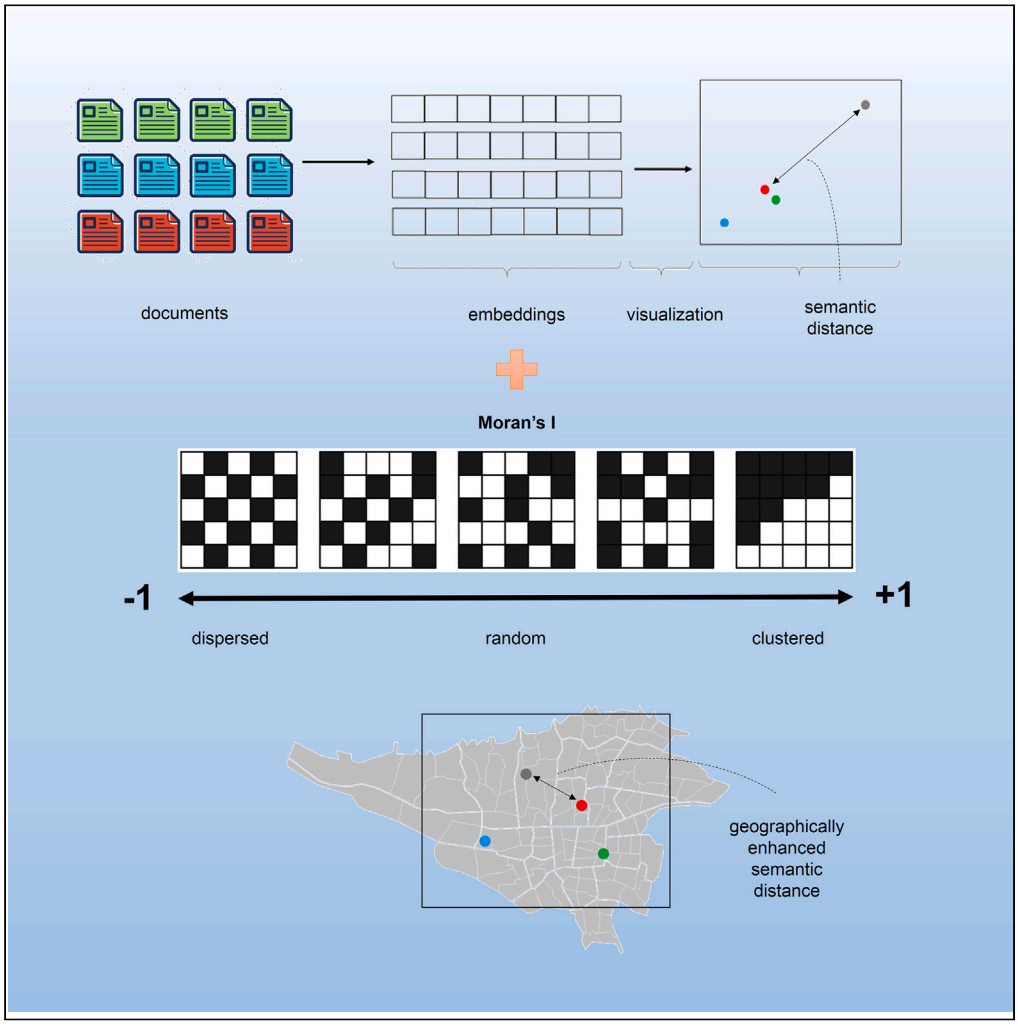
iScience

CellPress
OPEN ACCESS

Article

Semantic similarity is not enough: A novel NLP-based semantic similarity measure in geospatial context



Omid Reza Abbasi, Ali Asghar Alesheikh, Aynaz Lotfata

alesheikh@kntu.ac.ir (A.A.A.)
alotfata@ucdavis.edu (A.L.)

Highlights

Semantic similarity has no intrinsic correlation with geographical distance

Moran's I can be used to geographically enhance the embeddings in NLP

Users are more satisfied with the geographically enhanced embeddings

Abbasi et al., iScience 27, 109883
June 21, 2024
https://doi.org/10.1016/j.isci.2024.109883

iScience

CellPress
OPEN ACCESS

Article

Semantic similarity is not enough: A novel NLP-based semantic similarity measure in geospatial context

Omid Reza Abbasi,[1] Ali Asghar Alesheikh,[1,3,*] and Aynaz Lotfata[2,*]

SUMMARY

In this study, we addressed two primary challenges: firstly, the issue of domain shift, which pertains to changes in data characteristics or context that can impact model performance, and secondly, the discrepancy between semantic similarity and geographical distance. We employed topic modeling in conjunction with the BERT architecture. Our model was crafted to enhance similarity computations applied to geospatial text, aiming to integrate both semantic similarity and geographical proximity. We tested the model on two datasets, Persian Wikipedia articles and rental property advertisements. The findings demonstrate that the model effectively improved the correlation between semantic similarity and geographical distance. Furthermore, evaluation by real-world users within a recommender system context revealed a notable increase in user satisfaction by approximately 22% for Wikipedia articles and 56% for advertisements.

INTRODUCTION

User-generated content provides a vast amount of data for natural language processing (NLP) applications such as text similarity detection.[1] Text similarity is divided into two categories: syntactic similarity and semantic similarity. Syntactic similarity is a measure of how closely related two sentences are in terms of their syntax, irrespective of their meaning.[2,3] Semantic similarity measures how closely two sentences are related in meaning.[4] State-of-the-art models for semantic similarity computations typically involve transferring the documents to an embedding space and subsequently measuring the similarity between documents based on the distances within this space.[5]

Recent studies have shown that the establishment of the embedding space is influenced when the texts are associated with a specific domain.[6–8] A challenge in forming the semantic space for texts involves determining their similarity within a specific domain.[9] This challenge occurs because words used in a specific domain often refer to similar things, resulting in higher levels of similarity.[10] Furthermore, in applications such as location-based recommender systems, depending only on semantic similarity is not enough to provide relevant suggestions to users. For instance, in a recommender system for housing and apartment sales and rentals, users expect the suggested options to be near their initial search and also connected in meaning to their previous searches. As a result, a metric that considers the geographic indicativeness of the items can significantly improve location-based recommenders. This study introduces a geospatial semantic similarity approach wherein the computation of semantic similarity between texts is undertaken with a focus on prioritizing text qualities that exhibit greater geographically indicative characteristics.

A number of studies have looked into how to measure semantic similarity.[11–14] Wilcox et al.[15] developed a new method to improve the accuracy of medical record data by addressing encoding errors and missing information. The algorithm aims to create a reliable dataset essential for verifying individuals in medical scenarios. Results showed that SSIM outperforms other measures, especially in managing variations such as nicknames, abbreviations, and synonyms. Hendre et al.[16] explored the application of semantic similarity in an automated essay evaluation system, utilizing various text embedding methods, including deep neural embeddings such as Google Sentence Encoder (GSE), Embeddings for Language Models, and Global Vectors. Experimental findings show that GSE is superior to other methods in differentiating essays within the same or different sets, and its semantic similarity scores align with human-assessed essay ratings. Curiskis et al.[17] assessed document clustering and topic modeling methods for online social networks, focusing on Twitter and Reddit datasets. Four feature representations, combining TF-IDF matrices and word embedding models, were benchmarked with various clustering methods. Results revealed that clustering techniques applied to neural embedding feature representations outperform others across datasets.

Kim and Yoon[18] used taxi data to understand Seoul's regions. They used word2vec model to learn an embedding space of mobility data, where vectors in the embedding space represented urban zones. Then, they employed k-means to cluster similar zones and to form functional regions. To overcome the limitation of Place2vec model, Wang and Moosavi[19] suggested a new clustering mechanism called Place2vec that

[1]Department of Geospatial Information Systems, K. N. Toosi University of Technology, Tehran, Iran
[2]Department of Pathology, Microbiology, and Immunology, School of Veterinary Medicine, University of California, Davis, Davis, CA, USA
[3]Lead contact
*Correspondence: alesheikh@kntu.ac.ir (A.A.A.), alotfata@ucdavis.edu (A.L.)
https://doi.org/10.1016/j.isci.2024.109883

adjusts itself based on the density of places in a given context. The proposed model was trained on datasets from Yelp and OpenStreetMap, and the results showed significant improvements in performance compared to Place2vec. Dassereto et al.[20] introduced a method that captures geographical and hierarchical knowledge through embeddings. They argued that hyperbolic embeddings, particularly those utilizing the Poincaré disk model, are more suitable for this purpose than Euclidean embeddings. Hyperbolic embeddings maintain semantic relationships in fewer dimensions, making evaluations easier and reducing computational and memory demands.

