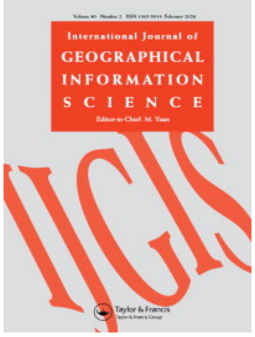


International Journal of Geographical Information Science

ISSN: 1365-8816 (Print) 1362-3087 (Online) Journal homepage: www.tandfonline.com/journals/tgis20

A semantic and context-dependent approach to the interpretation of ‘near’ in historical English Lake District narratives

Erum Haris, Anthony G. Cohn & John G. Stell

To cite this article: Erum Haris, Anthony G. Cohn & John G. Stell (19 Nov 2025): A semantic and context-dependent approach to the interpretation of ‘near’ in historical English Lake District narratives, International Journal of Geographical Information Science, DOI: 10.1080/13658816.2025.2588359

To link to this article: https://doi.org/10.1080/13658816.2025.2588359

Published online: 19 Nov 2025.

INTERNATIONAL JOURNAL OF GEOGRAPHICAL INFORMATION SCIENCE
https://doi.org/10.1080/13658816.2025.2588359

Taylor & Francis
Taylor & Francis Group

RESEARCH ARTICLE

OPEN ACCESS

A semantic and context-dependent approach to the interpretation of 'near' in historical English Lake District narratives

Erum Haris[a], Anthony G. Cohn[a,b] and John G. Stell[a]

[a]School of Computer Science, University of Leeds, United Kingdom; [b]The Alan Turing Institute, United Kingdom

ABSTRACT

A common and intuitive way of identifying the proximity relationship between two entities is to use the preposition 'near' (e.g. 'near (inn, village)'). However, the 'near' relation is vague, often asymmetric and context-dependent, and hence incorporating factors such as the effect of type, size and scale of reference object and associated features is required for cognitive modeling. In this work, we interpreted spatial proximity as described in the historic Corpus of Lake District Writing comprising travel narratives as early as the 16th century. At a time when modern transportation modes were not available, it is interesting to explore how proximity has been perceived and recounted with various context factors explicit or implied. We utilized pre-trained BERT and its variants to first identify the broader semantics of 'near' and generate contextual embeddings. We further identified the contextual factors of spatial nearness. Finally, we used quantitative distances between entities to measure the departure of context-dependent proximities from objective notions of distance.

ARTICLE HISTORY
Received 31 October 2024
Accepted 6 November 2025

KEYWORDS
Spatial narratives; digital humanities; natural language processing; spatial proximity

1. Introduction

The phenomenal spatial turn in the field of humanities has paved the way for scholars to study data from the lens of geographical information systems (GIS). As a consequence, the experimental arena of digital humanities has emerged with promising findings for a variety of historical data, including travel writings and narratives, with textual analysis and visualization being the most prevailing techniques (Murrieta-Flores and Martins 2019). The corpus of Lake District writing (CLDW) (Rayson et al. 2017), a valuable source of leisure accounts from the well-known lake poets and other writers, spans from the 16th to the 19th centuries. The corpus has been subjected to different analyses in prior studies, ranging from the most basic frequency and word-level

CONTACT Erum Haris e.haris@leeds.ac.uk

analysis to coordinate-based analysis, also termed geographical text analysis (GTA) (Donaldson et al. 2017), thematic analysis (Gregory et al. 2024), toponym recognition and geo-parsing (Rayson et al. 2017).

Recently, there has been a growing interest in studying and analysing the spatio-temporal details that appear in the contents of the CLDW (Ezeani et al. 2023, Haris et al. 2024). The corpus contains numerous vague, imprecise and relative geographical references, expressed using a variety of spatial terms representing spatial relations. The fundamental way of organizing such facts is a semantic triples network showing the toponyms as nodes and their specific spatial relation as an edge (Murrieta-Flores et al. 2019). However, beyond relation triples, a deeper analysis of spatial relations, backed by contextual interpretation, is essential for an accurate representation of the corpus content. In this regard, the context of spatial proximity or nearness relation is the core focus of this study.

The concept and sense of spatial proximity highly depend on the semantic analysis of spatial prepositions, as the conveyed sense could either be locative or non-locative. Consider some example instances of 'near' from the CLDW[1]:

> *"The island of Belle Isle, near Bowness, is a point of interest."*
>
> *"With this worthless man, his unhappy lady lived near twenty years."*
>
> *"The annual income of my chapel at present, as near as I can compute it, may amount to about £17, of which is paid in cash."*

The above examples demonstrate varying senses of nearness, where only the first one conveys information about the spatial nearness. This can be considered as a classification on a coarse-grained level. Consider another set of examples below, pointing to various contexts of spatial nearness that will be further discussed later. On a fine-grained level, specific contextual factors (Novel et al. 2020) from the literature help to further assess the indicators or drivers of locative proximity.

