Alto Ligonha pegmatite field, northeastern Mozambique, to predict zones of increased rare earth element potential. *Journal of African Earth Sciences, 62*(1), 26–34.

Enkhsaikhan, M., Holden, E.-J., Duuring, P., & Liu, W. (2021a). Understanding ore-forming conditions using machine reading of text. *Ore Geology Reviews, 135*, 104200.

Enkhsaikhan, M., Liu, W., Holden, E.-J., & Duuring, P. (2021b). Auto-labelling entities in low-resource text: A geological case study. *Knowledge and Information Systems, 63*(3), 695–715.

Floridi, L., & Chiriatti, M. (2020). GPT-3: Its Nature, scope, limits, and consequences. *Minds and Machines, 30*(4), 681–694.

Fruchterman, T. M. J., & Reingold, E. M. (1991). Graph drawing by force-directed placement. *Software: Practice and Experience, 21*(11), 1129–1164.

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

Giles, J. R. A., & Bain, K. A. (1995). The nature of data on a geological map. *Geological Society, London, Special Publications, 97*(1), 33–40.

Gomes, D. D. S. M., Cordeiro, F. C., Consoli, B. S., Santos, N. L., Moreira, V. P., Vieira, R., et al. (2021). Portuguese word embeddings for the oil and gas industry: Development and evaluation. *Computers in Industry, 124*, 103347.

González-Álvarez, I., Stoppa, F., Yang, X. Y., & Porwal, A. (2021). Introduction to the special Issue, insights on carbonatites and their mineral exploration approach: A challenge towards resourcing critical metals. *Ore Geology Reviews, 133*, 104073.

Harman, D. (1991). How effective is suffixing? *Journal of the American Society for Information Science, 42*(1), 7–15.

Hirschberg, J., & Manning, C. D. (2015). Advances in natural language processing. *Science, 349*(6245), 261–266.

Holden, E.-J., Liu, W., Horrocks, T., Wang, R., Wedge, D., Duuring, P., & Beardsmore, T. (2019). GeoDocA—Fast analysis of geological content in mineral exploration reports: A text mining approach. *Ore Geology Reviews, 111*, 102919.

Horton, J. D., San Juan, C. A., & Stoeser, D. B. (2017). *The state geologic map compilation (SGMC) geodatabase of the conterminous United States* (No. 1052). *Data Series*. U.S. Geological Survey. https://doi.org/10.3133/ds1052.

Howard, A. S., Hatton, B., Reitsma, F., & Lawrie, K. I. G. (2009). Developing a geoscience knowledge framework for a national geological survey organisation. *Computers & Geosciences, 35*(4), 820–835.

Huston, D. L., Champion, D. C., Czarnota, K., Duan, J., Hutchens, M., Paradis, S., et al. (2022). Zinc on the edge—Isotopic and geophysical evidence that cratonic edges control world-class shale-hosted zinc-lead deposits. *Mineralium Deposita*. https://doi.org/10.1007/s00126-022-01153-9.

Huston, D. L., Mernagh, T. P., Hagemann, S. G., Doublier, M. P., Fiorentini, M., Champion, D. C., et al. (2016). Tectono-metallogenic systems—The place of mineral systems within tectonic evolution, with an emphasis on Australian examples. *Ore Geology Reviews, 76*, 168–210.

Hvitfeldt, E., & Silge, J. (2021). *Supervised machine learning for text analysis in R* (1st ed.). Chapman and Hall/CRC.

Joshi, A. V. (2020). Amazon's machine learning toolkit: Sagemaker. In *Machine learning and artificial intelligence* (pp. 233–243). Springer. https://doi.org/10.1007/978-3-030-26622-6_24.

Kesler, S. E., Gruber, P. W., Medina, P. A., Keoleian, G. A., Everson, M. P., & Wallington, T. J. (2012). Global lithium resources: Relative importance of pegmatite, brine and other deposits. *Ore Geology Reviews, 48*, 55–69.

Kuhn, M. (2008). Building predictive models in R using the caret package. *Journal of Statistical Software, 28*, 1–26.

Lawley, C. J. M., McCafferty, A. E., Graham, G. E., Huston, D. L., Kelley, K. D., Czarnota, K., et al. (2022b). Data-driven prospectivity modelling of sediment–hosted Zn–Pb mineral systems and their critical raw materials. *Ore Geology Reviews, 141*, 104635.

