Mapping Semantic Regional Variation in Great Britain through Reddit Comments

Anonymous

Abstract

Observed regional variation in geotagged social media text is often attributed to dialects, where features in language are assumed to exhibit region-specific properties. While dialects are seen as a key component in defining the identity of regions, there are a multitude of other geographic properties that may be captured within natural language text. In our work, we consider locational mentions that are directly embedded within comments on the social media website Reddit, providing a range of associated semantic information, and enabling deeper representations between locations to be captured. Using a large corpus of Reddit comments from UK related local discussion subreddits, we first identify place names using a transformer-based named entity recognition model. Embedded semantic information is then generated from these comments and aggregated into local authority districts, representing the semantic footprint of these regions. These footprints broadly exhibit spatial autocorrelation, with clusters that conform with the national borders of Wales and Scotland. London, Wales, and Scotland also demonstrate notably different semantic footprints compared with the rest of the UK, partially explainable through the perception of national identity associated with these regions.

# 1. Introduction

While formal geographic regions within Great Britain are typically designed for administrative and political purposes, they are non-restrictive in how populations can move between them. The level of geographic cohesion between regions across the UK is often studied from the context of mobility, where data sources like Census or transport records describe the physical movement of populations and individuals across geographic space (Rae, 2009; Titheridge *et al.*, 2009), or through non-physical networks using phone records (Lambiotte *et al.*, 2008; Reades, Calabrese and Ratti, 2009; Sobolevsky *et al.*, 2013; Zheng, 2015), and social media (Sui and Goodchild, 2011; Lengyel *et al.*, 2015; Arthur and Williams, 2019). When these networks are examined, cohesive clusters develop, which broadly appear to correlate with administrative boundaries (Ratti *et al.*, 2010; Arthur and Williams, 2019).

As with social interactions, language is also known to exhibit regional variation (Trudgill, 2004), and certain dialects contribute to the concept of identity in regions across the UK (Haesly, 2005; Llamas, 2009; Llamas and Watt, 2014). While variations in dialects may be explored from the perspective of survey data (Clopper, Tamati and Pierrehumbert, 2016), alternative approaches have considered statistical methods for determining lexical variation in geographically bounded social media networks (Russ, 2012; Doyle, 2014; Gonçalves and Sánchez, 2014; Eisenstein *et al.*, 2014; Huang *et al.*, 2016; Pérez *et al.*, 2019; Arthur and Williams, 2019). Comparative analysis of corpora associated with geotagged locations have similarly been shown to exhibit regional properties; for example, Tweets originating in the North East of England are noticeably different compared with the South (Arthur and Williams, 2019).

While social media text has been shown to exhibit regional variation, these works prevalently rely on the use of geotagged posts. The content of geotagged posts is often unrelated to the geotagged location (Kropczynski *et al.*, 2018), meaning any geographically associated lexical variation is only attributable to dialects, despite only presenting a single component of language that may exhibit geographic heterogeneity. Instead, our paper explores the similarity of corpora with respect to locational mentions taken from social media text, embedded directly within a related semantic context. Place names may be attributed with coordinate information through a process called geoparsing, where names are identified automatically in text through a natural language processing technique called named entity recognition, and associated an entry in a gazetteer. This approach offers an alternative perspective for the analysis of geographic variation in social media, built directly from the semantic information associated with locations, rather than the location associated with social media users themselves. Semantic information therefore represents a broad range of topics and associations directly relating to geographic locations, rather than just the dialect of contributing users.

Our hypothesis is that related semantic information regarding locations from social media embeds a regional semantic footprint, representing the collective geographic knowledge of each individual user in our corpus. Additionally, any geographic knowledge extracted from social media represents an informal, place-based understanding of geography, which more appropriately reflects a cognitive understanding and perception of locations (Sui and Goodchild, 2011; Goodchild, 2011; Berragan, no date). Semantic footprints are represented through embeddings; numerical representations of text that capture semantic information, enabling a deeper analysis compared with lexicons, and allowing for a direct numerical comparison between them. To explore why certain regions in the UK appear more semantically isolated, we consider the link between identity, querying a large language model to determine the perception of the strength of national identities in UK regions.

[Section 2](#sec-literature) first gives an overview of work exploring semantic variation in social media text, place-based geographic knowledge, and regional identities. [Section 3](#sec-methodology) first describes our data, then outlines the processing used to generate regional embeddings for comparison and our method for identifying regional identities within these comments. [Section 4](#sec-results) presents our results and [Section 5](#sec-conclusion) concludes with suggestions for future work.

We have identified the following key aims that are addressed in this paper:

**Aim 1:** Generate cohesive clusters of regions in Great Britain based on embedded locational semantic information.

**Aim 2:** Calculate Semantic similarity measures between regional corpora to determine the least and most semantically connected regions.

**Aim 3:** Examine the distribution of perceived national identity across all British regions using a zero-shot classification.

# 2. Geographic Variation in Social Media Text

While it is well understood that despite the advancement of communication technologies, social ties still appear to be geographically restricted (Ratti *et al.*, 2010; Scellato *et al.*, 2010; Onnela *et al.*, 2011; Takhteyev, Gruzd and Wellman, 2012; Sobolevsky *et al.*, 2013), it is less well understood why these geographic restrictions appear to correlate with administrative boundaries (Ratti *et al.*, 2010; Stephens and Poorthuis, 2015; Yin *et al.*, 2017; Arthur and Williams, 2019). When considering the lexical properties of text associated with geotagged social media data, geographically cohesive groupings occur (Eisenstein *et al.*, 2014; Gonçalves and Sánchez, 2014; Huang *et al.*, 2016; Arthur and Williams, 2019), which conforms with the idea that dialects are an important component in the identity of regions (Haesly, 2005; Llamas, 2009; Llamas and Watt, 2014). Dialects however, only present a single component of language that may be associated with geographic locations.

