CEUR-WS.org/Vol-3385/paper1.pdf

Evaluating the Similarity of Location-based Corpora Identified in Reddit Comments

Cillian Berragan[1], Alex Singleton[1], Alessia Calafiore[2] and Jeremy Morley[3]

[1]University of Liverpool, Liverpool, L69 3BX, United Kingdom
[2]University of Edinburgh, Edinburgh, EH8 9YL, United Kingdom
[3]Ordnance Survey, Southampton, SO16 0AS, United Kingdom

Abstract

Social interaction is typically studied from the context of physical movement, where geographic distance and ease of connectivity influence the strength of interaction between regions. From the point of view of social media networks however, these limitations appear to still persist, despite interactions not being reliant on physical movement, suggesting non-physical geographic characteristics influence interaction between social communities. Unlike geotags, which provide explicit geographic information about social media users as coordinates, unstructured text presents an alternative perspective for the study of social interaction between regions, instead allowing for the comparison between the language used when mentioning locations in context. Our paper analyses the corpora associated with major cities across the UK, first vectorising Reddit comments through transformer-based embeddings, which capture semantic information, then using these to establish unsupervised clusters and similarity between them. We find that distinct groups emerge which broadly conform with established regional identities of locations across the UK, but with interesting deviations.

Keywords

Social media, Natural Language Processing, Social Interaction

1. Introduction

Social interaction is typically studied in the context of mobility, using data sources like Census or transport records, where physical movement is restricted by distance and ease of connectivity between two locations [1, 2]. In contrast to this, social interaction has also been studied using phone call data [3], and social media networks [4], where the spatial and temporal bounds of connectivity between two locations does not restrict interactions. Despite this however, many studies have found that geographic identities within communities still persist in these networks, with interaction strength influenced by the geographic distance between them [5, 6].

Social media also presents rich semantic information regarding locations through text associated with geotagged social media posts. Comparative analysis of corpora associated with

GeoExt 2023: First International Workshop on Geographic Information Extraction from Texts at ECIR 2023, April 2, 2023, Dublin, Ireland

C.Berragan@liverpool.ac.uk (C. Berragan); ucfnale@liverpool.ac.uk (A. Singleton); calafio@ed.ac.uk (A. Calafiore); jeremy.morley@os.uk (J. Morley)
0000-0003-2198-2245 (C. Berragan); 0000-0002-2338-2334 (A. Singleton); 0000-0002-5953-2891 (A. Calafiore); 0000-0002-3658-8796 (J. Morley)

© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons Licence Attribution 4.0 International (CC BY 4.0).
CEUR Workshop Proceedings (CEUR-WS.org)

geotagged locations similarly exhibit regionality; for example, tweets from the North East of England are statistically different compared with the South [5].

Our paper explores the similarity of corpora with respect to locational mentions from data taken directly from text, without relying on geotagged metadata. This approach offers an alternative perspective for the analysis of social interaction, built directly from the semantic information associated with locations, rather than the location associated with social media users themselves. Collective semantic information from social media embeds the regional identity of locations across a continuous spectrum, allowing for the direct comparison between these identities and their relationships.

2. Methodology

The following section gives an overview of our data source and the data processing methodology used in our paper. All code, analysis and data are available on our DagsHub repository.

Reddit is a public discussion, news aggregation social network, among the top 20 most visited websites in the United Kingdom. As of 2020, Reddit had around 430 million active monthly users, comparable to the number of Twitter users [7, 8]. Reddit is divided into separate independent *subreddits* each with specific topics of discussion, where *users* may submit *posts* which each have dedicated nested conversation threads that users can add *comments* to. Subreddits cover a wide range of topics, and in the interest of geography, they also act as forums for the discussion of local places. The United Kingdom subreddit acts as a general hub for related topics, notably including a list of smaller and more specific related subreddits. This list provides a 'Places' section, a collection of local British subreddits, ranging in scale from country level (`/r/England`), regional (`/r/thenorth`, `/r/Teeside`), to cities (`/r/Manchester`) and small towns (`/r/Alnwick`). In total there are 213 subreddits that relate to 'places' within the United Kingdom[1]. For each subreddit, every single historic comment was retrieved using the Pushshift Reddit archive [9]. In total 8,282,331 comments were extracted, submitted by 490,535 unique users, between 2011-01-01 and 2022-04-17.

