GSA TODAY

THE GEOLOGICAL SOCIETY OF AMERICA®

VOL. 32, NO. 11 | NOVEMBER 2022

Answering Geosciences Research Questions at Global Scale via a Hybrid Machine–Human Learning Approach: A Case Study of the Link between Climate and Volcanism

THE GEOLOGICAL SOCIETY OF AMERICA®

Answering Geosciences Research Questions at a Global Scale via a Hybrid Machine-Human Learning Approach: A Case Study of the Link between Climate and Volcanism

Seongjin Park*, Barbara Carrapa#, *University of Arizona, Tucson, Arizona 85721, USA;* ***Mihai N. Ducea,*** *University of Arizona, Tucson, Arizona 85721, USA, and Faculty of Geology and Geophysics, University of Bucharest, 010041, Bucharest, Romania;* ***Mihai Surdeanu, Robert Hayes, Dan Collins,*** *University of Arizona, Tucson, Arizona 85721, USA*

ABSTRACT

A common challenge in science is the human capability to evaluate the real impact of an observation and a data set. This is a complex task due to having only partial information and/or to the complexity of the problem, requiring different fields to be combined. In order to overcome these important limitations, we need to be able to review all the available data and interpretations. This would allow us to evaluate the global distribution of a specific process or phenomenon of interest. The increasing number of scientific publications prevents scientists from being able to keep up with all the available literature especially when scientific papers cross disciplines. These challenges prevent us from evaluating the global impact of a certain process and are particularly relevant today given the impact of our scientific assessment on one of the most pressing issues of our time, which is climate change and its impact on society. We present here an application of artificial intelligence to geosciences: We conduct a systematic analysis of geoscience literature through a hybrid machine-human approach. Such applications are more common in other fields such as biomedicine and are in their infancy in the geosciences because of various difficulties the machines encounter in parsing geologic literature. We describe here some of these limitations and how we overcame them. We then use the following case study as an example to test our approach: We ask whether climate is influenced by volcanism in the geologic past. Our case study results show, as expected, that most analyzed literature in this experiment conclude that volcanism influences climate change in deep time, but there is no complete consensus on this question. Similarly, any question of potential global significance, such as the impact of human activities on climate change, can be posed as an interrogating technique for our vast and fast-growing literature in the field of geosciences. Such an approach has the potential to be applied to a variety of complex problems, hence addressing some of the major limitations with cross-disciplinary research.

INTRODUCTION

One of the cornerstone theories in natural sciences, Darwin's evolutionism, states that the evolution of flora and fauna in the geologic past goes through temporally determined and irreversible extinctions corroborated with the development of new species. That theory has been vetted by innumerable observations and stands today because of that. However, most potentially groundbreaking research questions in natural sciences have a difficult time being resolved at global scales because of the complexity of observations. In order to answer such questions at a global scale, we need to have a global review of the scientific literature. This task has turned into a near impossible challenge in recent years due to the vast amount of scientific data that have been published, which exceeds human capacity for processing and interpretation. This is particularly problematic in fields like geosciences that require the interpretation of data and research questions on a global scale and over large time intervals. Whereas data pertaining to a specific field (e.g., regional geology) of a particular area can still be tracked by the interested geologist (the number of papers is still within reach of human processing), the importance of so many global-scale multidisciplinary interpretations is difficult to evaluate. For example, did erosion of Earth's surface increase globally since the Pliocene as the result of increased climate variability (Zhang et al., 2001; Herman et al., 2013)? Was tectonics the cause of CO_2 drawdown and global cooling in the Cenozoic (e.g., Raymo and Ruddiman, 1992; Gernon et al., 2021)? Did Earth's surface topography affect biodiversity through time (Badgley et al., 2017)? These are just a couple of examples of far-reaching but hard-to-evaluate research questions in a science that increasingly requires ingestion of too much information at a global scale and that commonly needs to be placed into a complex deep time–space framework.

To address these issues, we built a hybrid machine-human approach for the systematic analysis of scientific discoveries in geosciences. The proposed approach employs machine reading to ingest publications at scale and aggregate scientific discoveries. These models allow scientists to attempt a wider understanding of science, which facilitates the identification of (apparent) contradictions in scientific findings, as well as "blank spaces" in the research landscape.