> *"After that, we pass a doleful-looking house, which I should have thought deserted but for seeing a child at play near the door."*
>
> *"I went up a small hill near the inn, from whence I had a view of the whole of the lake."*
>
> *"Parton and Harrington, two small sea-ports, are near Moresby."*

From a computational perspective, these nuances of lexical semantics are hard to analyse. With the advent of transfer learning – the practice of leveraging knowledge gained from one task to improve performance on another – pre-trained large language models (LLMs) (Brown et al. 2020) have been extensively applied to various downstream natural language processing (NLP) tasks with minimal fine-tuning. Contextual word embedding incorporates the idea of modeling sentence-level semantics of a word which refers to the context in which a particular word appears in a sentence (Wiedemann et al. 2019). This approach, as opposed to the static word embedding, has significantly improved the ability of LLMs on associated NLP tasks, such as question answering, named entity recognition (NER), text classification and word sense disambiguation (WSD) (Wiedemann et al. 2019, Hadiwinoto et al. 2019). We seek to interpret the linguistic spatial proximity in the CLDW narratives using this approach and address the following research questions:

- Do contextual word embeddings focus on the syntactics or semantics of 'near' in the CLDW, a historical corpus with complex narrative structure?
- On a fine-grained level, which contextual factors appear in the CLDW shaping the sense of locative nearness?

This paper makes three main contributions. We first define a holistic framework for problem understanding by putting the analysis of spatial proximity based on context factors in the perspective of a place-based model. We then propose the exploitation of context-based embeddings to analyze the term 'near', which has an impoverished polysemy (Brenda 2017) compared to other common prepositions studied in the literature. Finally, we present a comparison of linguistically uttered or context-dependent proximity with quantitative distance to analyse the perception of distances in a historical corpus.

The rest of the paper is organized as follows: Section 2 covers background on the literature, the study area, and the CLDW. Section 3 describes our methodological approach. Section 4 contains the experimental findings and results. Section 5 presents a discussion on addressing research questions. Finally, Section 6 concludes the paper with limitations of our study and future directions.

2. Background

This section provides a structured review of the relevant research areas. It begins with an overview of processing historical corpora in digital humanities (Section 2.1), then focuses on the study area and corpus—the English Lake District and the CLDW (Section 2.2), including related work on CLDW processing (Section 2.2.1). The second half reviews the spatial relation 'near' and contextual factors (Section 2.3), covering its various senses (Section 2.3.1), spatial nearness (Section 2.3.2), and contextual interpretation (Section 2.3.3).

2.1. Analysis of historical corpora in digital humanities

Recent research has highlighted the representation of space through text and the various perspectives in which space and place are conceptualized and interrelated. One notable example is the Digital Periegesis project (Foka et al. 2020), which employs semantic geo-annotation to analyze spatial forms within the narratives of Pausanias's second-century CE Periegesis Hellados. After ancient Greek literature was rediscovered following the Enlightenment, these narratives received significant attention as they provided valuable guidance on Greece's classical times and shaped the routes taken by travelers and archaeologists. Foka et al. (2020) utilize different methods within spatio-temporal analysis that include geo-parsing, Linked Open Data (LOD), and network analysis, to rethink the geographies presented in this ancient textual collection, emphasizing topological connections rather than topographic proximity.

In another study (Candela et al. 2023), the authors propose a framework that converts digital collections into LOD. The methodology has been applied to indigenous and Spanish colonial archives, notably the Relaciones Geográficas de Nueva España, a collection containing documents and maps compiled from 1577 to 1585. As a key resource for understanding Early Colonial Mexico and Guatemala, this corpus has become invaluable to historians, archaeologists, and anthropologists studying the region's history. Taking a network analysis approach, the Hestia project (Barker et al. 2015) explores the use of GIS to visualize locations described in Herodotus's fifth-century BCE Histories, examining the application of modern technologies to represent ancient, non-cartographic spatial thinking. The project asserts that network graphs offered a better way to visualize the connections Herodotus made between places and their associated peoples. The resulting visualizations shed light on the narrative's spatial structure, which is based more on relational actions and influence than geographic positioning.

McDonough et al. (2019) argue that from a digital humanities perspective, GTA has been largely applied on contemporary English writings, utilizing lexicons and gazetteers suited to specific time frames for identifying and resolving place names. This poses a limitation for scholars studying the early modern period (1400–1800), often relying on general databases such as Geonames, which introduce historical inaccuracies and ethical concerns. To address these issues, McDonough et al. (2019) use entries from the eighteenth-century Encyclopédie to evaluate rule-based NER systems. They identify areas for improvement in processing historical corpora, particularly through nested and extended annotation of place information.