Lawley, C. J. M., Raimondo, S., Chen, T., Brin, L., Zakharov, A., Kur, D., et al. (2022a). Geoscience language models and their intrinsic evaluation. *Applied Computing and Geosciences, 14*, 100084.

Laxton, J. L. (2017). Geological map fusion: OneGeology-Europe and INSPIRE. *Geological Society, London, Special Publications, 408*(1), 147–160.

Laxton, J. L., & Becken, K. (1996). The design and implementation of a spatial database for the production of geological maps. *Computers & Geosciences, 22*(7), 723–733.

Leach, D. L., Bradley, D., Lewchuk, M. T., Symons, D. T., de Marsily, G., & Brannon, J. (2001). Mississippi Valley-type lead–zinc deposits through geological time: Implications from recent age-dating research. *Mineralium Deposita, 36*(8), 711–740.

Lebel, D. (2020). Geological Survey of Canada 8.0: Mapping the journey towards predictive geoscience. *Geological Society, London, Special Publications, 499*(1), 49–68.

Li, W., Ma, K., Qiu, Q., Wu, L., Xie, Z., Li, S., & Chen, S. (2021). Chinese word segmentation based on self-learning model and geological knowledge for the geoscience domain. *Earth and Space Science, 8*(6), e2021EA001673.

Lincoln, L. A., Benoit, K., Keyes, O., Selivanov, D., & Arnold, J. (2018). Fast, consistent tokenization of natural language text. *Journal of Open Source Software, 3*(23), 655.

Lombardo, V., Piana, F., & Mimmo, D. (2018). Semantics-informed geological maps: Conceptual modeling and knowledge encoding. *Computers & Geosciences, 116*, 12–22.

London, D. (2005). Granitic pegmatites: An assessment of current concepts and directions for the future. *Lithos, 80*(1–4), 281–303.

Loudon, T. V. (2009). Four interacting aspects of a geological survey knowledge system. *Computers & Geosciences, 35*(4), 700–705.

Lyons, T. W., Gellatly, A. M., McGoldrick, P. J., & Kah, L. C. (2006). Proterozoic sedimentary exhalative (SEDEX) deposits and links to evolving global ocean chemistry. In S. E. Kesler & H. Ohmoto (Eds.), *Evolution of early earth's atmosphere, hydrosphere, and biosphere-constraints from ore deposits* (Vol. 198, pp. 169–184). Geological Society of America. https://doi.org/10.1130/2006.1198(10).

Ma, K., Tian, M., Tan, Y., Xie, X., & Qiu, Q. (2021). What is this article about? Generative summarization with the BERT model in the geosciences domain. *Earth Science Informatics*. https://doi.org/10.1007/s12145-021-00695-2.

Ma, X. (2022). Knowledge graph construction and application in geosciences: A review. *Computers & Geosciences, 161*, 105082.

Mantovani, A., Piana, F., & Lombardo, V. (2020). Ontology-driven representation of knowledge for geological maps. *Computers & Geosciences, 139*, 104446.

McCaffrey, K. J. W., Jones, R. R., Holdsworth, R. E., Wilson, R. W., Clegg, P., Imber, J., et al. (2005). Unlocking the spatial dimension: Digital technologies and the future of geoscience fieldwork. *Journal of the Geological Society, 162*(6), 927–938.

McCauley, A., & Bradley, D. C. (2014). Thye global age distribution of granitic pegmatites. *The Canadian Mineralogist, 52*(2), 183–190.

McCuaig, T. C., Beresford, S., & Hronsky, J. (2010). Translating the mineral systems approach into an effective exploration targeting system. *Ore Geology Reviews, 38*(3), 128–138.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013a). Efficient estimation of word representations in vector space. arXiv:1301.3781 *[cs]*. http://arxiv.org/abs/1301.3781.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. arXiv:1310.4546 *[cs, stat]*. http://arxiv.org/abs/1310.4546.

Mitchell, J., & Lapata, M. (2010). Composition in distributional models of semantics. *Cognitive Science, 34*(8), 1388–1429.

Morrison, S. M., Liu, C., Eleish, A., Prabhu, A., Li, C., Ralph, J., et al. (2017). Network analysis of mineralogical systems. *American Mineralogist, 102*(8), 1588–1596.