The prevalence of social media data for use in geographic research has generated a renewed interest in the concept of ‘place’ (Westerholt *et al.*, 2018; Purves, Winter and Kuhn, 2019; Wagner, Zipf and Westerholt, 2020; Berragan, no date), as contributions to social media are theorised to capture informal knowledge that represents a ‘place-based’ understanding of geography (Goodchild, 2011; Sui and Goodchild, 2011). In the context of language, this place-based knowledge is generated through ‘vernacular geography’, which describes the natural language used when informally describing geographic locations (Waters and Evans, 2003; Hollenstein, 2008; Goodchild, 2011; Gao *et al.*, 2017). This informal knowledge incorporates associated biases regarding locations, better representing human associations between locations, compared with formal administrative aggregations. In this sense, associations of geography drawn from social media capture place through a ‘bottom-up’ approach, building knowledge through experience rather than administrative formalisations (Agnew, 2005; Sui and Goodchild, 2011). While many works consider the formalisation of place through social media data, few have considered how the semantic properties of text may reveal associations between regions, captured through vernacular geography.

## 2.1 Measuring Regional Variation in Language

Regional variations in linguistic characteristics are often quantified from a statistical analysis of lexicons, considering that regions exhibit subtle differences in dialects that may be captured. For example, studies have considered regional differences in common words on Twitter (Russ, 2012; Han, Cook and Baldwin, 2012; Eisenstein *et al.*, 2014; Gonçalves and Sánchez, 2014; Doyle, 2014; Huang *et al.*, 2016; Zheng, Han and Sun, 2018; Arthur and Williams, 2019), or identified spoken differences in dialects from active participation methods (Clopper, Tamati and Pierrehumbert, 2016). While few explicitly consider a comparative analysis of regional dialects, cosine similarity measures have been used to examine geotagged lexicons from Twitter within England and Wales (Arthur and Williams, 2019). Results demonstrated that the Welsh lexicon appeared to be the most dissimilar to all other regions, while Northern English regions were more similar to each other compared with Southern regions. Alternatively vague ‘vernacular’ regions have been defined through an analysis of dialects, where geotagged lexical alternations in Tweets were used to establish a clear North and South divide in the USA (Huang *et al.*, 2016).

Given dialects are considered key in defining regional identities, these observations likely in part explain the coherent regions that form through both physical and online social interactions. Users are more likely to interact with others that share a similar identity, which is in part captured through the lexical components of language. Despite this, there are a multitude of other properties that contribute to the perceived identity of geographic regions, which are broadly definable as non-specific cultural associations (Haesly, 2005; Middleton and Freestone, 2008). Studies that consider dialects on social media are limited by the ability to capture geographic properties from existing data sources, where the only geographic component is an associated geotag. Given the content of geotagged social posts does not necessarily relate to the geotagged location, semantic information extracted therefore represents a broad range of non-specific and unrelated topics. Any observed regional variation is only attributable to the dialect of the contributing author, with the assumption that the author is a resident in the geotagged location.

With the ability to accurately geoparse text, locations can instead be extracted that are embedded directly within a related semantic context, without relying on geotags. This method enables semantic information in text to be directly associated with defined geographic locations, providing a context covering a broad range of topics relating to locations, built from the experiential geographic knowledge of contributing authors. Given social media contributes informal geographic knowledge (Sui and Goodchild, 2011), observations generated from this data conforms with a general understanding of ‘place’, incorporating cultural knowledge and characteristics, alongside cognitive biases and perceptions of locations.

In this work we aim to generate a new comparative measure between regions in the UK through an examination of text associated with locations, extracted from comments on the social media website Reddit. While past work has examined variation between regions from the perspective of social media networks, or by examining lexicons associated with geotagged social media messages, we examine regional variations derived from text associated with directly embedded place names. Unlike using geotags, which ascribe linguistic features such as dialect to specific locations, our method instead captures any comment that mentions a location alongside its context. Quantified information therefore does not reflect dialects associated with locations, but common semantic associations, embedding cultural information, or location specific topics and opinions.

## 2.2 Embedding Semantic Footprints

Statistical comparisons between two or more distinct texts first relies on an appropriate method for processing the text into a numerical format. Typically, a TF-IDF approach is used to generate document embeddings (Daniel and James H, 2007), which assigns word importance based on the frequency of mentions within a corpus. TF-IDF however does not have the capability to capture broader semantic information, given there is no knowledge of the meaning behind words. Large Language Models (LLMs) instead are pre-trained on a very large corpus of natural language text, which, alongside their architecture, enables them to more appropriately consider semantic information (Devlin *et al.*, 2019). Similarly to TF-IDF, text is input into these models and output as a numerical representation, which embeds words as high dimensional vectors, capturing contextual semantic information.

This approach differs from past work that only considered a lexical analysis, where semantic information and context is not preserved, instead building vectors that act as semantic representations of locations identified in our corpus, which we name ‘semantic footprints’. Given semantic information is preserved, locational embeddings are able to reflect the deeper associations between geographic locations, built from a multitude of contexts and perspectives, forming an aggregate representation. Any geographically cohesive relationships between footprints therefore demonstrate a direct association between geography and language, which hasn’t been captured previously.

## 2.3 Zero Shot Classification

As LLMs are pre-trained on a large corpus of natural language text, they build representations of this text that emulate a human understanding of language. In theory these representations capture the collective knowledge of humans that contributed the natural language text used to build them. Therefore, in addition to factual information, when posed with non-deterministic questioning, these models are able to contribute the biased information that is incorporated into their model weights.