2.2. The English Lake District and the corpus of Lake District writing

The English Lake District (Figure 1), located in Cumbria in the northwest of England, is renowned for its mountainous terrain and scenic lakes. Covering approximately 2,362 square kilometers, it features a diverse range of landscapes, including deep valleys, rolling hills, and lakes, the largest of which is Lake Windermere[2]. The region's topography is largely shaped by ancient glacial activity, which carved out the U-shaped valleys and left behind the iconic ribbon lakes, such as Ullswater and Derwentwater. The Lake District is also the location of England's highest peak, Scafell Pike[3], standing at 978 meters. Surrounding the lakes are rugged fells (mountains), and the area's geology is complex, with ancient rocks such as slates and volcanic formations from millions of years ago. The region is also rich in biodiversity, with a variety of ecosystems including ancient woodlands, moorlands, and heathlands.

The Lake District, recognized as a UNESCO World Heritage cultural site, owes much of its cultural significance to its rich literary heritage. The combination of water bodies and craggy, weathered peaks creates a landscape of immense natural beauty, which has been a source of inspiration for generations of writers and artists. While the writers in the late 17th century popularize the lake district as a destination for travelers, notable Lake poets including William Wordsworth, Robert Southey and Samuel Taylor Coleridge (Donaldson et al. 2017), celebrated the natural beauty of the area and simultaneously impacting its cultural evolution in the late 18th and beginning of the 19th centuries.

The CLDW has been assembled by Lancaster University's Spatial Humanities project, and contains 80 texts from 1622 to 1900. The corpus spans major literary periods

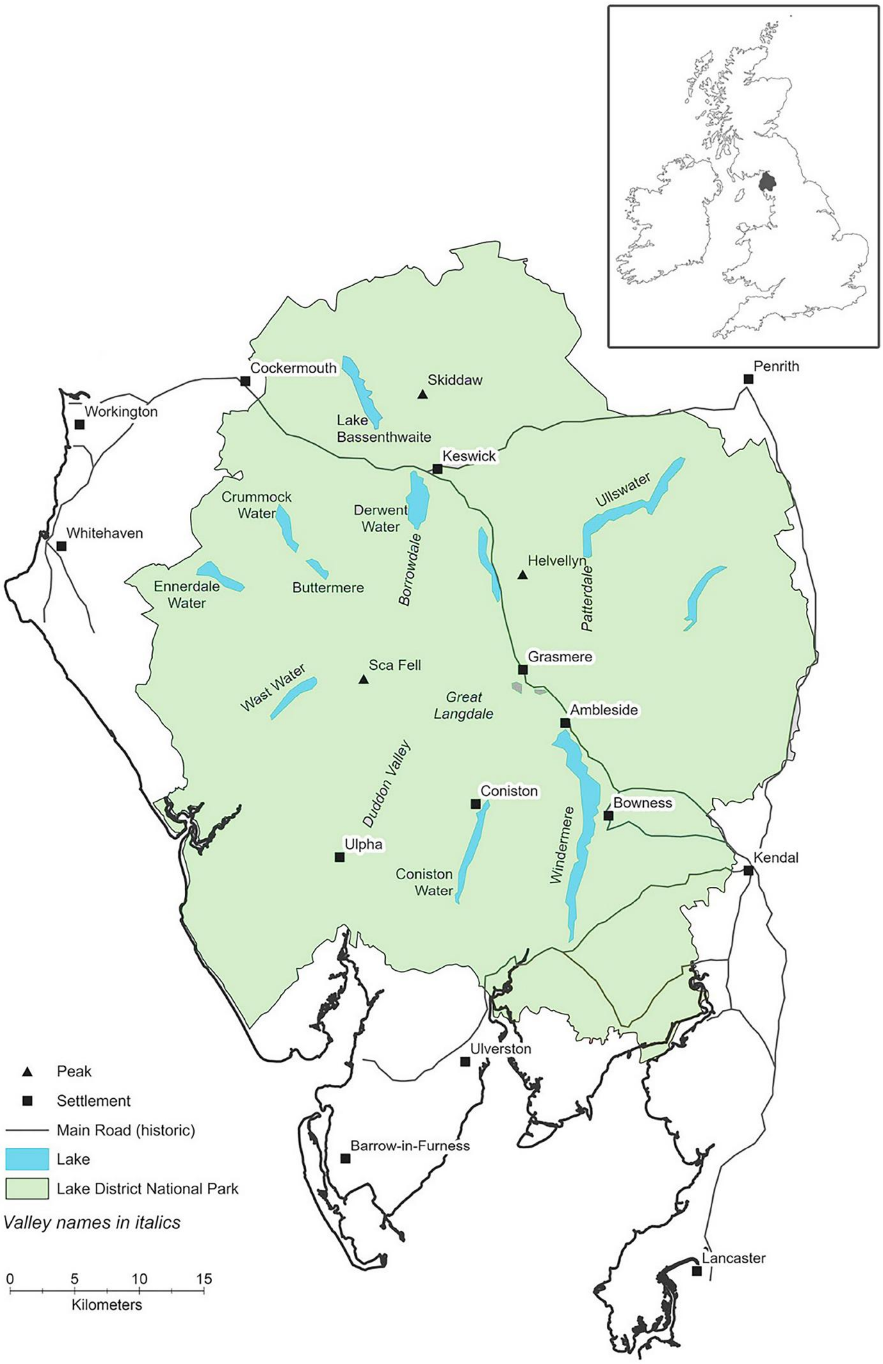


Figure 1. The English Lake District (Gregory et al. 2024).