Note that approaches that summarize scientific work already exist, such as SCITE (https://scite.ai) and SCITLDR (https://scitldr.apps.allenai.org), both of which are trained on previously published papers. However, their goals are different from what we aimed to achieve in our study. SCITE

GSA Today, v. 32, https://doi.org/10.1130/GSATG528A.1. BY-CC-NC.

*This author is currently affiliated with Educational Testing Service.

#Corresponding author: bcarrapa@arizona.edu

analyzes the relationship between citations and their textual context (i.e., whether the citation is used in a positive way or negative way). SCITLDR is used to create a short summary of the given paper (without truly understanding what the underlying content means). Our work is complementary to these directions, because we aim for deeper language understanding. That is, the purpose of the proposed approach is to spatially and temporally contextualize a given geoscience research question and to identify whether the content of the papers analyzed supports or negates it.

For this purpose, we developed an application to geosciences to demonstrate the potential of our proposed approach to experiment with the limitations of this type of literature and how they can be overcome. The application investigates the research question of whether there is a causal relationship between volcanism and climate change in the geologic record as seen through the lens of published literature. Specifically, we ask whether volcanism influenced climate change in the deep time geologic archive. We selected this question because several geological studies seem to support this link (e.g., Lee and Dee, 2019). Our results indicate more variability on whether or not available studies on the subject actually support this research question.

SYSTEMATIC MACHINE REVIEW OF GEOSCIENCE DATA

Since there was no pre-built corpus for this geosciences task, we extracted 1164 papers from the Web of Science website via the University of Arizona's library. These papers were selected because they contained keywords relevant to the research question at hand, such as *volcanism* or *magmatism*, and *climate change*. This was implemented as the Boolean query: (*volcanism* OR *magmatism)* AND "*climate change,*" where OR and AND are the disjunctive and conjunctive Boolean operators, and quotes indicate that the entire phrase must be present. This query extracted 1164 papers from the Web of Science. We then randomly chose 200 papers and extracted the abstract, introduction, and conclusion sections from each paper to be manually annotated with the information if they support or do not support the research question. Note that for this work we assume that the authors' data, interpretations, and conclusions are correct. The annotation task was conducted on FindingFive (https://www.findingfive.com), an online annotation platform. The papers were placed into one of four classes: SUPPORT, NEGATE, NEGATE&SUPPORT, and UNRELATED (see Table 1). The annotations for these four classes were collected by two of the coauthors of this effort, who are domain experts (i.e., geoscientists). The two annotators worked independently.

Next, we implemented a natural language processing (NLP) component for geosciences that extracts two types of information. First, we contextualized individual publications by extracting and normalizing the geospatial and temporal contexts addressed in these papers (e.g., *Pliocene*, *4 million years ago*, and *Bering Sea*). For example, *Tucson* and *Saguaro National Park* can be considered as the same geographic location (for the purposes of this analysis), even though they are described differently in text. To facilitate the consolidation of findings, we normalized the geospatial contexts to absolute latitude/longitude coordinates (see the next section for details). Similarly, temporal expressions such as *4 million years ago* were converted to geological eras or epochs (e.g., *Paleoproterozoic*) to have a better overall understanding of the relationship between volcanism and climate change on the geological time scale.

Second, we built a document classifier that is trained to determine whether any given paper supports the observation that "volcanism affected climate change," so that we could make a prediction on *new* papers. The results of these two components were aggregated into a publication knowledge base, which contains the publication itself, the prediction of our classifier (SUPPORT, NEGATE, NEGATE&SUPPORT, and UNRELATED—see Table 1 for details), the occurrence of geological eras and epochs (e.g., the frequency of *Pliocene* in a given paper), and the occurrence of geological locations (e.g., the frequency of *Africa* in a given paper). We used this knowledge base to visualize the evidence for the research question investigated on the world map to identify global temporal and geospatial patterns.

THE HYBRID MACHINE-HUMAN APPROACH

Below, we detail the three key components of our hybrid machine-human approach in this experiment.

Contextualizing Findings: Time and Site Identification

To analyze the relationship between volcanism and climate change at different times in the geological past and locations, we built a custom *Named Entity Recognizer* to extract spatial and temporal information from the analyzed text. Named entity recognition (NER) is a common NLP task that aims to identify named entities within the given text and classify or categorize those entities under various predefined classes. Our focus in this work is on the identification of locations and geological eras and epochs, which are necessary to contextualize the findings discussed in the papers.