Our research aligns more closely with studies[21–23] that specifically look at the semantic similarity of geospatial texts. Ballatore et al.[21] introduced the OSM Semantic Network through web crawling of the OSM Wiki website, enabling the computation of semantic similarity using co-citation measures. They developed a semantic tool and evaluated the cognitive plausibility of co-citation algorithms for computing semantic similarity in the context of crowdsourced geographic concepts, with SimRank algorithm showing the highest plausibility. Ballatore et al.[22] developed a knowledge-based approach to quantify the semantic similarity of lexical definitions in volunteered geographic information contexts, where definitions can be inconsistent and idiosyncratic. Grounded in the idea that similar geospatial objects are described using similar terms, the approach employs paraphrase-detection techniques and the lexical database WordNet. They evaluated the cognitive plausibility of this approach within the OSM Semantic Network, demonstrating high correlation with human judgments and providing practical guidelines for its usage in GIScience applications. Mai et al.[23] addressed the limitation of traditional topographic and thematic maps in directly expressing non-spatial relationships, such as semantic similarity among features. They proposed a semantically enriched geospatial data visualization and searching framework, employing techniques such as paragraph vector and clustering to produce a semantic distribution map of geographic features. They also developed an information retrieval model based on vector embedding, visualizing the results using both the semantic similarity-based map and a regular map to enhance users' understanding of latent relationships between seemingly unrelated geographic features. Recent studies have commonly employed embedding spaces using traditional methods such as doc2vec and then have applied a semantic similarity metric, such as Jaccard similarity, to the generated space.[24] Han et al.[25] evaluated different methods of determining semantic similarity and concluded that deep learning methods perform better than simple approaches such as doc2vec on short texts. Summa et al.[26] have acknowledged that most of the studies in NLP have only focused on textual content and ignored its spatiotemporal features. Likewise, Wang et al.[27] identified that the traditional models often fail in cases involving words with a geographical context. To overcome this, the authors proposed a method that measures the spatial semantic similarity using a sliding context window for geo-tagged photos. Ma et al.[28] presented a deep learning algorithm that can be used to match texts with their corresponding spatial objects. They applied their method to geological texts and were able to achieve better accuracy than conventional methods such as TF-IDF and doc2vec.

While the research mentioned earlier attempt to relate the semantics and geography of texts, they mostly fail to quantify document similarity based on both semantics and location. In this paper, we propose a method for determining the semantic similarity of texts in the geospatial context and attempt to distinguish between semantic similarity and spatial semantics. We specifically measure the similarity of texts based not only on their semantics, but also on their geographical location. By incorporating both semantic and geographical context, our approach addresses certain limitations inherent in current methodologies and aligns more closely with user expectations and behaviors in various practical applications involving textual content. Relying solely on semantic similarity fails to consider the geographical context. Despite having high semantic similarity, two documents may pertain to locations that are widely separated geographically. Furthermore, in numerous applications, particularly those related to location-based services such as location-based recommender systems, users anticipate results that not only exhibit semantic similarity but also manifest geographical proximity. For instance, in a real estate search, users seek property listings that align with their criteria and are situated in close proximity to their specified area of interest. Existing systems typically handle these two aspects independently: semantic similarity computations leverage texts and other descriptive fields of properties, while geographical proximity computations rely on explicit geographical coordinates. In contrast, our proposed methodology integrates geospatial criteria with semantic similarity computations to offer a more comprehensive and user-centric solution. In addition, while prior studies produced good accuracy when the datasets contained a variety of topics and issues, our suggested method is capable of computing the similarity of texts in a narrow domain with acceptable accuracy. Our research aims to merge semantics and geography for enhanced textual similarity. While this approach could benefit location-based recommendation systems, our primary focus is not on developing such a recommender system.

RESULTS

Classifying documents into meaningful topics

The proposed method was applied to both Divar and Wikipedia datasets. In the pre-processing step, all stop words were removed from the datasets. Then, the topic modeling methods were applied to each dataset to provide a list of topics, each containing three words. The number of topics were determined manually and based on their interpretability. Considering the diversity of the domains of each datasets, 10 topics were extracted for the Wikipedia articles and four topics were extracted for the Divar advertisements. The topics are shown in Figures 1 and 2.

As shown in Figures 1 and 2, the documents are intuitively grouped into the topics and, promisingly, can be used to be evaluated by real users in the evaluation step.

Identifying geographically indicative topics

In order to determine the most geographically indicative topics, Moran's I indices were calculated. According to our definition of geographic indicativeness, the geographically indicative topics tend to be more clustered. The probabilities provided by topic modeling were introduced

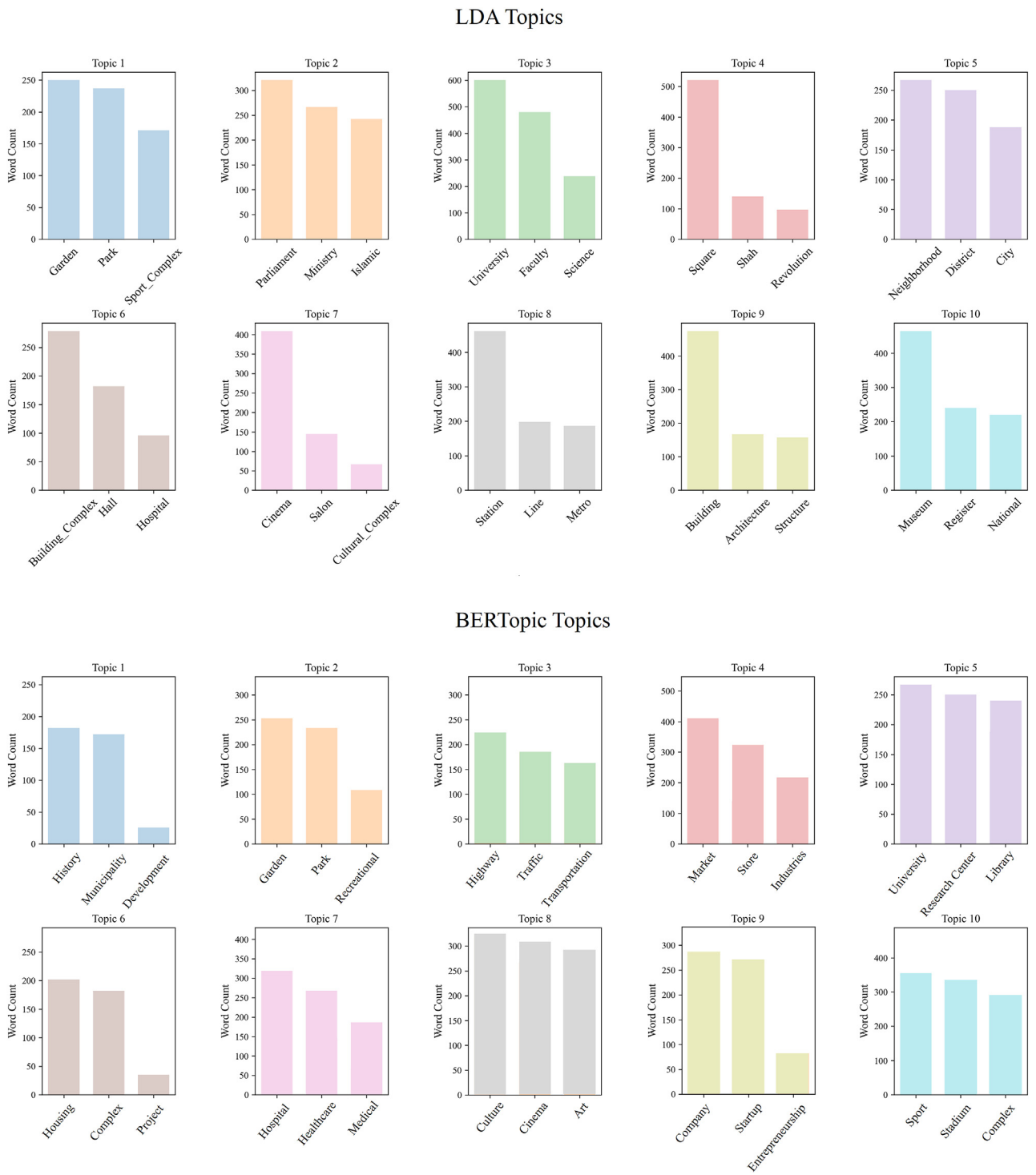


Figure 1. The topics found by the topic modeling for the Wikipedia dataset

as attributes in calculating Moran's I. Figures 3 and 4 illustrate the geographical distribution of the identified topics using latent Dirichlet allocation (LDA). For the Wikipedia articles, topics 10 and 2 are the most geographically indicative topics with the Moran's I values of 0.45 and 0.42, respectively. On the other hand, topic 5 is the least geographically indicative topic (Moran's I = 0.23). For the Divar dataset, the scenario is, to some extent, different. The topics are more dispersed than that of the Wikipedia articles, and the values of Moran's I are