Nykänen, V., Lahti, I., Niiranen, T., & Korhonen, K. (2015). Receiver operating characteristics (ROC) as validation tool for prospectivity models—A magmatic Ni–Cu case study from the Central Lapland Greenstone Belt, Northern Finland. *Ore Geology Reviews, 71*, 853–860.

Padarian, J., & Fuentes, I. (2019). Word embeddings for application in geosciences: Development, evaluation, and examples of soil-related concepts. *The Soil, 5*(2), 177–187.

Parker, R., Graff, D., Kong, J., Chen, K., & Maeda, K. (2011). English Gigaword, 5th edition. Linguistic Data Consortium. https://doi.org/10.35111/WK4F-QT80.

Parsa, M., Maghsoudi, A., & Yousefi, M. (2017). An improved data-driven fuzzy mineral prospectivity mapping procedure; cosine amplitude-based similarity approach to delineate exploration targets. *International Journal of Applied Earth Observation and Geoinformation, 58*, 157–167.

Pavlis, T. L., Langford, R., Hurtado, J., & Serpa, L. (2010). Computer-based data acquisition and visualization systems in field geology: Results from 12 years of experimentation and future potential. *Geosphere, 6*(3), 275–294.

Pebesma, E. (2018). Simple features for R: Standardized support for spatial vector data. *The R Journal, 10*(1), 439–446.

Pedersen, T. L. (2021). ggraph: An implementation of grammar of graphics for graphs and networks. https://CRAN.R-project.org/package=ggraph.

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532–1543). https://doi.org/10.3115/v1/D14-1.

Peters, S. E., Husson, J. M., & Czaplewski, J. (2018). Macrostrat: A platform for geological data integration and deep-time earth crust research. *Geochemistry, Geophysics, Geosystems, 19*(4), 1393–1409.

Peters, S. E., Zhang, C., Livny, M., & Ré, C. (2014). A machine reading system for assembling synthetic paleontological databases. *PLoS ONE, 9*(12), e113523.

Pollock, D. W., Barron, O. V., & Donn, M. J. (2012). 3D exploratory analysis of descriptive lithology records using regular expressions. *Computers & Geosciences, 39*, 111–119.

Qiu, Q., Xie, Z., Wu, L., & Li, W. (2018). DGeoSegmenter: A dictionary-based Chinese word segmenter for the geoscience domain. *Computers & Geosciences, 121*, 1–11.

Qiu, Q., Xie, Z., Wu, L., & Li, W. (2019). Geoscience keyphrase extraction algorithm using enhanced word embedding. *Expert Systems with Applications, 125*, 157–169.

R Core Team. (2023). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/.

Ramdeen, S. (2015). Preservation challenges for geological data at state geological surveys. *GeoResJ, 6*, 213–220.

Raymond, O. L., Duclaux, G., Boisvert, E., Cipolloni, C., Cox, S., Laxton, J., et al. (2012a). GeoSciML v3.0—A significant upgrade of the CGI-IUGS geoscience data model. *Geophysical Research Abstracts*, *14*, 2711. Presented at the EGU General Assembly Conference Abstracts.

Raymond, O. L., Liu, S., Gallagher, R., Highet, L., & Zhang, W. (2012b). Surface geology of Australia 1: 1 million scale dataset 2012b edition. *Geoscience Australia, Canberra*. https://doi.org/10.26186/74619.

Reed, J. C., Jr., Wheeler, J. O., Tucholke, B. E., Stettner, W. R., & Soller, D. R. (2005). Decade of North American geology geologic map of North America—Perspectives and explanation. In J. C. Reed Jr., J. O. Wheeler, B. E. Tucholke, W. R. Stettner, & D. R. Soller (Eds.), *Decade of North American geology geologic map of North America—Perspectives and explanation* (Vol. 1, pp. 1–28). Geological Society of America.

Reitsma, F., Laxton, J., Ballard, S., Kuhn, W., & Abdelmoty, A. (2009). Semantics, ontologies and eScience for the geosciences. *Computers & Geosciences, 35*(4), 706–709.

Schetselaar, E. M. (1995). Computerized field-data capture and GIS analysis for generation of cross sections in 3-D perspective views. *Computers & Geosciences, 21*(5), 687–701.