We extracted and geolocated all place names in this collection of comments using a custom built geoparsing pipeline. To identify place names, we used a BERT transformer-based NER model trained on the WNUT 2017 dataset [10], available on the HuggingFace Model Hub. We then implemented a disambiguation methodology using contextual place names and two gazetteers to geolocate place names; OS Open Names and 'natural' location types from the Gazetteer of British Place Names. Processed comments consist of a collection of geolocated place names, alongside their natural language context sentence.

2.1. Similarity of Place Corpora

Comparing the similarity between two or more distinct texts first relies on an appropriate method for processing the text into a numerical format. For each location we obtained a corpus of comments, consisting of sentences where each location is mentioned. These were then processed into a single vector, reflecting the semantic information attributed with locations.

[1] https://www.reddit.com/r/unitedkingdom/wiki/british_subreddits

Typically, a TF-IDF approach is used to generate document embeddings [11], however we found comparative analysis between embeddings did not always provide insightful information. Each vector shared similar properties, giving cosine similarities which did not result in any distinct variation between locations. This is likely a problem with the language between locations sharing similar properties, meaning the more nuanced semantic information is not captured through a TF-IDF method.

We therefore extracted embeddings from a deep neural network called a transformer. Unlike TF-IDF or simpler neural network models, transformers are able to use contextual information to generate word embeddings, meaning the same word in two different contexts will not share the exact same vector, capturing different embedded semantic information [12]. Additionally, transformers are *pre-trained* on a large corpus of text, meaning general information regarding the English language is already embedded within the model, allowing for improved understanding of semantic information. These core features mean that embeddings generated from transformers are likely to capture information that allows for more the accurate comparative analysis. We generated embeddings using the `all-mpnet-base-v2` model from the `sentence-transformers` library in Python [13]. Unlike a standard 'BERT'-like transformer, this library implements modifications to base models that more appropriately captures semantic information in their output embeddings.

Before calculating embeddings we first masked every mention of a location with a generic token 'PLACE', this ensured that when analysing embeddings, no explicit geographic information was captured accidentally. For example, Manchester and Liverpool may mention matching locations frequently in each of their comments because they are geographically close. To both remove noise and reduce the computational requirements for this work, only locations with over 10,000 unique mentions were kept, from these a random sample of 1,000 comments were selected for each. Once embeddings were generated for every comment in each city corpus, the mean for each corpus was generated, giving a vector 768 decimal values for each city.

With a single vector for each selected location, we first calculated K-Means clusters to determine whether distinct groupings of locations could be identified across the UK. To visualise these clusters we used a PCA decomposition to reduce the dimensionality from 768 down to 2 dimensions. Finally, we calculated the cosine similarity between each and every location vector.

3. Results & Discussion

Figure 1 gives K Means clusters for transformer embeddings decomposed into two dimensions with $k = 5$. These Clusters show corpora that share similar semantic properties, however, it is worth noting that while points that are closer together likely indicate increased similarity, the position of these points reflect PCA decomposed values, which capture less information compared with the clusters calculated on non-decomposed vectors. Notably London appears as a single value in a cluster, suggesting the corpus associated with the capital of the UK is semantically distinct from the rest of the country. There is also a single cluster associated with the four Scottish cities considered in our study (Cluster 1), as well as a cluster for Cambridge and Oxford (Cluster 5). Figure 1 (B) reveals that clusters do broadly appear to reflect distance-