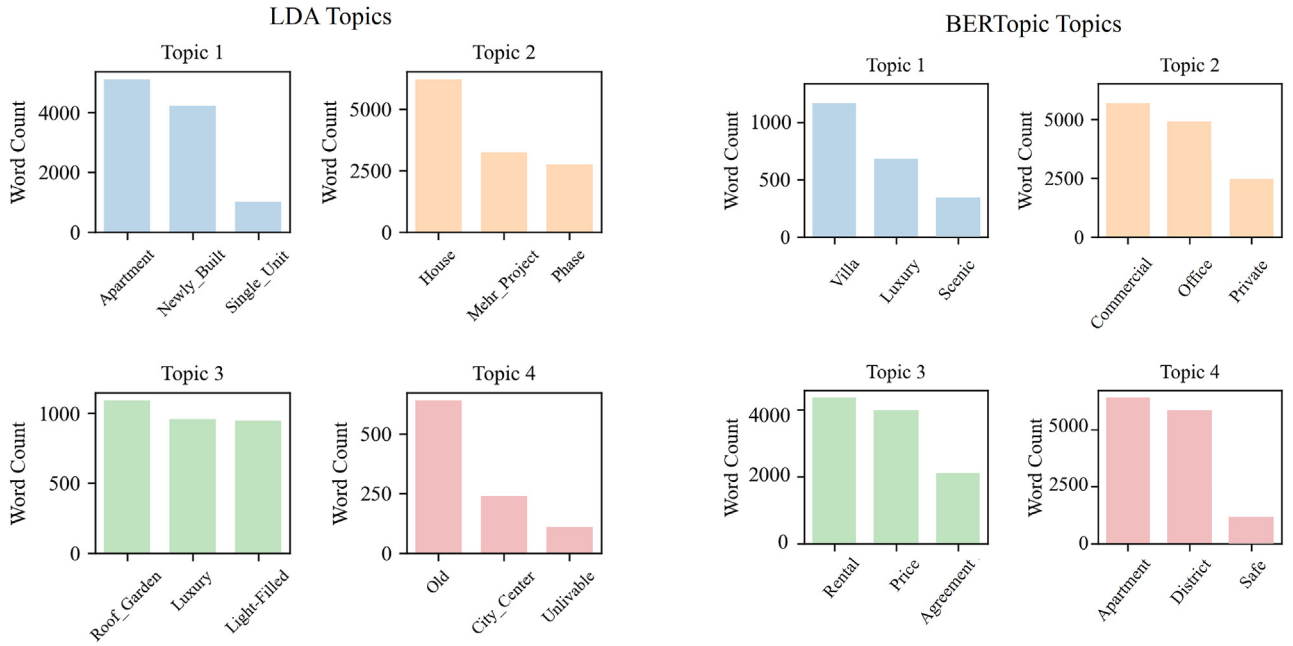


Figure 2. The topics found by the topic modeling for the Divar dataset

generally low. Topic 3 is the most geographically indicative topic by an index of 0.19. The other three topics are very dispersed over the study area and therefore cannot be considered as geographically indicative.

Figures 5 and 6 depict the geographical distribution of the identified topics using BERTopic. In the case of using BERTopic, topic 4 is the most geographically indicative topic in the Wikipedia dataset (Moran's I = 0.43). The least geographically indicative topic is topic 3 (Moran's I = 0.01). As can be seen in the figures, topic 4 comprises points that are located in a specific region in the study area. On the other hand, topic 3 is composed of transportation-related documents. The topic is not specific to a particular region and is distributed over the study area.

As a by-product of our proposed method, the topics can also be represented in an embedding space. By multiplying the documents' embedding (**M**) by the documents' topic distribution matrix (**T**), we get an embedding space for topics.

$$\text{topic2vec} = T^T \times M$$

Figure 7 illustrates the heatmap representation of the similarities between different topics extracted by LDA in the Wikipedia articles, found by cosine similarity applied on the topic2vec embedding. The similarity values make sense in many aspects. For instance, the topic consisting of words *station*, *line*, and *metro* is most similar to the topic comprised of words *neighborhood*, *district*, and *city*, which is intuitive as public transport is semantically very close to the regions of the city.

The correlation between semantic similarity and geographical distance

By multiplying the embedding by the Moran's I values, the fine-tuned embedding is yielded. Then, we computed the correlation between the geographical distances of documents and their corresponding semantic similarity. It was demonstrated earlier that employing standard semantic similarity measures in isolation yields virtually no correlation with geographic distance, indicating the independence of the two factors. Through an analysis of correlation before and after the fine-tuning process, the efficacy of the method in bridging semantics and geography is quantified. A high correlation indicates that documents located in close geographical proximity have been positioned closer to each other in the embedding space. The correlation values are shown in Table 1. The results show that the fine-tuning of the model has been able to significantly improve the correlation values. The proposed approach has had a greater impact on the Wikipedia dataset since its topics has been more geographically indicative than those of the Divar dataset. Conversely, fine-tuning with BERTopic modeling has led to stronger correlations in similarities. This is particularly noticeable in the Divar dataset with its distinct topics. The improved correlation indicates that the geospatial semantic measure gives higher similarity scores to text pairs close in both meaning and location, going beyond just lexical or topical resemblance.

We use the raw and fine-tuned models as part of a recommender algorithm to provide items (both rental properties and Wikipedia articles) for real users. The recommender algorithm employs a simple collaborative filtering approach where different users visit similar items based on the current item. The users were asked to rate how much they expected the recommendations to 10 given items, based on a one-to-ten scale. Twenty users participated in the evaluation process. Figure 8 demonstrates the average rating of different users for both datasets. In the case of the Wikipedia articles, the original model outperforms the fine-tuned models in four items. On the other hand, in the case of the Divar dataset, at least one fine-tuned model makes better recommendations to the users. Since the domain of the Wikipedia dataset is broad