Recent research has noted on the ability to perform zero-shot classification using LLMs, where class predictions may be made without the model ever having previously seen the labels (Wei, Tay, *et al.*, 2022; Wei, Bosma, *et al.*, 2022). While research has considered the use of questionnaires to query the strength of national identities within the UK (Haesly, 2005; Griffiths, 2022), an LLM may instead be used. For example, an LLM may be questioned whether it personally feels a sequence of text appears to be ‘British’, ‘English’, ‘Scottish’, or ‘Welsh’. Through this zero-shot classification, we are able to determine the perceived strength of national identity associated with each region in our work, to examine whether this appears to correlate with any cohesion between the semantic footprints that we generate.

# 3. Methodology

The following section first introduces our main data source; the social media website Reddit, from which we access a collection of user submitted comments. Following this, we detail our methodology for generating embeddings from each of these comments, and the processing that follows, to analyse the geographic properties of these embeddings.

[Reddit](https://reddit.com) is a public discussion, news aggregation social network, and among the top 20 most visited websites in the United Kingdom. In 2020, Reddit had around 430 million active monthly users, comparable to the number of Twitter users (Murphy, 2019; Statista, 2022). 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](https://reddit.com/r/unitedkingdom) 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 (/r/England), region (/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]](#footnote-27). We use the corpus generated by Berragan (no date), which consists of a collection of all Reddit comments taken from each UK related subreddit (Baumgartner *et al.*, 2020), with place names identified by a custom transformer-based named entity recognition model[[2]](#footnote-28). In total 8,282,331 comments were extracted, submitted by 490,535 unique users, between 2011-01-01 and 2022-04-17. Table gives an example entry from this geoparsed Reddit corpus.

There are a total of 52,169 unique locations in this corpus, with a highly skewed distribution in mentions. Most locations were only mentioned a single time, while ‘London’ was mentioned in almost 300,000 comments. To reduce this skew, we sampled any location mentioned more than 5,000 times, retaining only up to 5,000 randomly sampled comments per location. The goal with this processing was to ensure that our generated embeddings did not simply become biased towards the word embedding for a single location, and instead capture a broader sense of an aggregate region. In our data subset, we find that 1% of users (1,698) mention 29% of our place names.

Table gives an overview of the number of comments, word count and number of places that were identified within each administrative region of the UK. Our study concerns a subset of the full Reddit comment dataset, subsetting each location to a maximum of 5,000 mentions, leaving a total of 830,770 comments containing place names. Comments range from 1 to 3,555 words in length, with a mean length of 79.

## 3.1 Generating Embeddings

We first generate semantic embeddings for each comment in which a location was mentioned, using the sentence-transformers Python library (Reimers and Gurevych, 2019), with the all-mpnet-base-v2 model. With our selected embedding model, we then performed the following steps to generate embeddings for each Local Authority District (LAD) in Great Britain.

1. Masked any place name with a generic token: ‘PLACE’.
2. Generate sentence embeddings for each comment.
3. Group embeddings by LAD using identified locations, taking the mean embedding.

To visualise the outputs from this processing we consider an example comment , shown on Equation .

$$
\begin{aligned}
\mathit{s\_{i}} &= \text{'I live in \textit{London}'} \\
\textbf{1. }\downarrow \\
\mathit{s\_{i}} &= \text{'I live in \texttt{PLACE}'},
\end{aligned}
\qquad
\begin{aligned}
\textbf{2. }\mathit{s\_{i}} \rightarrow
\begin{bmatrix}
x\_{1} \\
x\_{2} \\
\vdots\\
x\_{n}
\end{bmatrix},
\end{aligned}
\qquad
\begin{aligned}
\textbf{3. }\mathit{LAD\_{j}} =
\begin{bmatrix}
x\_{1,1} & x\_{1,2} & \cdots & x\_{1,t} \\
x\_{2,1} & x\_{2,2} & \cdots & x\_{2,t} \\
\vdots & \vdots & \ddots & \vdots \\
x\_{n,1} & x\_{n,2} & \cdots & x\_{n,t}
\end{bmatrix} \rightarrow \begin{bmatrix}
\bar{x\_{1}} \\
\bar{x\_{2}} \\
\vdots \\
\bar{x\_{n}}
\end{bmatrix}
\end{aligned}
\qquad(1)$$

In Equation , is the sentence-transformers embedding dimension (768), and is the total number of unique comments that relate to a single LAD region (). Values () in step **2.** are model weights that represent the embedding for the comment , capturing semantic information. All comment embeddings associated with are then processed into one dimension by taking the mean (step **3.**), producing the semantic footprint. Reducing embedding dimensionality by taking the mean is common in NLP tasks. For example, sentence embeddings are simply the average of word embeddings, and are used in many tasks like topic identification or sentence similarity, with good results (Reimers and Gurevych, 2019).

By masking place names we ensure that no comment embeddings accidentally incorporated geographically grounded information. For example, comments in South Eastern local authorities are likely to frequently mention London, given they are geographically close. Embeddings for these locations would therefore capture an association through the mention of London, rather than general semantic information. For our work, we want to exclude any geographic information, ensuring that embeddings solely capture semantic associations.

Given that transformers are a fairly new architecture in natural language processing, and the creation of these models require significant computational resources and training time, their use to date has been limited in related research. Our choice to use the transformer architecture stems from the emphasis we place on the extraction of nuanced and contextual semantic information, which is lost with more statistical methods. It should be noted however that while TF-IDF lexical methods are less complex, they are typically more interpretable; for instance, words that contribute importance to an embedding may be extracted from a TF-IDF model. The numerical representations of any text generated by transformers are not directly interpretable in this manner.