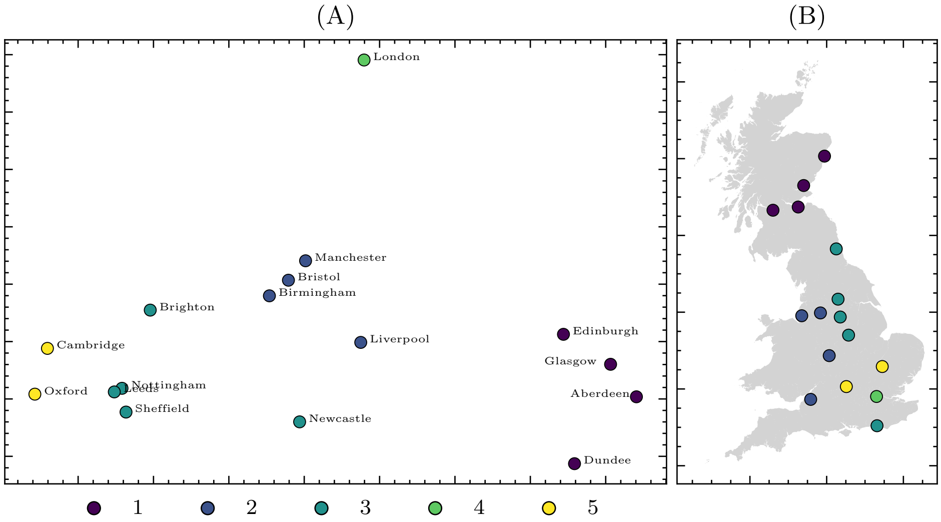

Figure 1: Average transformer vector associated with each location corpus coloured by K Means clusters where $K = 5$. (A) PCA decomposed into 2 dimensions. (B) Visualised with their easting and northing coordinates.

restricted geographic properties, while also capturing some divergences from this, with locations like London, Newcastle, Bristol and Brighton geographically distant from locations they share clusters with.

With our high dimensional transformer embeddings we compare the cosine similarity between them on Figure 2. The highest and lowest similarity score for each location is highlighted in red and green respectively. As with Figure 1, corpora in Scottish cities appear to largely share similarities, with Glasgow and Edinburgh sharing their highest similarity values. The city with the lowest similarity to the most other locations is Oxford, which shares low values with cities in Scotland, as well as Liverpool and Manchester. London again stands out, with overall very low similarities with all other cities, but the highest similarity with Manchester.

4. Conclusion

Our paper demonstrates the ability to compare Reddit comments relating to cities across the UK, using document embeddings generated from a transformer neural network. Instead of focussing on physical interactions between people or social media interactions, our work identifies relationships between cities through their semantic footprint, and analysing each corpus computationally allows for direct comparisons between cities through clustering and cosine similarity.

Our analysis reveals distinct clusters which largely reflect geographic proximity of locations, however, interesting deviations from proximity do emerge. Oxford and Cambridge are both clustered and share a high cosine similarity, but generate the lowest similarity with many other locations in the UK, including London. London in particular appears distinct from the rest of