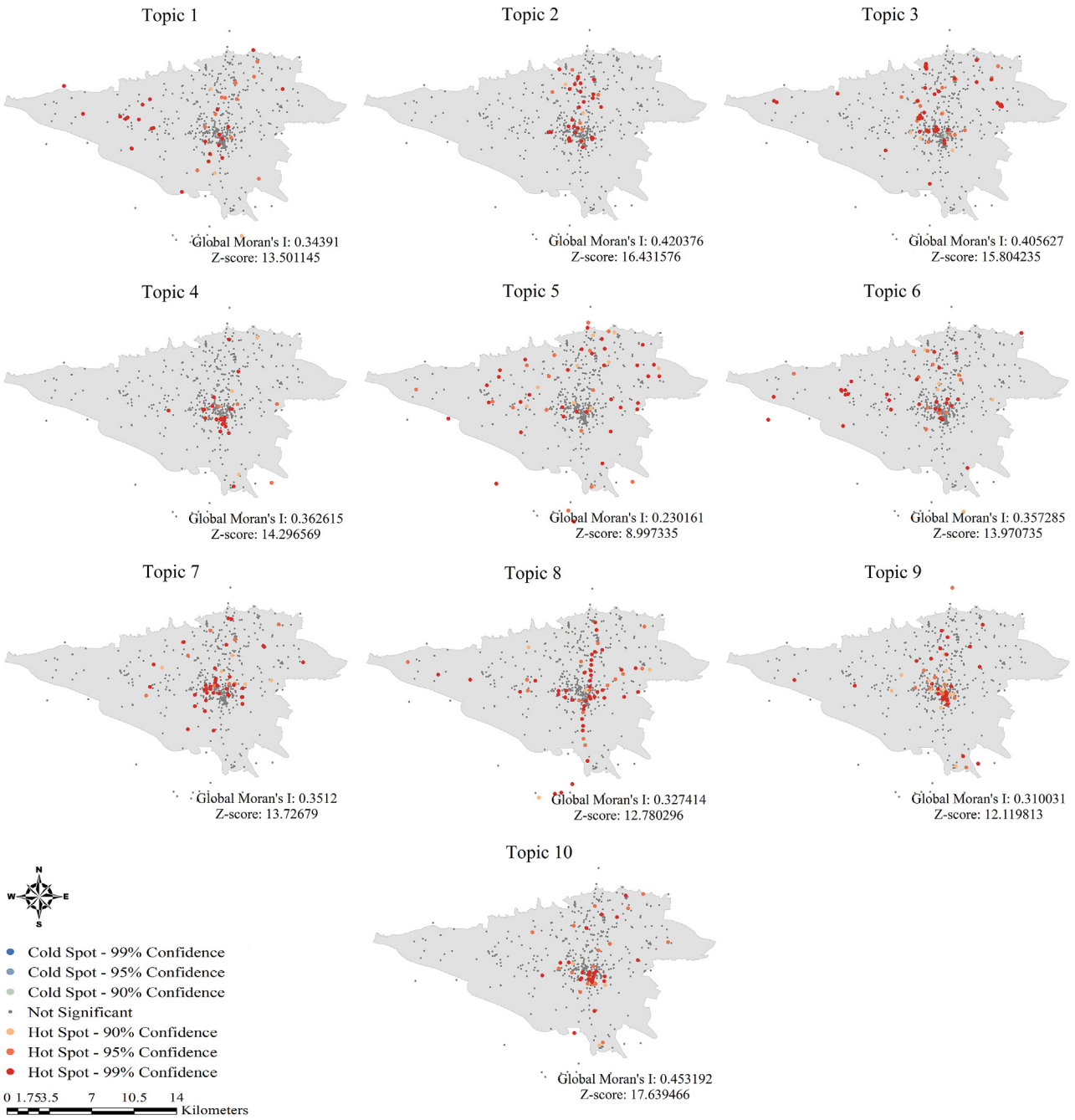


Figure 3. The geographical distribution of topics extracted by LDA in the Wikipedia dataset and their corresponding Moran's I index

in terms of the topics, the use of semantic similarity alone in a general recommender algorithm may satisfy the expectations of the users. When users encounter Wikipedia articles, they expect recommendations that semantically match the given item. In contrast, for rental property advertisements, people expect similar items located near the given item. The average accuracy of suggested Wikipedia items is 2.2% and 5.2% using BERTopic and LDA, respectively. Comparatively, the average accuracy of Divar items is 18.2% and 10.37% using BERTopic and LDA, respectively.

DISCUSSION

In this paper, we identified two problems that location-based recommenders might face when they use semantic similarity as a measure of recommendation. First, we show that the distances in the embedding space are not correlated to the geographical distance. This problem

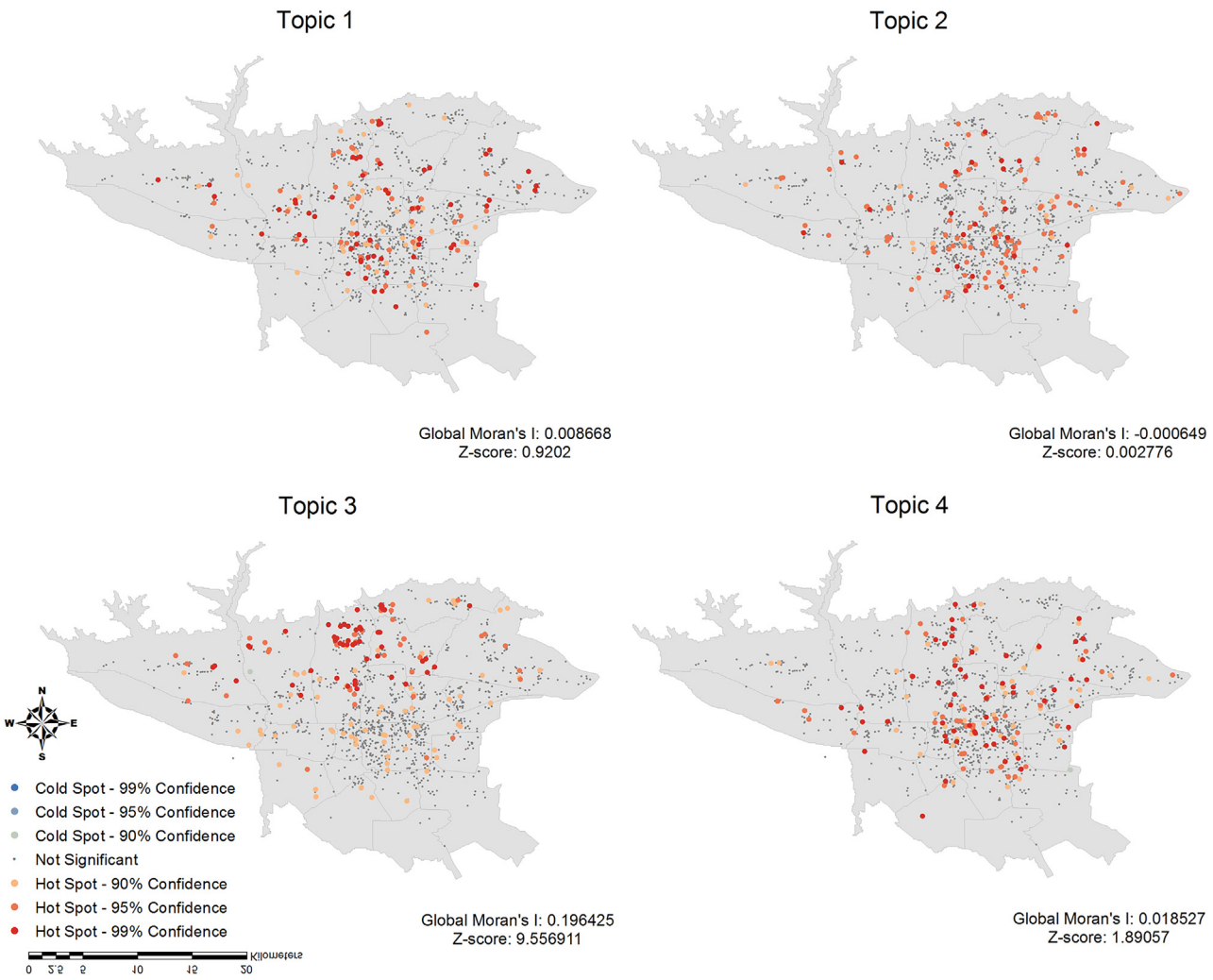


Figure 4. The geographical distribution of topics extracted by LDA in the Divar dataset and their corresponding Moran's I index

causes the recommended item to not be located in the proximity of their desired region. Then, we show that, due to the domain shift problem, the computed semantic similarities in the geospatial context are biased.

Given the identified problems, we proposed a method to measure the geospatial semantic similarity of texts. The method combines the Moran's I, as a measure of geographic indicativeness, with Sentence-BERT. We employed two datasets of different nature. A dataset consists of geo-located Persian Wikipedia articles on diverse topics, all related to the city of Tehran, Iran. The other dataset is extracted from Divar platform and contains more than 10,000 of real estate advertisements. Obviously, the diversity of the topics in this dataset is low, and the problem of domain shift appears vividly in this case. Therefore, whereas we identified 10 topics for the Wikipedia articles, only four meaningful topics were discovered from the advertisements. By applying the method on both datasets, the correlation between semantic similarity and geographical distance was computed. A high correlation is critical in applications such as location-based recommenders where the users intend to find semantically similar items near their desired item. The findings show that the proposed method can promisingly improve the correlation. Again, the amount of improvement is higher in the case of Wikipedia articles. This is in accordance with our expectations as the Wikipedia articles were more clustered than the rental property advertisements, leading to higher Moran's I values. Nevertheless, it should be noted that there was almost no correlation between semantic and geographical distances before fine-tuning.

Besides, the evaluation of the model was conducted by real users. The evaluation uncovers different patterns for each dataset. First, in the case of Wikipedia articles, the user ratings are rather high even when the original model is used. The cause of high ratings can be attributed to the fact that the users focus on the content of the articles in their judgments. For example, two Wikipedia articles about two different museums are considered very similar to each other regardless of their geographical distance. However, when users face a rental property advertisement, they consider both content, that is the characteristics of the property, and the location of the property. This is why the accuracy of the results of the model for the second dataset was twice as high.