## 3.2 Spatial Clustering and Autocorrelation

It is reasonable to assume that there are LADs within our corpora that generate embeddings that share similar properties. A typical method to group unlabelled multi-variate data based on shared properties uses unsupervised clustering. Therefore, to explore whether geographically cohesive clusters appear within our semantic embeddings (Aim 1), we generate hierarchical clusters, which are non-geographically bounded.

Agglomerative clustering was used to generate clusters of LAD embeddings, a type of hierarchical clustering method which we found performed better compared with K Means. Agglomerative clustering is able to automatically select the optimal number of clusters based on a distance threshold, which in our data was determined to be 1.5, giving three clusters. These Clusters were visualised geographically to examine whether geographically cohesive groupings occurred. The proportion of clusters presented within each administrative region (RGN)[[3]](#footnote-31) in Great Britain was also plotted to determine whether clusters appeared to correlate with administrative boundaries.

To further explore aim 1, we consider whether embeddings demonstrate spatial autocorrelation through the Moran’s I metric, and identify geographic clusters of shared semantic attributes, using a Local Indicators of Spatial Autocorrelation (LISA) analysis (Anselin, 1995; Rey, Arribas-Bel and Wolf, 2023). As spatial autocorrelation analysis is only univariate, we consider embeddings UMAP decomposed into 2 dimensions for this analysis, rather than the original 768 dimensional embeddings. It is important to note that the numerical value of this decomposed embedding does not convey any definable information based on its magnitude, meaning unlike typical LISA analysis, HH and LL values are only distinguished to highlight non-specific differences in semantic information.

### 3.2.1 Semantic Similarity

We determine the semantic similarity between regions across Great Britain (Aim 2), by grouping LAD embeddings into their respective RGN, taking the mean embedding value across each region. We then calculate the cosine similarity between each and every embedding, demonstrating the inter-region cohesion across Great Britain.

Cosine similarity is a common metric for comparing embeddings, as it is invariant to the magnitude of the vectors, and only considers the direction. This is important as the magnitude of embeddings is not meaningful, and only the direction of the vector conveys information. For example, the embedding for ‘London’ cannot be twice as important as the embedding for ‘Manchester’, and therefore the magnitude of the embedding is not meaningful.

## 3.3 Zero-Shot national identities

To determine whether the distribution of identities correlate with our findings (Aim 3), we examine the relative national identity strength based on the semantic information embedded within comments. Semantic information is expected to capture both explicit information contributed by users; for example stating ‘London is a British city’, in addition to implicit semantic information that exists within language. For this reason, we do not mask place name mentions in these embeddings. For example the phrase ‘bonnie Scotland’ may suggest a strong identity due to the inclusion of Scottish slang[[4]](#footnote-34).

To identify regional identities through semantic information, we build on the emergent properties of large language models, which enable a task known as ‘Zero-Shot Classification’. This allows models to predict a class that was not seen during training, by generating a prompt that contains the labels required. For our task an example prompt may be constructed as follows:

Classify the following input text into one of the following four categories:  
[British, English, Scottish, Welsh]  
  
Input Text: I feel more Welsh than British.

The output would then be given as a sequence of confidence values for each label:

'labels': ['Welsh', 'British', 'English', 'Scottish']  
'scores': [0.9902838468551636,  
 0.007325825747102499,  
 0.0013424316421151161,  
 0.0010479396441951394]

# 4. Results

|  |
| --- |
| Figure 1: Average transformer vector associated with each location corpus coloured by hierarchical Agglomerative clusters where . (a) LAD embeddings UMAP decomposed into 2 dimensions. (b) Proportion of clusters by RGN. (c) Geographic location of clusters. |

[Figure 1](#fig-clusters) (a) shows clusters of LAD transformer embeddings UMAP decomposed into two dimensions, indicating embeddings that share similar semantic properties. These clusters appear to broadly correlate with three distinct regions within the UK, where cluster 0 most closely identifies with Scotland and Wales, 1 with England, and 2 with London and surrounding areas ([Figure 1](#fig-clusters) (b-c)). The few areas that appear as cluster 1 in Wales and Scotland are major urban centres like Cardiff, Glasgow, and Edinburgh, in addition to Pembrokeshire in Wales. Overall these clusters appear to be geographically restricted, and even broadly conform with administrative regions like the Welsh and Scottish borders.

These findings appear to correlate with past work that has observed strong ‘boundary effects’, where lexical similarity between geotagged Tweets often correlates with administrative boundaries (Yin *et al.*, 2017; Bailey *et al.*, 2018; Arthur and Williams, 2019; Li *et al.*, 2021). This observation in our results appears particularly strong at the borders of Scotland and Wales, suggesting that the prominent sense of identity within these countries has generated semantic coherence (Haesly, 2005). Note that unlike past work that has demonstrated this effect from the perspective of dialects or interviews from residents, our results instead capture the perceived sense of identity associated with these countries, built from all users in our corpus.

As Glasgow, Edinburgh and Cardiff share a cluster with English LADs rather than their respective country, these locations are more semantically connected with the rest of the UK, compared with other regions. This observation mirrors the results of work that considered co-occurring locational mentions between cities, where shared city mentions in text often appear irrespective of distance, and across administrative borders (Berragan, no date). However, while Pembrokeshire is not a major urban centre, the alignment with English LADs is likely due to its interest as a holiday destination.

Cluster 2 generated surrounding London suggests a distinct perceived separation with this region of the UK. This is interesting given London’s extensive physical connections through high speed rail, and general sense of strong association with other cities, given it is the capital city (Berragan, no date). Our results therefore suggest that despite London’s importance nationally, semantic information is able to capture a deeper context that dissociates it from other regions in the UK.

The following section formalises the level of geographic coherence our embeddings exhibit, and explores the key locations that drive the relationship between text and geography.