(Donaldson et al. 2017), including the Age of Sensibility (1740–1788), the Romantic period (1789–1836), and the Victorian era (1837–1901), each period significantly shaped contemporary perceptions of the Lake District. The corpus encompasses a variety of genres such as guidebooks, travelogues, and topographical literature, providing a rich, multilayered view of the area. The texts include works by prominent figures like William Wordsworth and Beatrix Potter, as well as lesser-known authors and anonymous publications. This vast collection of over 1.5 million words is georeferenced and tagged using the Edinburgh Geoparser, an automated tool that identifies toponyms, thus allowing locations in the texts to be linked to physical places in a place-name gazetteer (Rayson et al. 2017).

2.2.1. A Review of existing studies

The CLDW offers a comprehensive view of how the Lake District evolved as both a physical and cultural landscape. Through its blend of famous and obscure works, it reflects the complex interplay of nature, literature, and tourism that defined the region's development across centuries. These aspects have remained a center of attention for researchers. Donaldson et al. (2017) discuss an interdisciplinary approach, termed GTA, to examine the connection of CLDW literature with aesthetic aspects and physical geography of the region. The study uses a combination of corpus linguistics techniques, including automated geoparsing and GIS to analyze how key aesthetic terms 'beautiful', 'picturesque', 'sublime' and 'majestic' were used in describing the region, showing a link between aesthetic conceptions of the 18th century and the language used in the CLDW. Chesnokova et al. (2019) focus on references to silence and tranquility in the CLDW narratives and contemporary Geograph Project data. Through sentiment analysis, they reveal that mentions of silence and peaceful sounds tended to have more positive connotations than random text excerpts and mapping these historical locations emphasizes the lasting influence of authors, such as Wordsworth, on the portrayal of serenity in the Lake District.

Butler et al. (2017) address the significant challenge of recognizing historical place names due to spelling variations and non-standardized naming conventions in CLDW, where the spatial connection to the natural landscape is critical for understanding key literary ideas. The authors argue that using geo-historical datasets not only enhances the corpus but could also serve as a broader resource for identifying onomastic variations, potentially enriching other geospatial resources. Similarly, Rayson et al. (2017) emphasize the importance of expanding open-access toponym corpora and developed the annotated CLDW, a deeply labeled corpus. This resource surpasses the capabilities of standard NER, disambiguation, and geoparsing tools by offering a more comprehensive annotation scheme that effectively links historical and spelling variants of place names, while also distinguishing between different geographical entities using nineteen geospatial categories.

Recent methods for analyzing CLDW narratives from a GIS perspective have extended beyond conventional coordinate based geographical analysis. Gregory et al. (2024) expand toponym references to include 'geo-nouns', which relate places to geographic features. Additionally, the study analyzes adjectives, nouns, and verbs that authors associate with specific locations to capture the 'sense of place' in the texts, providing a more nuanced understanding of places by incorporating both geographical features and the way these places are described in the literature. Steiner et al. (2023) introduce the concept of 'spatio-textual regions', defined as clusters of toponyms within a contiguous section of text describing those places. Their application of spatial clustering to the CLDW reveal eight major clusters, as well as an 'outside region' indicating locations beyond the Lake District. Haris et al. (2023) propose the use of qualitative spatial representation (QSR) to extract and interpret spatial relationships from the CLDW, enabling spatial reasoning to uncover new knowledge. Along this line, Ezeani et al. (2023) develop an extensible framework that incorporates NLP, QSR, and visual analytics for the analysis of spatial narratives. The framework focuses on extracting, analyzing, and visualizing 'location, locale, and sense of place' described in the texts. Using an 'Extractor' that draws upon standard NLP entity extraction libraries, they identify key spatial elements in the CLDW. In another work, Haris et al. (2024) integrate spatial information theory with a structured approach to analyze the CLDW. The framework builds a conceptual foundation using spatial ontology and a custom gazetteer (Rayson et al. 2017) for a semantic exploration of the corpus. They assess three LLMs, evaluating their ability to extract spatial relations from a historical archive. The framework is further strengthened by enhancing the extracted triples with gazetteer-based object classification, enriching spatial profiling of the network. These studies provide a foundation for processing spatial narratives in the CLDW. Building on this, the present work applies contextualized BERT embeddings to analyze the spatial preposition near, identifying context-sensitive meanings and comparing them to geographic distances. This approach offers a more nuanced account of spatial language use in historical texts and demonstrates how modern language models can support deeper semantic interpretation in literary and cultural geography.