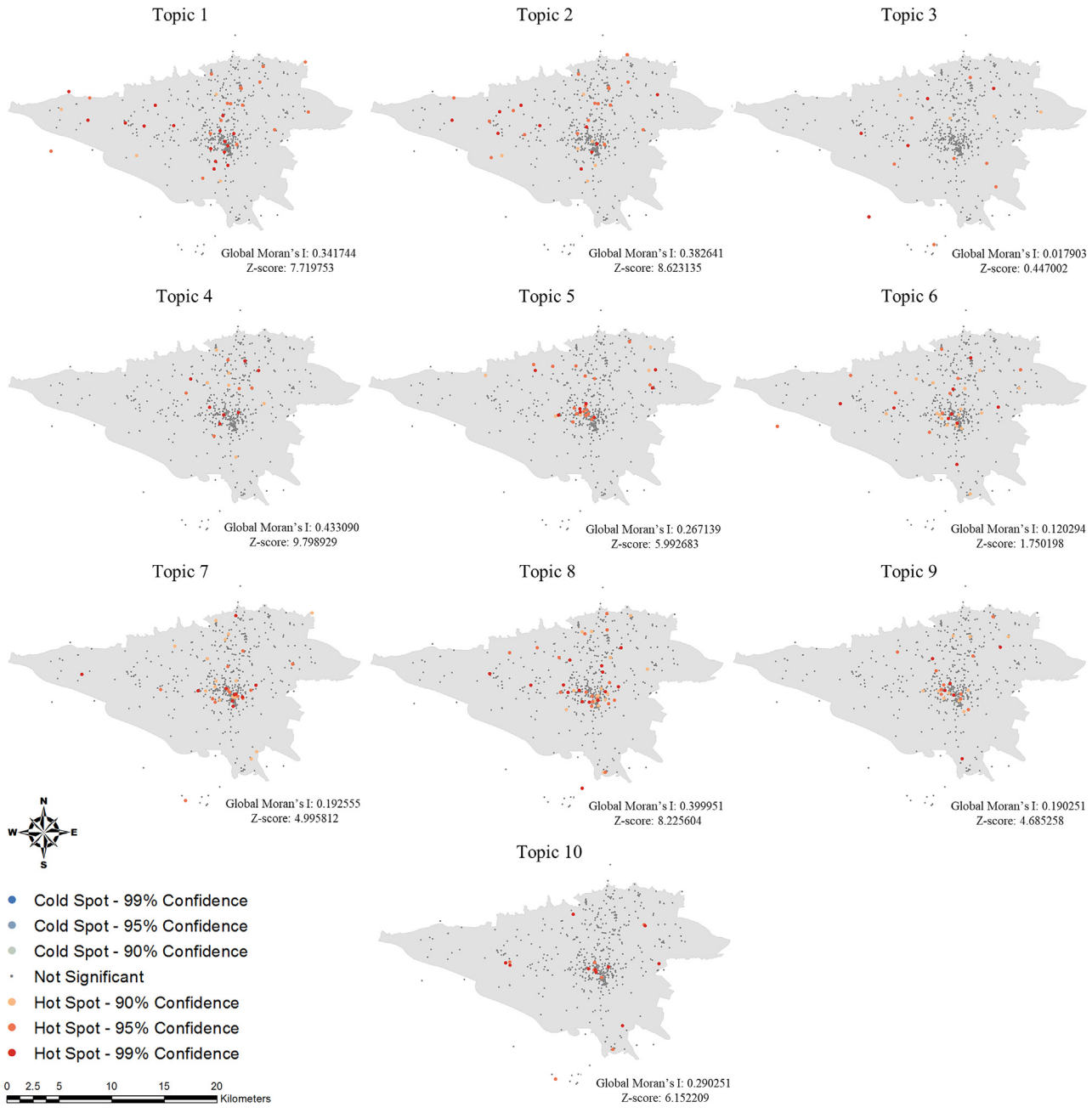


Figure 5. The geographical distribution of topics extracted by BERTopic in the Wikipedia dataset and their corresponding Moran's I index

The results found by Wang et al.[27] align with our results. They used a spatial context window to measure the spatial similarity of word pairs and compared it with the semantic similarity values. They estimated a Pearson correlation coefficient of 0.09. This estimation is compatible with our findings in Table 1 which shows that the relationship between semantic and spatial similarity is not linear. However, our proposed method was able to calculate the semantic similarity in a way that improved the correlation between the two to 0.32 in the case of Wikipedia data, and to 0.15 in the case of Divar dataset. It should be mentioned that they have used a dataset of Flickr tags resulting in texts with distinguishable and probably more clustered topics. In addition, Li et al.[29] evaluated the semantic similarity from the perspective of the first law of geography[30] and concluded that there was a distance at which near things were less related than distant things. This result is compatible with our correlation analysis.

Beyond its use in location-based recommender systems, our suggested approach can aid in clustering and classifying textual data, benefiting areas like marketing and social media analysis. Future research could explore the potential of incorporating other factors such as cultural differences or language barriers into the model to further improve its accuracy and applicability in diverse contexts. Another

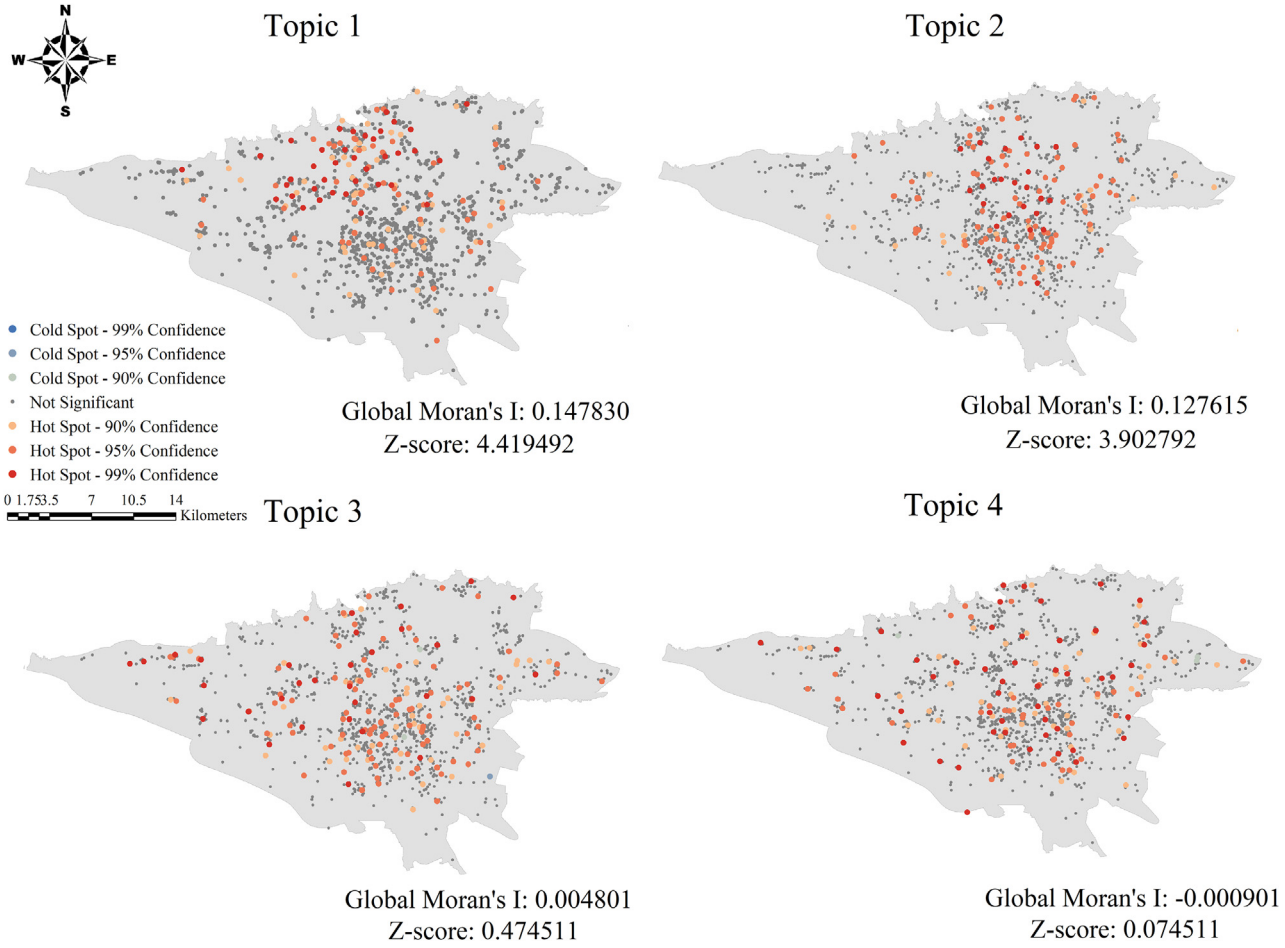


Figure 6. The geographical distribution of topics extracted by BERTopic in the Divar dataset and their corresponding Moran's I index

direction for future research could be to investigate the impact of different embedding methods on the relationship between semantic similarity and geographical distance and identify which methods work best for different types of texts and datasets. Finally, this approach could be extended to other modalities such as images or audio, where the concept of spatial distance can still be relevant but may need to be defined differently. This could open up new avenues for research in multimodal NLP.