Existing NER tools such as Stanford's CoreNLP (Manning et al., 2014) or spaCy (Honnibal and Montani, 2017) focus on generic locations, times, and dates rather than geoscience-specific ones. For example, when we fed the sample sentence "Clay mineral assemblages and crystallinities in sediments from IODP Site 1340 in the Bering Sea were analyzed in order to trace sediment sources and reconstruct the paleoclimatic history of the Bering Sea since Pliocene (the last 4.3 Ma)" into the Stanford CoreNLP NER, the result was:

Clay mineral assemblages and crystallinities in sediments from IODP Site [1340] DATE in the [Bering Sea]LOCATION were analyzed in order to trace sediment sources and reconstruct the [paleoclimatic] MISC history of the [Bering Sea] LOCATION since Pliocene (the last [4.3] NUMBER Ma).

Even though the Stanford CoreNLP NER correctly identified *Bering Sea* as a LOCATION, it did not recognize geo-sciences-specific expressions, and, further, it classified expressions into the incorrect

TABLE 1. NAMES AND DESCRIPTIONS OF THE LABELS USED DURING THE MACHINE CLASSIFICATION PROCESS*

Classification label	Definition
Support	The given text supports the relationship between volcanism and climate change.
Negate	The given text negates the relationship between volcanism and climate change.
Negate&Support	The same overall text both supports and negates the relationship between volcanism and climate change, with different paragraphs discussing each relationship.
Unrelated	The given text is unrelated to the topic at hand, i.e., the relationship between volcanism and climate change.

*See text footnote 1.

entity types. For example, *IODP Site 1340* (IODP stands for Integrated Ocean Discovery Program) refers to a certain location, but the recognizer identified only *1340*, and classified it incorrectly as a DATE. The recognizer missed the term *Pliocene*, which means "the geologic time scale that extends from 5.333 million to 2.58 million years B.P." *Ma* in geosciences articles usually means *million years ago*, but the CoreNLP NER did not identify it as TIME.

To recognize expressions that were not identified by CoreNLP or Spacy, we used the Odin event extraction framework and rule language (Valenzuela-Escárcega et al., 2016); henceforth, Odin), and added custom rules to capture geoscience-specific expressions. In particular, we developed rules to capture:

Temporal Information

As mentioned, initially we utilized the named entity recognition tool in Stanford's CoreNLP (Manning et al., 2015); henceforth, CoreNLP) to identify time information. However, since CoreNLP was trained on general text data, it does not recognize geological temporal expressions, such as Paleocene or Jurassic. In addition, in geosciences papers, there were abbreviations such as *M.y.r.* and *M.a.*, which mean *millions of years* (duration), and *million years ago* (absolute time). Thus, we wrote custom rules to recognize geological temporal expressions and built a custom time normalizer to convert actual times (e.g., *170 M.y.r.*, or *1.5 million years ago*) to relevant geological time scale (e.g., *Jurassic*, *Quaternary*) (see supplemental document 1[1] for specific details on these rules).

Site Information

Similar to temporal information, there were domain-specific spatial expressions that could not be captured by existing NERs such as Stanford's CoreNLP. Further, some of these expressions did not have any information about the actual locations that they indicate. Thus, we wrote scripts to extract spatial expressions, disambiguate geoscience-specific spatial expressions (e.g., *IODP Site U1360*), and normalize these expressions by aligning them with specific latitude-longitude bounding boxes that indicate the actual location of the corresponding spatial expressions on the world map (see supplemental document 2 [see footnote 1]).

CLASSIFYING THE SUPPORT FOR THE RESEARCH QUESTION OF INTEREST

Even though these spatial and temporal expressions are important to contextualize the findings of a publication, they provide no information on our key research question: whether volcanism affected climate change. To make a prediction of whether the given paper supports or negates the relationship between volcanism and climate change, it is necessary to build a machine learning classifier that infers if the observation is supported (or not) from the text of these publications.

Among the wide variety of text classification methods, in this work we focused on four methods that have been shown to perform well for text classification, including "traditional" statistical methods as well as deep learning. To represent the traditional "camp," we used Support Vector Machines (Cortes and Vapnik, 1995) and Naïve-Bayes SVMs (NB-SVMs) (Wang and Manning, 2012). For the deep learning field, we implemented a Multi-Layer Perceptron (henceforth, MLP) that operates on the same features as the above SVM variants. Last, we implemented an ensemble strategy that combines the outputs of these three individual models.