2.3. Semantics of 'near' and the role of context

2.3.1. Senses of 'near'

Spatial prepositions define the search domain for locating an object specified in a linguistic expression (Brenda 2017). For the simplest form of spatial linguistic structure 'x preposition y', where x is a trajector (TR) and y is a landmark (LM) (also referred to as locatum (LO) and relatum (RO), respectively), Miller and Johnson-Laird (1976) define two distinct search domains, where the first search domain is used to locate the LM, while the second search domain acts as a subdomain and is established based on the spatial relationship implied by the preposition along with properties specific to the LM. In the case of the preposition 'near', the search domain naturally extends to encompass the surrounding region of the LM, forming a localized area of proximity based on the spatial relation (Brenda 2017). From a psychometric perspective, human construal and interpretation of a spatial scene involves the role of certain facets, namely specificity, prominence, and perspective (Langacker 2008) as they are responsible for activating different spatial conceptual domains within the human mind, that is to say, spatial objects, spatial relations and notion of nearness. The preposition 'near' encapsulates this proximity schema, where the focus is on the relative closeness of two entities, with TR being more central in the spatial relationship (Fabregat 2022).

Examining the semantics of 'near' in depth, Brenda (2017) studies the polysemy of the preposition 'near' and identifies five regions with greater semantic density compared to other senses: 'in the vicinity', 'interaction', 'temporal', 'approaching', further categorized as 'approximately'; each sense either reflects a distinct geometric positioning of TR and LM or a unique metaphorical connotation. The primary sense of 'near' encodes the geographic proximity between a TR and a LM (Tyler and Evans 2003; Fabregat 2022), where the observer adopts a specific construal, selectively focusing on aspects of the scene to communicate proximity as mentioned above. Referring back to the description of search domains, the primary sense, named 'in the vicinity', points to the scenario where TR is to be searched in the region around the LM. Hence, various spatial objects are generally used to localize the surrounding area of LM and narrow down the search domain to locate the TR. These spatial objects act as markers and include named entities, such as human settlements of varying demography and geographic nouns such as mountains, rivers, lakes and forests. With regards to the non-spatial use cases of 'near', this transformation occurs when the core spatial meaning of 'near' metaphorically shifts to abstract domains such as emotion, illustrating how language uses spatial constructs to describe feelings and relationships. The original near-far schema, which denotes physical proximity, now exhibits emotional contexts, where closeness is often associated with a sense of intimacy and friendliness (Brenda 2017).

2.3.2. The spatial 'near'

In a spatial sense, the preposition 'near' holds topological spatial relation properties and maintains this relation independent of the precise distance between two spatial objects, as defined by Euclidean metric (Coventry and Garrod 2004). Owing to this topological nature, 'near' also does not get influenced by an observer's frame of reference, standpoint or object's orientation. Instead, it encodes spatial relations from a canonical, deictically neutral position, reflecting a construal detached from specific viewpoints (Langacker 2008). All these characteristics add to the generalization capability and variable behavior of preposition 'near'; as a consequence Zwarts (2017) notes that in some instances, we find 'near' functions as preposition 'in', 'on', 'at' which are locative or topological prepositions, while in other instances, we find 'near' functions as prepositions 'behind', 'above' which are used to specify distances. Novel et al. (2020) suggest that this variable functioning of 'near' can be categorized into two types depending upon the underlying proximity question to be addressed, that is to say, whether the textual description of 'near' poses a question for which 'near' specifies a region or a distance. In the first scenario, the spatial objects are considered as regions associated topologically and independent of directional context, whereas the second scenario is contrastive in which distance is evaluated relative to the objects' orientation.

Nevertheless, in any case, the preposition 'near' encodes a level of spatial detail to express proximity (Brenda 2017) and this leads us to define proximity from contextual perspectives. Langacker (2008) asserts that while the scope of 'near' is inherently constrained, it can be flexibly adapted to different distances and thus, to spatial scenes at different scales as well. Contrary to this simplification, Burigo and Coventry (2010) state that certain factors do affect the selection of linguistic terms used to define a spatial relation that include the size of spatial objects and the extent of physical space in which objects are placed. Hence, it depends on the objects' size that the term 'near' is replaced with 'far' when we describe relative distance of two toy cars placed one meter apart as opposed to massive objects for which one-meter distance may signal a menacing situation. Similarly, changing the scale of proximity interpretation will influence the usage of 'near' while keeping the objects' size same. The functional geometric framework proposed by Coventry and Garrod (2004) further supports this fact, stating that the physical and functional properties directly influence spatial 'near', such as the relevance of objects' sizes in a proximity relation, which means that objects of comparable sizes are associated as 'near' to each other. Similarly, the static object in a proximity relation is used as the LM or reference and movable object's location is defined with respect to it. Finally, the underlying purpose of nearness association is another relevant factor that affects the interpretation of 'near'. With regards to symmetry, the 'near' relation between static spatial objects of comparable sizes will always be symmetric, in other words, if *'a is near b'* then the opposite also holds. When one of the objects is significantly bigger, it will serve as the LM or reference and provide search domain to locate the other object.