Conclusions

In this paper, we developed a method to measure the geospatial semantic similarity of texts. The method is based on the combination of the BERT architecture and Moran's I. The method classifies each document into topics using a topic modeling approach. The findings show that the proposed method can improve the correlation between semantic similarity and geographical distance. Our study deepens our understanding of the differences between semantic similarity and geographical distance and how the users perceive them. The proposed method can be used in location-based recommender systems where users expect similar items near their desired locations.

Limitations of the study

This study has a few limitations. First, pre-trained language models for Persian text are rare and a few options are available. Therefore, the authors had to choose among a few options. On the other hand, crawling text from web is time-consuming. Specifically, in this study, we needed georeferenced text to evaluate the results. In addition, in order to illustrate the effect of domain's diversity on the semantic similarity computations, two datasets with different topic's diversity were required.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE

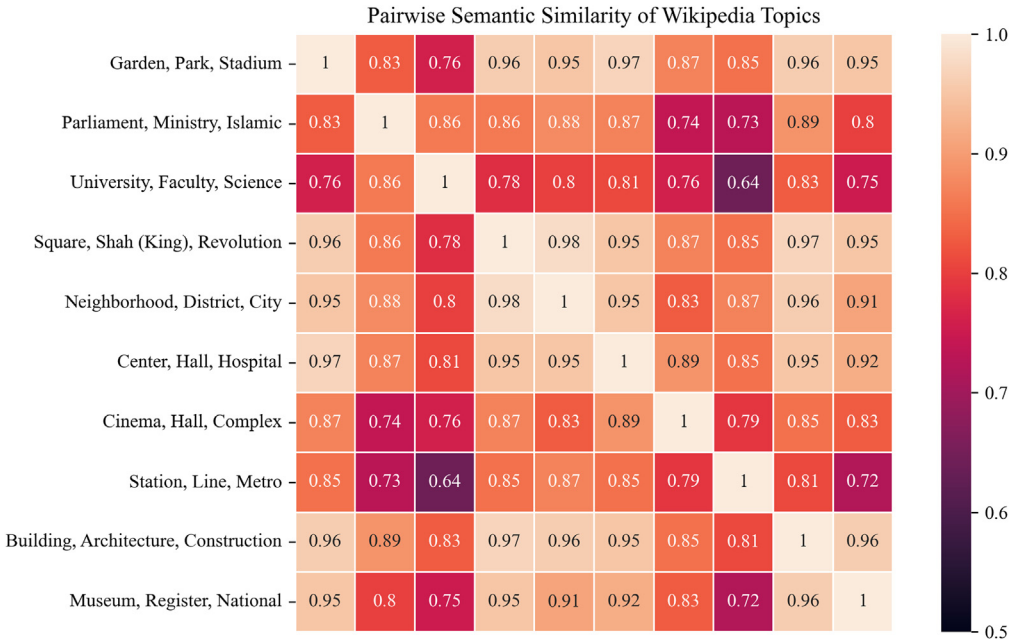


Figure 7. The pairwise semantic similarity between the Wikipedia articles in our dataset

- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- METHOD DETAILS
  - Dataset description
  - NLP problems in the geospatial context
  - Domain shift problem
  - Semantic similarity is not enough
  - Proposed method
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2024.109883.

ACKNOWLEDGMENTS

We thank the editor and the anonymous reviewers for their excellent suggestions to improve the original draft of the manuscript.

AUTHOR CONTRIBUTIONS

O.R.A. performed the computations and analysis, drafted the manuscript, and designed the figures. A.A.A. has made contributions to interpreting the results and revising the final manuscript. A.L. has contributed to original draft reviewing and editing.

Table 1. The correlation between geographical distance and semantic similarity before and after the fine-tuning

		Correlation after fine-tuning	
	Correlation before fine-tuning	LDA	BERTopic
Wikipedia	−0.063	0.324	0.338
Divar	0.015	0.127	0.166

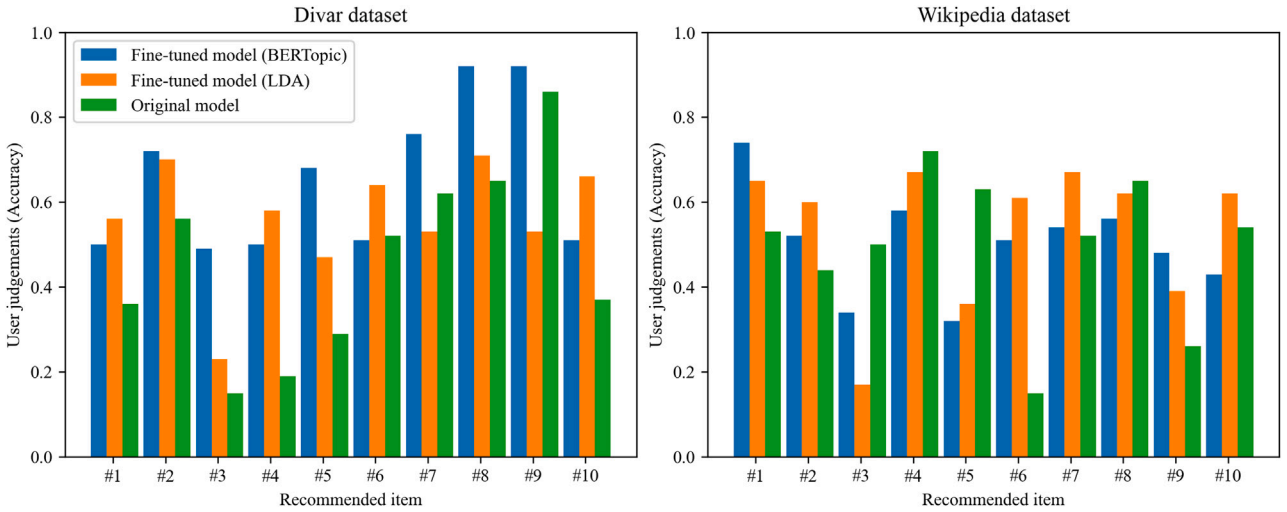


Figure 8. Evaluation of recommendations based on user judgments

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: June 4, 2023
Revised: January 20, 2024
Accepted: April 30, 2024
Published: May 3, 2024