Selivanov, D., & Wang, Q. (2016). text2vec: Modern text mining framework for R. https://cran.r-project.org/web/packages/text2vec.

Sen, M., & Duffy, T. (2005). GeoSciML: Development of a generic GeoScience Markup Language. *Computers & Geosciences, 31*(9), 1095–1103.

Sharpe, T. (2015). The birth of the geological map. *Science, 347*(6219), 230–232.

Shen, D., Wang, G., Wang, W., Min, M. R., Su, Q., Zhang, Y., et al. (2018). Baseline needs more love: on simple word-embedding-based models and associated pooling mechanisms. In *Proceedings of the 56th annual meeting of the association for computational linguistics (Volume 1: Long Papers)* (pp. 440–450). Presented at the ACL 2018. Association for Computational Linguistics. https://doi.org/10.18653/v1/P18-1041.

Silge, J., & Robinson, D. (2016). tidytext: Text mining and analysis using tidy data principles in R. *Journal of Open Source Software, 1*(3), 37.

Simons, B., Boisvert, E., Brodaric, B., Cox, S., Duffy, T. R., Johnson, B. R., et al. (2006). GeoSciML: Enabling the exchange of geological map data. *ASEG Extended Abstracts, 2006*(1), 1–4. https://doi.org/10.1071/aseg2006ab162.

Stephenson, M., Wang, C., Cheng, Q., Shen, S., Fan, J., & Oberhansli, R. (2022). Deep-time digital earth programme of the international union of geological sciences: Connecting and harmonising deep-time data (Vol. 2022, pp. 1–5). Presented at the 83rd EAGE annual conference & exhibition. European Association of Geoscientists & Engineers. https://doi.org/10.3997/2214-4609.202210348.

Thorleifson, H. (2005). Geological map of the future: digital, interactive, and three-dimensional. In *The current role of geological mapping in geosciences* (pp. 23–24). Presented at the NATO advanced research workshop on innovative applications of GIS in geological cartography. Springer. https://doi.org/10.1007/1-4020-3551-9_3.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. arXiv: 1706.03762 *[cs]*. http://arxiv.org/abs/1706.03762.

Wang, B., Ma, K., Wu, L., Qiu, Q., Xie, Z., & Tao, L. (2022). Visual analytics and information extraction of geological content for text-based mineral exploration reports. *Ore Geology Reviews, 144*, 104818.

Wang, C., Ma, X., Chen, J., & Chen, J. (2018). Information extraction and knowledge graph construction from geoscience literature. *Computers & Geosciences, 112*, 112–120.

Wheeler, J., Hoffman, P., Card, K., Davidson, A., Sanford, B., Okulitch, A., & Roest, W. (1996). Geological map of Canada/Carte géologique du Canada. *Geological Survey of Canada, ''A'' Series Map 1860A,* 3 sheets; 1 CD-ROM. https://doi.org/10.4095/208175.

Whitmeyer, S., Nicoletti, J., & De Paor, D. (2010). The digital revolution in geologic mapping. *GSA Today*. https://doi.org/10.1130/GSATG70A.1.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., et al. (2019). Welcome to the Tidyverse. *Journal of Open Source Software, 4*(43), 1686.

Wieting, J., Bansal, M., Gimpel, K., & Livescu, K. (2016). Towards universal paraphrastic sentence embeddings. arXiv. https://doi.org/10.48550/arXiv.1511.08198.

Wilson, F. H., Hults, C. P., Mull, C. G., & Karl, S. M. (2015). Geologic map of Alaska. *U.S. Geological Survey Scientific Investigations Map 3340, Pamphlet, 196*, 2. https://doi.org/10.3133/sim3340.

Woolley, A. R., & Kjarsgaard, B. A. (2008). Paragenetic types of carbonatite as indicated by the diversity and relative abundances of associated silicate rocks: Evidence from a global database. *The Canadian Mineralogist, 46*(4), 741–752.

Wyborn, L. A. I., Heinrich, C. A., & Jaques, A. L. (1994). Australian proterozoic mineral systems: essential ingredients and mappable criteria. In *The AusIMM annual conference* (Vol. 1994, pp. 109–115). AusIMM Darwin.

Zuo, R., & Wang, Z. (2020). Effects of random negative training samples on mineral prospectivity mapping. *Natural Resources Research, 29*(6), 3443–3455.