2.3.3. Nearness in context

Dey (2001) describes the notion of 'context' as the knowledge available to characterize the circumstances or factors associated with an entity. Considering the conceptual understanding of 'near', there are various linguistic terms synonymously used to define nearness, as discussed in Section 2.3.2 with examples. Hence, in general, 'near' is a nuanced and vague notion, requires interpretation beyond topological or projective conceptualization (McKenzie and Hu 2017). This usually necessitates the acquisition of supplementary knowledge to characterize the actual concept of nearness in a given situation (Novel et al. 2020); the supplementary knowledge is essentially the contextual factors that influence the interpretation of nearness in different contexts. If we specifically consider the linguistic term 'near', the inherent vagueness can be perceived in several ways, such as 'near' does not provide indication of intent and mode of travel (for example, walking or flying to the reference location). It also does not explicitly inform about aspects related to the reference location, such as geometrical (for example, border or center) or topological aspects (for example, top or bottom) or the path linking the observer to the reference location (McKenzie and Hu 2017). Some notable studies in Geographic Information Science (GIScience) have focused on exploring the structured association between the concept of 'near' and the role of context in its interpretation, recent works include studies conducted by Brennan and Martin (2012), Minock and Mollevik (2013), Xu and Klippel (2013), Hahn et al. (2016), Stock and Hall (2018), and Novel et al. (2020). It should be noted that context not only provides added or implied information on the factors related to the perceived proximity but also on the perceiver's personal attributes (Yao and Thill 2005). Hence, the question of choosing context factors generally boils down to selection based on these two broad aspects as they are fundamental ways in which human beings think and reason about distance and proximity.

Moving on to explicitly recognizing these context factors, psychometric studies about human interpretation of proximity provide preliminary foundations; their results exhibit that the association between physical distance and perceived nearness in observer's mental map is influenced by various intricate spatial aspects related to the objects participating in a proximity relation, such as their relative positions and the extent of physical space involved (Yao and Thill 2005). This fact led Gahegan (1995) to recognize certain context factors that are meaningful in reasoning about physical and perceived proximity. These factors include the scale or geographic span of the region or object along with the size, the derived attractiveness based on the type of object, the spatial arrangement of objects and reachability, say in a transportation network scenario. Yao and Thill (2005) add another context factor to proximity analysis that is the activity linked with intended travel. In essence, it is the motivation for the kind of activity associated with a certain object that is in play behind the mental processing and evaluation of physical distance. This implicitly encompasses the scale factor as well, in a sense that human mind can instinctively adjust the physical span in their thought process when informed about activities of different nature and types. Brennan and Martin (2012) also point out the significant relevance of object size in nearness interpretation, while Novel et al. (2020) further draw attention to transportation and demographic information for contextualizing proximity. They suggest to consider different modes of travel, such as walking or using vehicle, and typical higher-level governing structures of urban and rural areas in the case of demographic aspects. Stock and Hall (2018) use a layered approach to classify context factors into six broad categories, each category can be considered as a layer and organized as entities interacting in a cohesive framework. These categories cover a range of context factors connected with the object environment, the observer with a focus on intended objectives and activities, the particular spatial relation and finally the object and its related spatial features.

3. Proposed approach

3.1. Conceptualizing contextual factors in a place-based model

As this study lies at the intersection of spatial humanities, GIS and NLP, it is important to link the geographical analysis concepts to the theoretical domain, which refers to the description of a place-based model and conceptualizing the notion of context factor in it. This approach allows us to present a comprehensive formal framework for the applied domain i.e., geographical analysis of spatial narratives. With regards to the CLDW, some peculiar themes as described below, stand out when exploring how contextual factors can influence spatial proximity interpretation in the Lake District travel writing.

3.1.1. Environmental and functional context

3.1.1.1. Descriptions of the landscape. The geography of the Lake District, with its lakes, valleys, and mountains, features prominently in the CLDW. These descriptions provide a sense of spatial scale and proximity, such as the compactness of valleys or the expansiveness of the fells. For example, the proximity of water bodies such as *'Windermere'* or *'Derwentwater'* to the surrounding mountains can create a feeling of enclosed beauty. Similarly, the role of the Lake District as a tourist destination influences descriptions of spatial proximity.