REFERENCES

1. Khurana, D., Koli, A., Khatter, K., and Singh, S. (2023). Natural language processing: State of the art, current trends and challenges. Multimed. Tools Appl. *82*, 3713–3744.
2. Nayak, S., Kanetkar, A., Hirudkar, H., Ghotkar, A., Sonawane, S., and Litake, O. (2022). Suggesting Relevant Questions for a Query Using Statistical Natural Language Processing Technique. Preprint at arXiv. https://doi.org/10.48550/arXiv.2204.12069.
3. Yang, J., Li, Y., Gao, C., and Zhang, Y. (2021). Measuring the short text similarity based on semantic and syntactic information. Future Gener. Comput. Syst. *114*, 169–180.
4. Oussalah, M., and Mohamed, M. (2022). Knowledge-based sentence semantic similarity: algebraical properties. Prog. Artif. Intell. *11*, 43–63.
5. Chandrasekaran, D., and Mago, V. (2021). Evolution of semantic similarity—a survey. ACM Comput. Surv. *54*, 1–37.
6. Nguyen, H.T., Duong, P.H., and Cambria, E. (2019). Learning short-text semantic similarity with word embeddings and external knowledge sources. Knowl. Based. Syst. *182*, 104842.
7. Wei, C., Wang, B., and Kuo, C.-C.J. (2022). Task-specific dependency-based word embedding methods. Pattern Recognit. Lett. *159*, 174–180.
8. Deb, S., and Chanda, A.K. (2022). Comparative analysis of contextual and context-free embeddings in disaster prediction from Twitter data. Mach. Learn. Appl. *7*, 100253.
9. Peinelt, N., Nguyen, D., and Liakata, M. (2020). tBERT: Topic models and BERT joining forces for semantic similarity detection. In Proceedings of the 58th annual meeting of the association for computational linguistics, pp. 7047–7055.
10. Deng, W., Zheng, L., Sun, Y., and Jiao, J. (2021). Rethinking triplet loss for domain adaptation. IEEE Trans. Circuits Syst. Video Technol. *31*, 29–37.
11. Pedersen, T., Pakhomov, S.V.S., Patwardhan, S., and Chute, C.G. (2007). Measures of semantic similarity and relatedness in the biomedical domain. J. Biomed. Inform. *40*, 288–299.
12. Garla, V.N., and Brandt, C. (2012). Semantic similarity in the biomedical domain: an evaluation across knowledge sources. BMC Bioinform *13*, 261.
13. Harispe, S., Sánchez, D., Ranwez, S., Janaqi, S., and Montmain, J. (2014). A framework for unifying ontology-based semantic similarity measures: A study in the biomedical domain. J. Biomed. Inform. *48*, 38–53.
14. Sousa, R.T., Silva, S., and Pesquita, C. (2020). Evolving knowledge graph similarity for supervised learning in complex biomedical domains. BMC Bioinform *21*, 6.
15. Wilcox, C., Djahel, S., Giagos, V., Welsh, K., and Costen, N. (2023). A New Semantic Similarity Scheme for more Accurate Identification in Medical Data. In 2023 IEEE International Smart Cities Conference (ISC2), pp. 1–7.
16. Hendre, M., Mukherjee, P., Preet, R., and Godse, M. (2020). Efficacy of deep neural embeddings based semantic similarity in automatic essay evaluation. Int. J. Comput. Digit. Syst. *9*, 1–11.
17. Curiskis, S.A., Drake, B., Osborn, T.R., and Kennedy, P.J. (2020). An evaluation of document clustering and topic modelling in two online social networks: Twitter and Reddit. Inf. Process. Manag. *57*, 102034.
18. Kim, N.W., and Yoon, Y. (2021). Representation learning of urban regions via mobility-signature-based zone embedding: A case study of Seoul, South Korea. In Proceedings of the 5th ACM SIGSPATIAL International Workshop on Location-based Recommendations, Geosocial Networks and Geoadvertising, pp. 1–4.
19. Wang, Z., and Moosavi, V. (2020). From Place2Vec to Multi-Scale Built-Environment Representation: A General-Purpose Distributional Embedding for Urban Data Analysis. In Proceedings of the 4th ACM SIGSPATIAL Workshop on Location-Based Recommendations, Geosocial Networks, and Geoadvertising, pp. 1–12.
20. Dassereto, F., Di Rocco, L., Guerrini, G., and Bertolotto, M. (2020). Evaluating the effectiveness of embeddings in representing the structure of geospatial ontologies. In Geospatial Technologies for Local and Regional Development: Proceedings of the 22nd AGILE Conference on Geographic Information Science, pp. 41–57.
21. Ballatore, A., Bertolotto, M., and Wilson, D.C. (2013). Geographic knowledge extraction and semantic similarity in OpenStreetMap. Knowl. Inf. Syst. *37*, 61–81.

22. Ballatore, A., Wilson, D.C., and Bertolotto, M. (2013). Computing the semantic similarity of geographic terms using volunteered lexical definitions. Int. J. Geogr. Inf. Sci. *27*, 2099–2118.
23. Mai, G., Janowicz, K., Prasad, S., and Yan, B. (2018). Visualizing the semantic similarity of geographic features. In Proceedings of the Conference: AGILE 2018, pp. 12–15.
24. Ajao, O., Bhowmik, D., and Zargari, S. (2018). Content-aware tweet location inference using quadtree spatial partitioning and jaccard-cosine word embedding. In 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), pp. 1116–1123.
25. Han, M., Zhang, X., Yuan, X., Jiang, J., Yun, W., and Gao, C. (2021). A survey on the techniques, applications, and performance of short text semantic similarity. Concurr. Comput. *33*, e5971.
26. Summa, A., Resch, B., and Strube, M. (2016). Microblog emotion classification by computing similarity in text, time, and space. In Proceedings of the workshop on computational modeling of people's opinions, personality, and emotions in social media (PEOPLES), pp. 153–162.
27. Wang, B., Fei, T., Kang, Y., Li, M., Du, Q., Han, M., and Dong, N. (2020). Understanding the spatial dimension of natural language by measuring the spatial semantic similarity of words through a scalable geospatial context window. PLoS One *15*, e0236347.
28. Ma, K., Wu, L., Tao, L., Li, W., and Xie, Z. (2018). Matching descriptions to spatial entities using a Siamese hierarchical attention network. IEEE Access *6*, 28064–28072.
29. Li, T.J.-J., Sen, S., and Hecht, B. (2014). Leveraging advances in natural language processing to better understand Tobler's first law of geography. In Proceedings of the 22nd ACM SIGSPATIAL international conference on advances in geographic information systems, pp. 513–516.
30. Tobler, W.R. (1970). A computer movie simulating urban growth in the Detroit region. Econ. Geogr. *46*, 234–240.
31. Reimers, N., and Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. Preprint at arXiv. https://doi.org/10.48550/arXiv.1908.10084.
32. Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. Preprint at arXiv. https://doi.org/10.48550/arXiv.2203.05794.
33. Divar Platform. https://divar.com.
34. Kouw, W.M., and Loog, M. (2018). An introduction to domain adaptation and transfer learning. Preprint at arXiv. https://doi.org/10.48550/arXiv.1812.11806.
35. Sun, S., Shi, H., and Wu, Y. (2015). A survey of multi-source domain adaptation. Inf. Fusion *24*, 84–92.
36. Hu, Y. (2018). Geo-text data and data-driven geospatial semantics. Geogr. Comp. *12*, e12404.
37. HuggingFace. https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2.
38. Farahani, M., Gharachorloo, M., Farahani, M., and Manthouri, M. (2021). Parsbert: Transformer-based model for persian language understanding. Neural Process. Lett. *53*, 3831–3847.
39. Liao, W., Zhang, Q., Yuan, B., Zhang, G., and Lu, J. (2022). Heterogeneous multidomain recommender system through adversarial learning. IEEE Trans. Neural Netw. Learn.
40. Vandecasteele, A., and Devillers, R. (2015). Improving volunteered geographic information quality using a tag recommender system: the case of OpenStreetMap. In OpenStreetMap in GIScience (Springer), pp. 59–80.
41. Tobler, W. (2004). On the first law of geography: A reply. Ann. Assoc. Am. Geogr. *94*, 304–310.
42. Vayansky, I., and Kumar, S.A. (2020). A review of topic modeling methods. Inf. Syst. *94*, 101582.
43. Moran, P.A.P. (1948). The interpretation of statistical maps. J. Roy. Stat. Soc. B *10*, 243–251.
44. Lee, J., and Li, S. (2017). Extending moran's index for measuring spatiotemporal clustering of geographic events. Geogr. Anal. *49*, 36–57.
45. Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., and Zhao, L. (2019). Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. Multimed. Tools Appl. *78*, 15169–15211.
46. Abbasi, O.R., and Alesheikh, A.A. (2023). A Place Recommendation Approach Using Word Embeddings in Conceptual Spaces. IEEE Access *11*, 11871–11879.
47. Farahani, A., Voghoei, S., Rasheed, K., and Arabnia, H.R. (2021). A brief review of domain adaptation. In Advances in Data Science and Information Engineering: Proceedings from ICDATA 2020 and IKE 2020, pp. 877–894.
48. Xu, H., Ebner, S., Yarmohammadi, M., White, A.S., Van Durme, B., and Murray, K. (2021). Gradual fine-tuning for low-resource domain adaptation. Preprint at arXiv. https://doi.org/10.48550/arXiv.2103.02205.
49. Alam, F., Joty, S., and Imran, M. (2018). Domain adaptation with adversarial training and graph embeddings. Preprint at arXiv. https://doi.org/10.48550/arXiv.1805.05151.
50. Sun, F., Wu, H., Luo, Z., Gu, W., Yan, Y., and Du, Q. (2019). Informative feature selection for domain adaptation. IEEE Access *7*, 142551–142563.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Software and algorithms		
BERT	Devlin et al.[51]	https://arxiv.org/abs/1810.04805v2
Sentence-BERT	Reimers et al.[31]	https://ar5iv.labs.arxiv.org/html/1908.10084
BERTopic	Grootendorst[32]	https://maartengr.github.io/BERTopic/index.html
Python version 3.8	Python Software Foundation	https://www.python.org

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Ali Asghar Alesheikh (alesheikh@kntu.ac.ir).