To prevent the classifiers from overfitting on the training data, we used L2 regularization when training the statistical classifiers that support it (i.e., SVM, NB-SVM, and MLP classifiers). Intuitively, regularization aims to "zero out" the features that are not critical to the task, which reduces the potential of overfitting, or "hallucinating a classifier" (Domingos, 2015). All document classification routines are detailed in supplemental document 3 (see footnote 1).

Data Annotation

Data annotation was performed via FindingFive. Two hundred papers were randomly chosen from the set of 1157 downloaded papers, and then title, abstract, introduction, conclusion/discussion sections of 200 papers were presented to the two

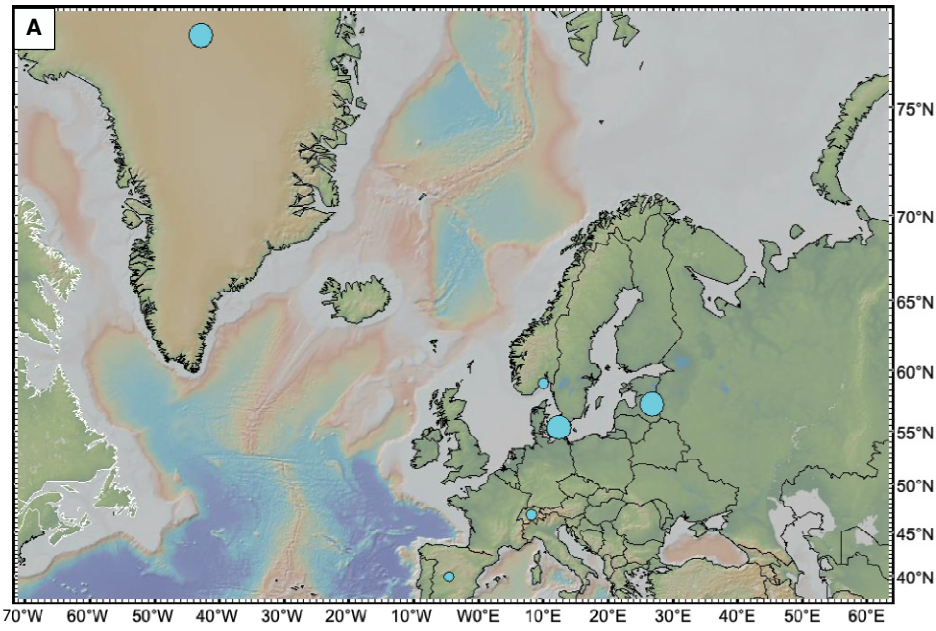

Figure 1. (A) Topographic map of Europe with circles representing the most frequent location found in each paper where the relationship between volcanism and climate change has been tested during the Cenozoic. Light blue circles indicate the locations where the impact of volcanism on climate change was verified, and pink circles indicate the locations where previous research negated the relationship between volcanism and climate change. The size of the circles represents its frequency; i.e., the number of publications supporting it. (B) Topographic map of North America with circles representing the top three most frequent locations found in each paper where the relationship between volcanism and climate change has been tested during the Cenozoic. (C) Topographic map of northern Europe with circles representing the most frequent location found in each paper where the relationship between volcanism and climate change has been tested during the Phanerozoic. (D) Topographic map of Europe and Asia with circles representing the top three most frequent locations found in each paper where the relationship between volcanism and climate change has been tested during the Cenozoic. **(*Continued on following page*.)**

[1]Supplemental Material. Supplemental Documents 1–3. Go to https://doi.org/10.1130/GSAT.S.20030015 to access the supplemental material; contact editing@geosociety.org with any questions.