3.1.1.2. Seasonal variations. The Lake District experiences significant seasonal changes that are often noted in travel writings. Descriptions of the region during different seasons can alter perceptions of space, such as the lush, dense foliage of summer versus the bare, open landscapes of winter.

3.1.1.3. Walking paths. The network of trails and footpaths in the Lake District often appears in travel narratives, emphasizing physical proximity and accessibility. Descriptions of popular routes such as *'the Cumbria Way'* or *'the ascent of Scafell Pike'* provide insight into how proximity is experienced through movement and journey.

3.1.2. Narrative and descriptive style

3.1.2.1. Detailed observations. Many travel writers offer meticulous descriptions of the Lake District's flora, fauna, and geographical features, which can create a vivid sense of spatial intimacy and immediacy. Phrases such as *'overlooking Derwentwater'*, *'nestled in a close-knit village'* versus *'secluded on a remote island'*, convey different senses of proximity without using a specific preposition (near, close etc.).

3.1.2.2. Literary references and poetic language. The Lake District is closely associated with the Romantic poets, such as William Wordsworth. Writings that reference these poets often highlight a cultural context where proximity to nature is highly valued, influencing how readers perceive the closeness of natural elements such as lakes, mountains, and wood. The association of the Lake District with Romantic poetry influences the way travel writers describe the region. Poetic language can enhance the sense of proximity to natural beauty, creating a deeper connection with the landscape.

Figure 2 presents a place facets-based model developed on the findings of Hamzei et al. (2020) where a narrative description is composed of four different types of semantic relations as described linguistically. The circles denoting the physical and functional semantic relations correspond to the above defined environment and functional context, while the circle showing emotive relation is related to the narrative and descriptive style. Finally, the circle for spatial relation is the proximity relation under analysis, the relation 'near'. Table 1 attempts to integrate the issues discussed in this section. It should be noted that though the category of literary references and poetic language plays an important role in shaping the perception of nearness as described, since the nearness is expressed using subjective linguistic terms, such statements are not part of the analysis described in the paper.

3.2. Framework

The overall framework consists of three main stages as depicted in Figure 3. In the first stage, a collection of raw text containing sentences with instances of 'near' in CLDW is passed to the pre-trained BERT model to generate contextual embeddings for the word 'near'. This stage further analyses the resulting embeddings with clustering and dimensionality reduction for visualization. It also determines k-NN (k-nearest neighbors) for input sentences from the resultant embeddings. Once we obtain a set

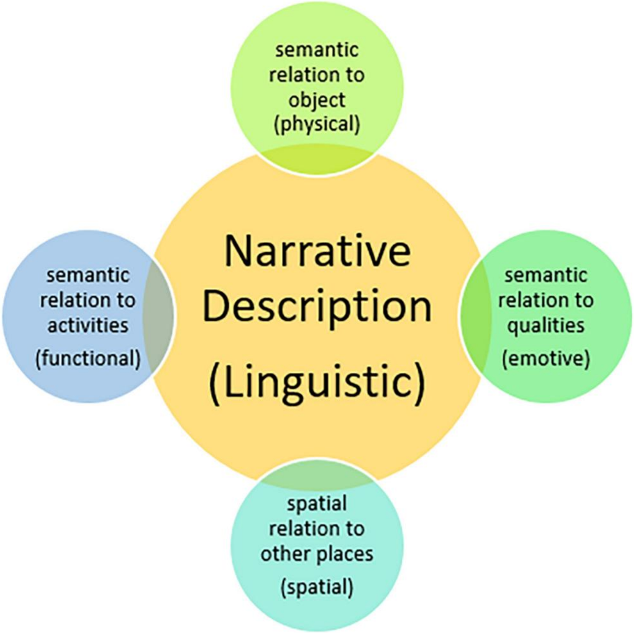

semantic relation to object (physical)
semantic relation to activities (functional)
Narrative Description (Linguistic)
semantic relation to qualities (emotive)
spatial relation to other places (spatial)

Figure 2. A place facets-based model for spatial narratives analysis.

Table 1. Selected context factors (Gahegan 1995, Yao and Thill 2005, Brennan and Martin 2012, Xu and Klippel 2013, Stock and Hall 2018, Novel et al. 2020).

Context factor	Association	Place-based model
Type	Object	Physical
Scale	Object	Physical
Size	Object	Physical
Feature/sub-part	Object	Physical
Geography/demography	Environment/Object	Physical
Means of travel	Object	Functional
Activity	Object/Observer/Audience	Functional

of sentences with spatial nearness instances, the second stage activates with the identification of selected context factors of spatial proximity. The final result is a list of triples with extended information about features or activity if any, grouped by the relevant context factors. The LO and RO geographic coordinates are determined and used to calculate absolute and relative distances.