## 4.1 Moran’s I Analysis

|  |
| --- |
| Figure 2: Moran’s I Plot: LAD embeddings decomposed into 2 dimensions and standardised against their spatial lag. |

To quantify whether our embeddings demonstrate spatial autocorrelation, we consider the Moran’s I metric, which considers the spatial relationship between each observation and its neighbours (Anselin, 1995). Given this analysis requires univariate data, we explore global spatial autocorrelation of our UMAP decomposed embeddings by plotting both dimensions against their spatial lag on [Figure 2](#fig-morans). The Moran’s I values of <esda.moran.Moran at 0x7ef6b23cb3a0> and <esda.moran.Moran at 0x7ef6949a8eb0> indicate a reasonably strong spatial autocorrelation with both embedding dimensions, confirming that semantic information is often similar with nearby locations. While the Moran’s I values for both dimensions are similar, their cosine similarity is low (array([[-0.09880356]], dtype=float32)), meaning it is likely that these two decomposed dimensions capture distinctly different semantic traits.

While spatially coherent results have been demonstrated from the perspective of dialects on social media (Russ, 2012; Doyle, 2014; Gonçalves and Sánchez, 2014; Eisenstein *et al.*, 2014; Huang *et al.*, 2016; Pérez *et al.*, 2019; Arthur and Williams, 2019), we have demonstrated that this phenomenon can also be captured from general semantic information. Dialects have always been considered to have strong geographical grounding (Trudgill, 2004), but unlike dialects, it is more surprising that general semantic information regarding locations similarly exhibits this relationship.

|  |
| --- |
| Figure 3: Local Indicators of Spatial Auto-correlation (LISA). (a/d) 1 dimensional embedding values. (b/e) Local Moran’s I values (). (c/f) LISA HH and LL significant values (), both are included as the value of embeddings do not convey information. |

To explore local indicators of spatial autocorrelation (LISA) we plot each decomposed embedding on [Figure 3](#fig-lisa) (a/d), each local Moran’s I value on (b/e) and all significant () HH and LL LISA quadrants on (c/f). Note that only selecting significant values on [Figure 3](#fig-lisa) (c/f) ensures that no regions are included that have values that could demonstrate autocorrelation even if randomly distributed geographically. From [Figure 3](#fig-lisa) (c/f), we can see that notable areas with strong spatial correlation include;

* Scotland
* Wales
* London and surrounding LADs
* the South West; towards Cornwall

When we compare the first and second UMAP dimension they appear capture different semantic information. London for example only appears in dimension 1, while dimension 2 captures more broad spatial autocorrelation across Scotland and Wales. In Scotland we can see that from the dimension 1 LISA, both Glasgow and Edinburgh represent areas of HL/LH, where semantic information in these cities is not the same as surrounding LADs, and effect that is also captured in some LADs surrounding London. For England, there appears to be fewer geographically cohesive semantic associations, most notably in the South West and part of the Midlands.

These results again demonstrate the geographically cohesive and administrative boundary effects that have been observed in past work, with respect to Scotland and Wales. While in addition we have captured a notable grouping in the South West, which potentially reflects the Cornish identity (Deacon, 2007), and a grouping associated with London.

## 4.2 Semantic Similarity of Regions

|  |
| --- |
| Figure 4: Scaled cosine similarity of embeddings for administrative regions across the UK. Higher values indicate greater cosine similarity. Regions shown in descending order by mean cosine similarity value. |

On [Figure 4](#fig-similarity) we compare the cosine similarity between each RGN embedding, allowing for inter-regional cohesion to be explored. Both the North and South West have the overall highest levels of cosine similarity with each other region, displaying comparatively high similarity with each other region. London has the lowest overall similarity, only sharing high cosine values with the South and South East of England. As expected, Scotland and Wales have low overall cosine values, with Wales sharing even lower similarity with respect to London compared with Scotland. Mean values show clearly that the least cohesive regions appear to be London, Wales, and Scotland, three regions that are also those with the strongest levels of spatial autocorrelation.

Excluding London, the North East is the region in England with the lowest overall cosine similarity with the rest of the UK. This is perhaps reflective of the sense of inter-region identity that is often noted with this region (Middleton and Freestone, 2008), as well as the lack of physical interaction with the rest of the country due to a reduction in local industry and jobs (Kalantaridis, 2010). Alternatively, given the North West is home to nationally important cities like Manchester and Liverpool, it shares strong similarity across the whole of the UK, excluding Wales. Other patterns emerge in regions like Yorkshire, where there is notably low similarity with the South Western regions of England, and while Wales shares low overall cosine similarities with England, Scotland has a slightly increased similarity, particularly with London. Major urban centres in Scotland are relatively well connected through rail routes, and Edinburgh and Glasgow are historically important UK cities, while Wales in this sense is less directly associated with the rest of the UK.

## 4.3 Regional Identities

Finally, we consider whether the concept of a perceived regional identity generated from a large language model aligns with the observed regional variation in our analysis. [Figure 5](#fig-identity) shows the distribution of regional identities identified through our zero-shot classification methodology.