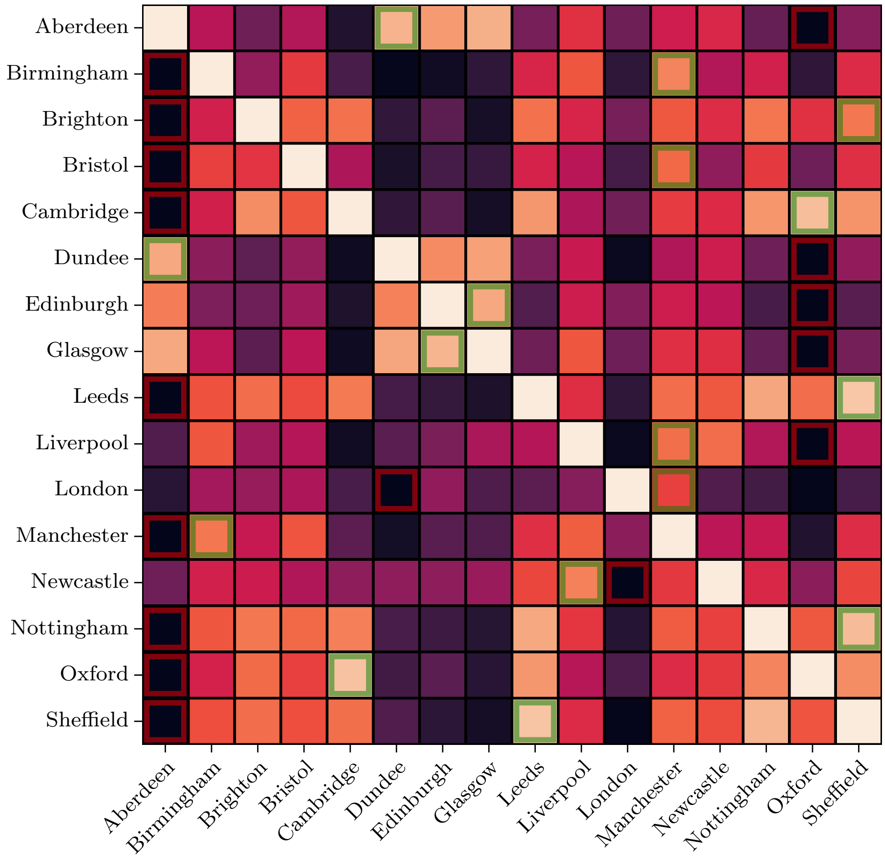

Figure 2: Cosine similarity between each and every location related transformer vector embedding. Values scaled between 0 and 1. Green highlights indicate the highest value in each row, while red indicates the lowest value in each row.

the UK, while cities that are not geographically close exhibit clustering and high similarity, such as Liverpool and Newcastle.

The information generated through our work presents an alternative view of relationships between cities that are not captured by existing data sources, all of which rely on explicit geographic coordinate information. Instead, we build similarities and clusters directly from the semantic information that exists within their respective corpora. Unlike traditional data, which captures objective social interactions between regions, the deviations from the restriction of geographic distance between several cities in our work appears to reflect the more subjective language that shapes the cultural and perceived identity of regions, and the relationships between them.

While our work enables the direct numerical comparison between city-based corpora, it cannot explain the similarities and dissimilarities between them. Additional work may explore the use of topic-modelling to identify shared topics between locations, and differences in the sentiment towards these topics may explain dissimilarity.

References

[1] A. Rae, From spatial interaction data to spatial interaction information? Geovisualisation and spatial structures of migration from the 2001 UK census, Computers, Environment and Urban Systems 33 (2009) 161–178. doi:10.1016/j.compenvurbsys.2009.01.007.

[2] H. Titheridge, K. Achuthan, R. L. Mackett, J. Solomon, Assessing the extent of transport social exclusion among the elderly, Journal of Transport and Land Use 2 (2009). doi:10.5198/jtlu.v2i2.44.

[3] S. Sobolevsky, M. Szell, R. Campari, T. Couronné, Z. Smoreda, C. Ratti, Delineating Geographical Regions with Networks of Human Interactions in an Extensive Set of Countries, PLoS ONE 8 (2013) e81707. doi:10.1371/journal.pone.0081707.

[4] B. Lengyel, A. Varga, B. Ságvári, Á. Jakobi, J. Kertész, Geographies of an Online Social Network, PLOS ONE 10 (2015) e0137248. doi:10.1371/journal.pone.0137248.

[5] R. Arthur, H. T. P. Williams, The human geography of Twitter: Quantifying regional identity and inter-region communication in England and Wales, PLOS ONE 14 (2019) e0214466. doi:10.1371/journal.pone.0214466.

[6] C. Ratti, S. Sobolevsky, F. Calabrese, C. Andris, J. Reades, M. Martino, R. Claxton, S. H. Strogatz, Redrawing the Map of Great Britain from a Network of Human Interactions, PLoS ONE 5 (2010) e14248. doi:10.1371/journal.pone.0014248.

[7] N. Murphy, Reddit's 2019 Year in Review - Upvoted, https://www.redditinc.com/blog/reddits-2019-year-in-review/#content, 2019.

[8] Statista, Most popular social networks worldwide as of January 2022, ranked by number of monthly active users, https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/, 2022.

[9] J. Baumgartner, S. Zannettou, B. Keegan, M. Squire, J. Blackburn, The Pushshift Reddit Dataset, 2020. arXiv:arXiv:2001.08435.

[10] L. Derczynski, E. Nichols, M. van Erp, N. Limsopatham, Results of the WNUT2017 Shared Task on Novel and Emerging Entity Recognition, in: Proceedings of the 3rd Workshop on Noisy User-generated Text, Association for Computational Linguistics, Copenhagen, Denmark, 2017, pp. 140–147. doi:10.18653/v1/W17-4418.

[11] J. Daniel, M. James H, Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, prentice hall, 2007.

[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention Is All You Need, arXiv:1706.03762 [cs] (2017). arXiv:1706.03762.

[13] N. Reimers, I. Gurevych, Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 3980–3990. doi:10.18653/v1/D19-1410.