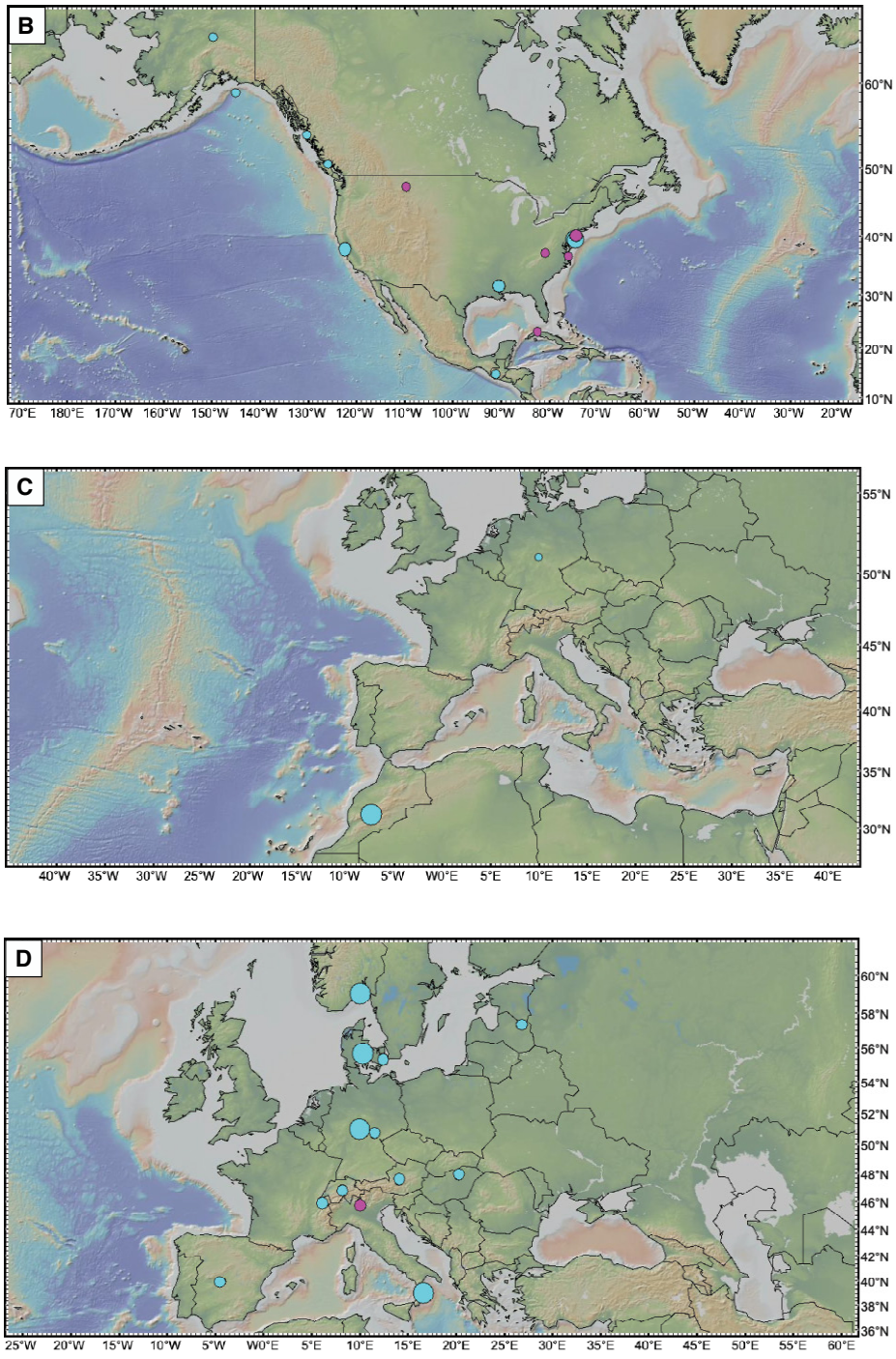

Figure 1 (*continued from page 6*).

annotators. After reading the provided text, the annotators determined whether the given paper supported or negated the relationship between volcanism and climate change. As a result, we produced 400 annotation results (200 papers × 2 annotators). All of 400 annotation results were used as a data set to train, validate, and evaluate the proposed system. Thus, even the disagreement between two annotators was used as data so that the proposed system could learn the ambiguity (see supplemental document 3). Before conducting the annotation session, authors discussed annotation criteria using papers that were *not* selected for annotation. To measure the agreement between annotators, Cohen's kappa score (Cohen, 1968) was measured. Cohen's kappa score is a commonly used metric to measure the agreement between two annotators. The Kappa result was 0.523, which showed moderate agreement between annotators (Landis and Koch, 1977). In other words, the two annotators somewhat agreed on whether a given paper supported or negated the observation that "volcanism affected the climate change." This "moderate" agreement is often found in this type of annotation task since the research question itself is quite complex and only part of the papers (e.g., abstract, introduction, conclusion) was provided to the annotators.