Across each region, the ‘English’ identity is always lower than ‘British’, suggesting that regions within England typically identify more strongly with the United Kingdom than solely England. Unlike English regions however, both Scotland and Wales identify most strongly with their respective nationalities. While the Welsh identity is reasonably weak, below the weakest British identity in the UK, the Scottish identity is by far the most strong of any identity across every region. Of the English regions, London has a distinctly higher average confidence value of both British and English identities compared with all other regions. These results also broadly agree qualitative interviewing, where the national identity of both Scotland and Wales is typically stronger compared with their British identity, but the Welsh identity is typically weaker than the Scottish identity (Haesly, 2005; Llamas, 2009; Carman, Johns and Mitchell, 2014; Llamas and Watt, 2014). Given these three regions have the lowest overall cosine similarity values indicated on [Figure 4](#fig-similarity) (Mean), these perceived differences in identity are a likely component in their semantic differences.

|  |
| --- |
| Figure 5: Zero Shot classification of each corpus into regional identities; [B]ritish, [E]nglish, [S]cottish, [W]elsh. Values show mean confidence value across each comment, lines indicate 95% confidence intervals. Descending order by [B]ritish confidence. |

To identify why the model makes certain predictions, we input a selection of phrases to identify the model confidence outputs, shown on Table . The model makes some interesting choices depending on the sentiment of the input text. For example if we input ‘I love Scotland.’ the model labels the text as Scottish, while if we change the text to ‘I hate Scotland!’ the model labels the text as British, with similar results of Wales. These results perhaps suggest that the model understands that personal national sentiment is typically more likely to be positive. As demonstrated on Twitter for example, negative perceptions of locations are generally directed towards locations external to a users home location (Butler, Schafran and Carpenter, 2018). However, ‘I hate England!’ and ‘I love England!’ are interestingly both considered British or English, but with low confidence. In many cases, our model appears mostly reluctant to label any text explicitly talking about English locations or topics as English, and prefers to use British, perhaps reflecting the general strong association of England with the UK identity, which is as strong in either Scotland and Wales.

# 5. Conclusion

Our paper demonstrates the ability to compare aggregate semantic information for local authorities and regions within the UK, from Reddit comments that mention geoparsed locations. When examining the semantic properties of each LAD in the UK, we find that geographically cohesive clusters appear, which exhibit a relatively strong spatial autocorrelation. Clusters broadly conform with the national borders of Scotland and Wales, while London also appears to be semantically distinct from the rest of the country. When examining the perception of national identities associated with regions, we find that these identities conform with these distinct geographic groupings.

As demonstrated in past work that has examined both physical and non-physical networks, our observed semantic information similarly appears to correlate with pre-defined administrative boundaries, particularly the national boundaries of Scotland and Wales (Yin *et al.*, 2017; Bailey *et al.*, 2018; Arthur and Williams, 2019; Li *et al.*, 2021). Additionally, the distinct difference in footprints between each constituent country in the UK likely relates to the perceived national identity that we generate through our zero-shot transformer classification. Scotland in particular is noted as having a strong national identity that is distinct from Britain, while similar is true in Wales, but to a lesser extent (Haesly, 2005; Carman, Johns and Mitchell, 2014). Similar to how dialect and social interaction networks conform with administrative regions, semantic information likely embeds the cultural perception of identity that causes distinct borders between regions. Interestingly however, despite having clear semantic associations, given they share a cluster on [Figure 1](#fig-clusters), and generate similar decomposed embedding values on [Figure 3](#fig-lisa), both Wales and Scotland are demonstrated to have distinct national identities on [Figure 5](#fig-identity).

Despite most locations across Scotland and Wales appearing disconnected with the rest of the UK, major cities like Glasgow and Edinburgh are more semantically similar, a distinction that was also observed when the distance decay of locational co-occurrences in text was examined (Berragan, no date). This observation suggests that these cities do appear to be typically more semantically cohesive regardless of geographic distance, while other locations typically share semantic properties within the same nation, captured through a stronger spatial autocorrelation.

When considering the cosine similarity between RGN embeddings, we see evidence of an alternative North-South divide that is often considered in England, which from a semantic context appears to be more related to proximity to London. Unlike typical representations of this divide therefore, the South West of England appears to be distinct from the South East, with a stronger association with the North. South Eastern regions however do share lower similarity to the Midlands and North of England, which conforms with a typical view of the English North-South divide (Jewell, 1994).

Geoparsing methods capture a geographic dimension to non-geotagged social media data, enabling a much larger repository of informal natural language geographic text to be used for research. Future work may consider the use of Reddit comment data to derive notable urban areas of interest for example (Chen, 2019). This area of research in particular would benefit from methodologies focussing on the extraction of fine-grained locations from text, which at present is a challenging task (Han *et al.*, 2018).

# References

Agnew, J. (2005) ‘Space: Place in Cloke P and Johnston R eds Spaces of geographical thought’.

Anselin, L. (1995) ‘Local Indicators of Spatial Association’, *Geographical Analysis*, 27(2), pp. 93–115. Available at: <https://doi.org/10.1111/j.1538-4632.1995.tb00338.x>.

Arthur, R. and Williams, H.T.P. (2019) ‘The human geography of Twitter: Quantifying regional identity and inter-region communication in England and Wales’, *PLOS ONE*. Edited by E. Ferrara, 14(4), p. e0214466. Available at: <https://doi.org/10.1371/journal.pone.0214466>.

Bailey, M. *et al.* (2018) ‘Social Connectedness: Measurement, Determinants, and Effects’, *Journal of Economic Perspectives*, 32(3), pp. 259–280. Available at: <https://doi.org/10.1257/jep.32.3.259>.

Baumgartner, J. *et al.* (2020) ‘The Pushshift Reddit Dataset’. arXiv. Available at: <https://arxiv.org/abs/2001.08435> (Accessed: 18 May 2022).

Berragan, C. (no date) ‘Mapping Cognitive Place Connectivity within the United Kingdom through Online Discussion on Reddit’.

Butler, A., Schafran, A. and Carpenter, G. (2018) ‘What does it mean when people call a place a shithole? Understanding a discourse of denigration in the United Kingdom and the Republic of Ireland’, *Transactions of the Institute of British Geographers*, 43(3), pp. 496–510. Available at: <https://doi.org/10.1111/tran.12247>.

Carman, C., Johns, R. and Mitchell, J. (2014) *More Scottish than British: The 2011 Scottish Parliament Election*. Springer.