Materials availability

Any additional information related to materials generated in this study is available from the lead contact upon request.

Data and code availability

- Some data reported in this study cannot be deposited in a public repository due to confidentiality reasons, which are mandatory according to the Ethical Committee. However, they might be available upon request to the lead contact.
- This paper does not report original code.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

METHOD DETAILS

Dataset description

We employed a dataset of over 700 Persian Wikipedia articles and a dataset of over 10,000 real estate advertisements from Tehran, Iran. The Wikipedia dataset contains the whole text of articles and the location information. It includes diverse topics from museums and cultural sites to metro stations, governmental organizations, and universities. The dataset was collected using SPARQL Protocol and RDF Query Language (SPARQL) through the Wikidata Query Service. The advertisement dataset includes text and location details, focusing exclusively on real estate and buildings. The dataset were crawled for six months, from September 2019 to February 2020, from a popular platform in Iran, called Divar.[33] Table S1 shows information about the datasets used in this study.

NLP problems in the geospatial context

In this section, we demonstrate empirically that current approaches for evaluating semantic similarity confront two challenges in geographical applications. First, we demonstrate that pre-trained models cannot be applied directly to the geographical domain. Then, we show that in some applications, such as location-based recommender systems, semantic similarity alone is insufficient as a criterion for recommending and fails to meet users' expectations.

Domain shift problem

A primary challenge for transfer learning algorithms is adapting to diverse domains.[34] This challenge arises when an algorithm trained on a general dataset needs to be applied to a dataset from a different domain with a distinct distribution.[35] Since most models in NLP have been trained using datasets with general contexts,[36] they have to be adapted to geospatial context in location-based applications. To illustrate the domain shift problem in geospatial context, we compute the paired semantic similarity on both datasets using the Sentence-BERT method.[31] The details and the architecture of BERT and Sentence-BERT is described in Figures S1 and S2, respectively. The Sentence-BERT algorithm was employed using three distinct models: a general pre-trained model for English known as "all-MiniLM-L6-v2", a pre-trained multilingual model known as "paraphrase-multilingual-mpnet-base-v2",[37] and a monolingual model specifically trained for the Persian language called ParsBERT.[38] Preliminary findings indicated that ParsBERT generally yields higher-quality results in discerning semantic similarity. Table S2 provides an example in both English and Persian, demonstrating that the monolingual model produces more intuitive results. The comparison involved two sentences,

"The house is lush with greenery" (in Persian: "خانه سرسبز است.") and "The house is in a garden" (in Persian: "خانه در باغ قرار دارد."), utilizing the three models.

To demonstrate the difficulty of transitioning from a general domain to the geospatial context, 100 texts from each dataset were chosen at random and their pairwise semantic similarity was determined. Figure S3 depicts the heat map corresponding to the similarity values. The results show that for the dataset of Wikipedia articles, which contains more diverse topics than real estate advertisements, the similarity values are also more diverse. On the contrary, for real estate advertisements, the amount of similarity obtained is very high. Therefore, as the target domain's range becomes narrower, the problem of domain shift becomes more apparent. The mean and standard deviation of the similarity values for the Wikipedia dataset are equal to 0.53 and 0.48, and for the Divar dataset are equal to 0.82 and 0.06, respectively.

Semantic similarity is not enough

Aside from domain shift, semantic similarity in geographical applications confronts an additional obstacle. This issue manifests itself in applications such as recommender systems.[39] In non-spatial applications, since users are merely looking for semantically similar content, a text-based recommender system suggests items to users based on semantic similarity. However, in geospatial applications, the user not only is looking for similar content, but also intends to find items nearby.[40] According to Tobler's first law of geography,[41] "all items are related, but closer items are more related." Accordingly, the geographical distances of all pairs of items in both datasets were calculated and their correlation with their corresponding semantic similarity values were determined. For the Wikipedia dataset, the correlation is equal to -0.063, and for the Divar dataset, it is equal to 0.015. These results show that, regardless of the domain's range, the correlation of semantic similarity with geographical distance is close to zero. In other words, there is no significant relationship between the computed similarity values and geographical distance.

Proposed method

In the previous section, we identified that the semantic similarities found in the textual contents do not necessarily correlate with geographical proximity. Moreover, we showed that in certain contexts the computed semantic similarities are higher than expected due to the domain shift problem. In this section we propose a method to overcome domain shift and the incompatibility of semantic similarity with geographical distance. Figure S4 shows the workflow of the proposed method.

To solve the problem of the lack of correlation between semantic similarity and geographical distance, a new method based on topic modeling[42] and Moran's I measure[43,44] is proposed. In this paper, we employ two commonly used topic modeling approaches, namely LDA and BERTopic. LDA, as an unsupervised method, categorizes documents into topics, each represented by a collection of words $\mathbf{w}$. Assume that a document set consists of k topics and contains V words, the probability of assigning each topic to a document D with N words is calculated as.[45]

$$p(\mathbf{w}|\alpha,\beta) = \int_{\theta} p(\theta|\alpha)\left(\prod_{n=1}^{N}\sum_{z_n=1}^{k} p(w_n|z_n,\beta)p(z_n|\theta)\right)d\theta,$$

where $\mathbf{w}$ is the set of words in document D, and θ is a k-dimensional random variable selected from a Dirichlet distribution. Also, α and β are parameters that control the multinomial distribution of topics in the entire set of documents and the multinomial distribution of words in each topic, respectively.

The BERTopic algorithm is a novel approach to topic modeling that leverages the power of BERT model.[32] Unlike traditional topic modeling methods such as LDA, BERTopic utilizes contextualized word embeddings generated by pre-trained BERT models to capture the meanings of words in context. The algorithm follows a multi-step process: it first embeds the documents into a high-dimensional vector space using BERT embeddings (Figure S1). Next, it employs clustering techniques to group similar documents together, forming clusters that represent distinct topics. Finally, representative words or phrases are extracted from each cluster to provide a coherent and interpretable representation of the topics. One of the key advantages of BERTopic is its ability to capture semantic relationships and context, making it particularly effective in tasks such as document clustering, summarization, and topic extraction. This algorithm has found applications across various domains where understanding the context and nuances of language is crucial for accurate analysis and interpretation.

After obtaining the topics, it should be determined which topics are more geographically indicative. Naturally, topics that are less scattered in the study area refer to more specific regions.[46] We use Moran's I to identify the geographically indicative topics. In the realm of spatial analysis, Moran's I stands as a conventional metric for gauging spatial autocorrelation, enabling the quantitative assessment of spatial clustering in accordance with the first law of geography. The numerical value derived from Moran's I serves as a robust indicator of geographic dispersion, facilitating the identification of geographically indicative topics and documents. We utilize Moran's I as a metric that operates on the premise that topics exhibiting a higher degree of spatial clustering inherently possess a greater degree of geographic indicativeness. As a result, the values derived from Moran's I serve as weights to enhance the embeddings, leading to more contextually relevant depiction of spatial relationships in the data. Moran's I is computed as.

$$I = \frac{n \sum_{i=1}^{n} \sum_{j=1}^{n} w_{i,j}(x_i - \overline{X})(x_j - \overline{X})}{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{i,j} \sum_{i=1}^{n} (x_i - \overline{X})^2},$$

where x_i is the topic probability value for document i calculated from LDA, and $w_{i,j}$ is the spatial weight between the points representing documents i and j.

In NLP, domain adaptation seeks to improve a model's performance when applied to data from a domain different from its training set. This technique involves adjusting the model's parameters or features to better fit the new domain while preserving its ability to generalize to other domains.[47] Domain adaptation can be achieved through various methods such as fine-tuning,[48] adversarial training,[49] and feature selection.[50] The goal is to make the model more robust and accurate in handling data from different domains. Accordingly, in this study, the Moran's I values is used as weights in the domain adaptation approach. To apply the weights vector to the embedding, we simply multiply the embedding by the weights vector. It scales each dimension of the embedding by the corresponding weight value, effectively giving more importance to certain features and facilitating the discovery of geospatially more related items for the recommendation algorithm.

QUANTIFICATION AND STATISTICAL ANALYSIS

No statistical analysis was employed in this study.