Classification of Results

We evaluated the quality of the proposed classifiers that were trained on the annotations by comparing the micro-F1 score calculated using 10-fold cross validation. More formally, we collected the algorithm's predictions on each test partition, and calculated the micro-F1 score (see supplemental material, including a formal definition of these measures in document 3) from *all* these predictions.

In these experiments, we observed that the MLP classifier outperforms both the NB-SVM and SVM classifiers, and that the ensemble approach does not improve over the performance of the MLP method (see supplemental document 3 for all these results). Informed by these results, we used the MLP model to classify all the 957 remaining papers in the collected data set on whether they supported/negated or were unrelated to the research question at hand.

Aggregation of Results for Visualization

With the two components described above that (a) place a scientific finding in its proper geospatial and temporal context, and (b) identify if publications support or negate the research question at hand, we can aggregate and visualize results at scale. To further simplify the visualizations, we used the *geopy* (https://pypi.org/project/geopy/) Python library to convert IODP sites to latitudes and longitudes, and we converted the identified specific geological periods and epochs into broader (larger time intervals) geological eras. For each paper analyzed, we used the most frequent top k (where $k = 1$, or $k = 3$) spatial and temporal entities for context.

Figure 1 shows several visualizations of the results, with light blue indicating support for the observation that volcanism impacts climate change and pink negating the observation. The sizes of the circles were determined based on the number of papers that the classifier predicted the corresponding label (i.e., light blue for

SUPPORT, and pink for NEGATE). Figure 1A shows the most frequent locations during the Cenozoic in Europe, and Figure 1 shows the top three most frequent locations during the Cenozoic in North America. When manually inspecting the machine prediction results from the MLP model, the domain experts observed that 11 out of 17 data points within the North American continent were correctly identified and visualized on the world map. Out of the six errors, four data points were from simulation papers, and two data points were based on incorrect predictions by the MLP classifier, as identified by the domain experts. For example, one pink circle (i.e., the corresponding paper was classified as not supporting the observation that volcanism impacts climate change) was incorrectly predicted when the actual paper was unrelated with respect to this observation.

These figures immediately highlight several important observations:

- Our data processing reduces the search space by almost two orders of magnitude (from ~1,000 papers that are shallowly related to the topic of interest to 17 that validate/invalidate the current observation that volcanism affects climate change), while our visualizations allow the scientist to quickly draw important conclusions that would not be easily available otherwise. For example, our figures show that while the majority of publications support the hypothesis investigated that volcanism impacts climate change, not all do.
- Similarly, this bird's-eye-view of a scientific question allows one to quickly identify "blank spaces" in research, i.e., topics that are insufficiently investigated. For example, our visualizations show that while support for our research question is well represented for the North American continent, it is scarce in other continents.
- Further, this work allows one to identify (potential) contradictions in scientific findings quickly, which provides opportunities for better science. For example, Figure 1B shows apparent contradictions in findings from the East coast of the North American continent in the Cenozoic.
- Lastly, the fact that 11 out the 17 identified papers are correctly classified is not surprising considering that none of the automated components (i.e., the module that extracts temporal and spatial context, and the research question classifier) are perfect. However, this result emphasizes that the human/machine interaction must continue if this system is to be improved.

All in all, this experiment finds strong support in favor of feedbacks existing between volcanism and climate change. However, the precise correlation is not a simple one. Our literature parsing system suggests that we do not yet have a clear and complete understanding of how volcanic events affect climate change.

CONCLUSIONS

The result of this preliminary work introduced a methodology to automatically provide a global review of the geoscientific literature and to evaluate the impact of specific research questions (i.e., understand if the question is [mostly] supported or rejected by the literature), in this case the causal relationship between volcanism and climate change. We show the promises and limitations of this approach to the geoscience literature with this admittedly simplistic example. This approach helps us process and interpret a large amount of published scientific papers, without the need for human annotators to invest time in reading and parsing all of the papers. In addition, with the visualization, researchers are able to investigate chronological changes in the relationship between volcanism and climate change. This approach could be expanded to any number of queries in the geoscience literature for the systematic analysis of various observations and ideas by examining a large body of previously published papers. Results can be further plotted on reconstructed various sample or study locations using paleogeographic maps.