Chen, M. (2019) ‘Understanding the dynamics of urban areas of interest through volunteered geographic information’, p. 21. Available at: <https://doi.org/10.1007/s10109-018-0284-3>.

Clopper, C.G., Tamati, T.N. and Pierrehumbert, J.B. (2016) ‘Variation in the strength of lexical encoding across dialects’, *Journal of Phonetics*, 58, pp. 87–103. Available at: <https://doi.org/10.1016/j.wocn.2016.06.002>.

Daniel, J. and James H, M. (2007) *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. prentice hall.

Deacon, B. (2007) ‘County, Nation, Ethnic Group? The Shaping of the Cornish Identity’, *The International Journal of Regional and Local Studies*, 3(1), pp. 5–29. Available at: <https://doi.org/10.1179/jrl.2007.3.1.5>.

Devlin, J. *et al.* (2019) ‘BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding’, *arXiv:1810.04805 [cs]* [Preprint]. Available at: <https://arxiv.org/abs/1810.04805> (Accessed: 11 February 2020).

Doyle, G. (2014) ‘Mapping Dialectal Variation by Querying Social Media’, in *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*. Gothenburg, Sweden: Association for Computational Linguistics, pp. 98–106. Available at: <https://doi.org/10.3115/v1/E14-1011>.

Eisenstein, J. *et al.* (2014) ‘Diffusion of Lexical Change in Social Media’, *PLoS ONE*. Edited by R.C. Berwick, 9(11), p. e113114. Available at: <https://doi.org/10.1371/journal.pone.0113114>.

Gao, S. *et al.* (2017) ‘Constructing gazetteers from volunteered Big Geo-Data based on Hadoop’, *Computers, Environment and Urban Systems*, 61, pp. 172–186. Available at: <https://doi.org/10.1016/j.compenvurbsys.2014.02.004>.

Gonçalves, B. and Sánchez, D. (2014) ‘Crowdsourcing Dialect Characterization through Twitter’, *PLoS ONE*. Edited by T. Preis, 9(11), p. e112074. Available at: <https://doi.org/10.1371/journal.pone.0112074>.

Goodchild, M.F. (2011) ‘Formalizing Place in Geographic Information Systems’, in L.M. Burton et al. (eds) *Communities, Neighborhoods, and Health*. New York, NY: Springer New York, pp. 21–33. Available at: <https://doi.org/10.1007/978-1-4419-7482-2_2>.

Griffiths, J.D. (2022) ‘Scrutinizing Relative Territorial Identity Measures’, *Publius: The Journal of Federalism*, 53(1), pp. 133–151. Available at: <https://doi.org/10.1093/publius/pjac011>.

Haesly, R. (2005) ‘Identifying Scotland and Wales: Types of Scottish and Welsh national identities’, *Nations and Nationalism*, 11(2), pp. 243–263. Available at: <https://doi.org/10.1111/j.1354-5078.2005.00202.x>.

Han, B., Cook, P. and Baldwin, T. (2012) ‘Geolocation Prediction in Social Media Data by Finding Location Indicative Words’, p. 18.

Han, J. *et al.* (2018) ‘Linking Fine-Grained Locations in User Comments’, *IEEE Transactions on Knowledge and Data Engineering*, 30(1), pp. 59–72. Available at: <https://doi.org/10.1109/TKDE.2017.2758780>.

Hollenstein, L. (2008) ‘Capturing vernacular geography from georeferenced tags’.

Huang, Y. *et al.* (2016) ‘Understanding U.S. Regional linguistic variation with Twitter data analysis’, *Computers, Environment and Urban Systems*, 59, pp. 244–255. Available at: <https://doi.org/10.1016/j.compenvurbsys.2015.12.003>.

Jewell, H.M. (1994) *The North-south Divide: The Origins of Northern Consciousness in England*. Manchester University Press.

Kalantaridis, C. (2010) ‘In-migration, entrepreneurship and ruralurban interdependencies: The case of East Cleveland, North East England’, *Journal of Rural Studies*, 26(4), pp. 418–427. Available at: <https://doi.org/10.1016/j.jrurstud.2010.03.001>.

Kropczynski, J. *et al.* (2018) ‘Identifying Actionable Information on Social Media for Emergency Dispatch’, p. 11.

Lambiotte, R. *et al.* (2008) ‘Geographical dispersal of mobile communication networks’, *Physica A: Statistical Mechanics and its Applications*, 387(21), pp. 5317–5325. Available at: <https://doi.org/10.1016/j.physa.2008.05.014>.

Lengyel, B. *et al.* (2015) ‘Geographies of an Online Social Network’, *PLOS ONE*. Edited by W.-X. Zhou, 10(9), p. e0137248. Available at: <https://doi.org/10.1371/journal.pone.0137248>.

Li, Z. *et al.* (2021) ‘Measuring global multi-scale place connectivity using geotagged social media data’, *Scientific Reports*, 11(1), p. 14694. Available at: <https://doi.org/10.1038/s41598-021-94300-7>.

Llamas, C. (2009) *Language and Identities*. Edinburgh University Press.

Llamas, C. and Watt, D. (2014) ‘Scottish, English, British?: Innovations in Attitude Measurement’, *Language and Linguistics Compass*, 8(11), pp. 610–617. Available at: <https://doi.org/10.1111/lnc3.12109>.

Middleton, C. and Freestone, P. (2008) ‘The Impact of Culture-led Regeneration on Regional Identity in North East England’.

Murphy, N. (2019) ‘Reddit’s 2019 Year in Review - Upvoted’. https://www.redditinc.com/blog/reddits-2019-year-in-review/#content.

Onnela, J.-P. *et al.* (2011) ‘Geographic Constraints on Social Network Groups’, *PLOS ONE*, 6(4). Available at: <https://doi.org/10.1371/journal.pone.0016939>.