It is vital to emphasize that the proposed methodology is hybrid, requiring direct collaboration between humans and machines. For example, geoscientists were required to provide training data for our research question classifier. Further, as discussed, our resulting classifier is only ~80% accurate, which means that, in order to improve it, it needs continuous feedback from the scientists using it. Longer term, we envision a community-wide effort in which such classifiers are created and deployed in the cloud to mine an arbitrary number of observations and are continuously improved over time by their human end users.

REFERENCES CITED

Badgley, C., Smiley, T.M., Terry, R., Davis, E.B., DeSantis, L.R.G., Fox, D.L., Hopkins, S.S.B., Jezkova, T., Matocq, M.D., Matzke, N., McGuire, J.L., Mulch, M., Riddle, B.R., Roth, V.L., Samuels, J.X., Strömberg, C.A.E., and Yanites, B.J., 2017, Biodiversity and topographic complexity: Modern and geohistorical perspectives: Trends in Ecology & Evolution, v. 32, no. 3, p. 211–226, https://doi.org/10.1016/j.tree.2016.12.010.

Cohen, J., 1968, Weighed kappa: Nominal scale agreement with provision for scaled disagreement or partial credit: Psychological Bulletin, v. 70, no. 4, p. 213–220, https://doi.org/10.1037/h0026256.

Cortes, C., and Vapnik, V., 1995, Support-vector networks: Machine Learning, v. 20, no. 3, p. 273–297, https://doi.org/10.1007/BF00994018.

Domingos, P., 2015, The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World: New York, Basic Books, 352 p.

Gernon, T.M., Hincks, T.K., Merdith, A.S., Rohling, E.J., Palmer, M.R., Foster, G.L., Bataille, C.P., and Müller, R.D., 2021, Global chemical weathering dominated by continental arcs since the mid-Palaeozoic: Nature Geoscience, v. 14, p. 690–696, https://doi.org/10.1038/s41561-021-00806-0.

Herman, F., Seward, D., Valla, P.G., Carter, A., Kohn, B., Willett, S.D., and Ehlers, T.A., 2013, Worldwide acceleration of mountain erosion under a cooling climate: Nature, v. 504, p. 423, https://doi.org/10.1038/nature12877.

Landis, J.R., and Koch, G.G., 1977, The measurement of observer agreement for categorical data: Biometrics, v. 33, no. 1, p. 159–174, https://doi.org/10.2307/2529310.

Lee, C-T., and Dee, S., 2019, Does volcanism cause warming or cooling?: Geology, v. 47, no. 7, p. 687–688, https://doi.org/10.1130/focus072019.1.

Honnibal, M., and Montani, I., 2017, spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing: https://spacy.io/.

Manning., C.D., 2015, Computational linguistics and deep learning: Computational Linguistics, v. 41, no. 4, p. 701–707, https://doi.org/10.1162/COLI_a_00239.

Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., and McClosky, D., 2014, The Stanford CoreNLP Natural Language Processing Toolkit: https://doi.org/10.3115/v1/p14-5010.

Raymo, M.E., and Ruddiman, W.F., 1992, Tectonic forcing of late Cenozoic climate: Nature, v. 359, p. 117–122.

Valenzuela-Escárcega, M.A., Hahn-Powell, G., and Surdeanu, M., 2016, Odin's Runes: A rule language for information extraction, *in* Proceedings of the 10th International Conference on Language Resources and Evaluation, LREC 2016, https://aclanthology.org/L16-1050.

Wang, S., and Manning, C.D., 2012, Baselines and bigrams: Simple, good sentiment and topic classification, *in* 50th Annual Meeting of the Association for Computational Linguistics, ACL 2012—Proceedings of the Conference, https://doi.org/https://dl.acm.org/doi/10.5555/2390665.2390688.

Zhang, P., Molnar, P., and Downs, W.R., 2001, Increased sedimentation rates and grain sizes 2–4 Myr ago due to the influence of climate change on erosion rates: Nature, v. 410, p. 891–897, https://doi.org/10.1038/35073504.

Manuscript received 16 Nov. 2021
Revised manuscript received 6 May 2022
Manuscript accepted 23 May 2022