Pérez, J.M. *et al.* (2019) ‘Exploiting user-frequency information for mining regionalisms from Social Media texts’. arXiv. Available at: <https://arxiv.org/abs/1907.04492> (Accessed: 23 May 2023).

Purves, R.S., Winter, S. and Kuhn, W. (2019) ‘Places in Information Science’, *Journal of the Association for Information Science and Technology*, 70(11), pp. 1173–1182. Available at: <https://doi.org/10.1002/asi.24194>.

Rae, A. (2009) ‘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(3), pp. 161–178. Available at: <https://doi.org/10.1016/j.compenvurbsys.2009.01.007>.

Ratti, C. *et al.* (2010) ‘Redrawing the Map of Great Britain from a Network of Human Interactions’, *PLoS ONE*. Edited by O. Sporns, 5(12), p. e14248. Available at: <https://doi.org/10.1371/journal.pone.0014248>.

Reades, J., Calabrese, F. and Ratti, C. (2009) ‘Eigenplaces: Analysing cities using the space time structure of the mobile phone network’, *Environment and Planning B: Planning and Design*, 36(5), pp. 824–836. Available at: <https://doi.org/10.1068/b34133t>.

Reimers, N. and Gurevych, I. (2019) ‘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)*. Hong Kong, China: Association for Computational Linguistics, pp. 3980–3990. Available at: <https://doi.org/10.18653/v1/D19-1410>.

Rey, S., Arribas-Bel, D. and Wolf, L.J. (2023) *Geographic data science with python*. CRC Press.

Russ, B. (2012) ‘Examining large-scale regional variation through online geotagged corpora’, in.

Scellato, S. *et al.* (2010) ‘Distance matters: Geo-social metrics for online social networks’, pp. 8–8.

Sobolevsky, S. *et al.* (2013) ‘Delineating Geographical Regions with Networks of Human Interactions in an Extensive Set of Countries’, *PLoS ONE*. Edited by Y. Moreno, 8(12), p. e81707. Available at: <https://doi.org/10.1371/journal.pone.0081707>.

Statista (2022) ‘Most popular social networks worldwide as of January 2022, ranked by number of monthly active users’, *Statista*. https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/.

Stephens, M. and Poorthuis, A. (2015) ‘Follow thy neighbor: Connecting the social and the spatial networks on Twitter’, *Computers, Environment and Urban Systems*, 53, pp. 87–95. Available at: <https://doi.org/10.1016/j.compenvurbsys.2014.07.002>.

Stuart-Smith, J. (2008) ‘Scottish English: Phonology’, *Varieties of English*, 1, pp. 48–70.

Sui, D. and Goodchild, M.F. (2011) ‘The convergence of GIS and social media: Challenges for GIScience’, *International Journal of Geographical Information Science*, 25(11), pp. 1737–1748. Available at: <https://doi.org/10.1080/13658816.2011.604636>.

Takhteyev, Y., Gruzd, A. and Wellman, B. (2012) ‘Geography of Twitter networks’, *Social Networks*, 34(1), pp. 73–81. Available at: <https://doi.org/10.1016/j.socnet.2011.05.006>.

Titheridge, H. *et al.* (2009) ‘Assessing the extent of transport social exclusion among the elderly’, *Journal of Transport and Land Use*, 2(2). Available at: <https://doi.org/10.5198/jtlu.v2i2.44>.

Trudgill, P. (2004) *Dialects*. 2nd edn. London: Routledge. Available at: <https://doi.org/10.4324/9780203314609>.

Wagner, D., Zipf, A. and Westerholt, R. (2020) ‘Place in the GIScience Community an Indicative and Preliminary Systematic Literature Review’. Available at: <https://doi.org/10.5281/zenodo.3628855>.

Waters, T. and Evans, A.J. (2003) ‘Tools for web-based GIS mapping of a "fuzzy" vernacular geography’, p. 10.

Wei, J., Tay, Y., *et al.* (2022) ‘Emergent Abilities of Large Language Models’. arXiv. Available at: <https://doi.org/10.48550/arXiv.2206.07682>.

Wei, J., Bosma, M., *et al.* (2022) ‘Finetuned Language Models Are Zero-Shot Learners’. arXiv. Available at: <https://doi.org/10.48550/arXiv.2109.01652>.

Westerholt, R. *et al.* (2018) ‘Towards the Statistical Analysis and Visualization of Places (Short Paper)’, pp. 7 pages. Available at: <https://doi.org/10.4230/LIPICS.GISCIENCE.2018.63>.

Yin, W. *et al.* (2017) ‘Comparative Study of CNN and RNN for Natural Language Processing’, *arXiv:1702.01923 [cs]* [Preprint]. Available at: <https://arxiv.org/abs/1702.01923> (Accessed: 2 January 2020).

Zheng, X., Han, J. and Sun, A. (2018) ‘A Survey of Location Prediction on Twitter’, *IEEE Transactions on Knowledge and Data Engineering*, 30(9), pp. 1652–1671. Available at: <https://doi.org/10.1109/tkde.2018.2807840>.

Zheng, Y. (2015) ‘Trajectory Data Mining: An Overview’, *ACM Transactions on Intelligent Systems and Technology*, 6(3), pp. 1–41. Available at: <https://doi.org/10.1145/2743025>.

1. https://www.reddit.com/r/unitedkingdom/wiki/british\_subreddits [↑](#footnote-ref-27)
2. https://huggingface.co/cjber/reddit-ner-place\_names [↑](#footnote-ref-28)
3. The highest tier of sub-national division in England. For Scotland and Wales we use the full national extents. [↑](#footnote-ref-31)
4. See ‘Scottish English’ or ‘Scots’; (Stuart-Smith, 2008) [↑](#footnote-ref-34)