3.2.1. Contextual word embeddings for analyzing semantics of 'near'

Word embeddings refer to fixed-dimensional vector representations of words, where each word is assigned a distinct vector based on its usage in large corpora. Traditional word embeddings methods, namely Word2Vec (Mikolov et al. 2013) and GloVe (Pennington et al. 2014), generate a single static embedding per word. These embeddings capture semantic similarities by placing words with similar meanings closer together in the vector space, but they fail to differentiate between different senses of the same word. Contextual word embeddings (Wiedemann et al. 2019) dynamically adjust word representations based on the surrounding context, overcoming limitations of static embeddings.

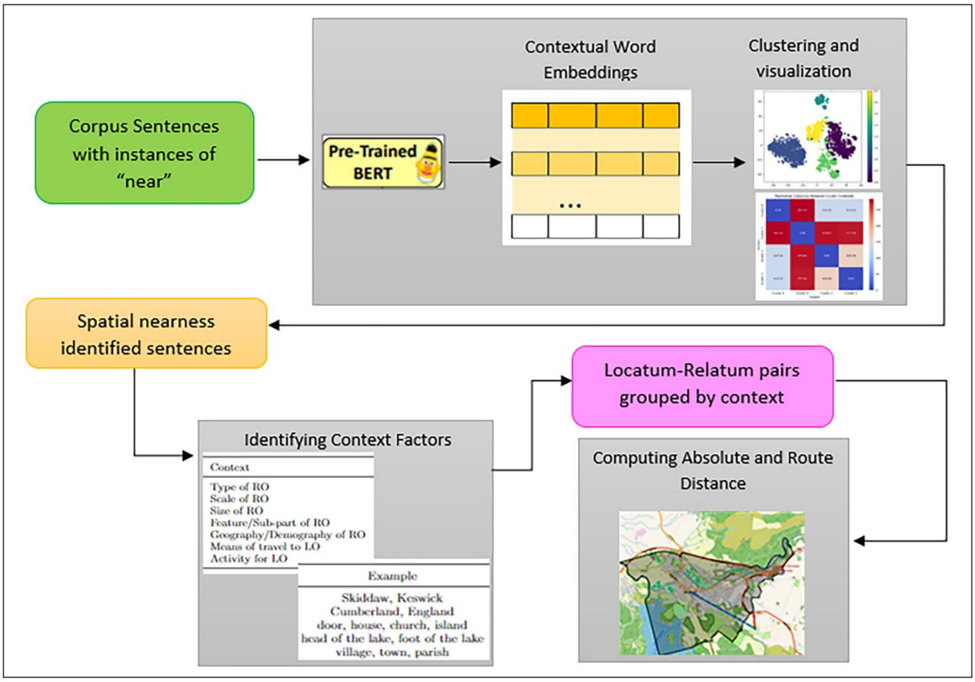

Corpus Sentences with instances of "near"
Pre-Trained BERT
Contextual Word Embeddings
...
Clustering and visualization
Spatial nearness identified sentences
Identifying Context Factors
Context
Type of RO
Scale of RO
Size of RO
Feature/Sub-part of RO
Geography/Demography of RO
Means of travel to LO
Activity for LO
Example
Skiddaw, Keswick
Cumberland, England
door, house, church, island
head of the lake, foot of the lake
village, town, parish
Locatum-Relatum pairs grouped by context
Computing Absolute and Route Distance

Figure 3. The Overall framework.

BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al. 2019) generates contextual word embeddings using a bidirectional transformer architecture that enables it to consider the left as well as the right context of a word simultaneously, capturing the detailed semantic meaning of words in context, which is crucial for tasks involving polysemy and semantic nuances. For instance, in the phrases *'He went to the bank to withdraw money'* versus *'The river bank was eroded'*, BERT can generate different contextual embeddings for 'bank' because it processes both preceding and succeeding tokens. However, unlike a noun, such as the word 'bank' with several meanings, the basic sense of 'near' whether used spatially or non-spatially (e.g. *'near the ocean'* or *'near completion'*), indicates closeness, making its disambiguation a challenging task.

With regards to the process of fine-tuning, recent NLP research increasingly advocates for fine-tuning large Transformer-based architectures used in supervised learning tasks. Nonetheless, Mayfield and Black (2020) argue that there exists a trade-off between performance achieved via fine-tuning and resource utilization, that is to say, fine-tuned BERT performs comparatively similar to conventional models, however, at a considerably higher computational cost. Peters et al. (2019) propose text processing with an untuned BERT model by directly using the [CLS] token's final activations as contextual embeddings. Similarly, Nadeem et al. (2019) employ an untuned BERT model to harness its embedded world knowledge, eliminating the need for fine-tuning and making it suitable for CPU-based predictions. We have followed a similar approach to directly utilize the contextual embedding for clustering which is an unsupervised task, with different variants of BERT and final embedding representations or heuristics.

3.2.2. Identification of context factors of spatial 'near'

Having developed the clusters of similar senses, the next task is to select the sentences relevant to spatial nearness and extract fine-grained information on contextual factors as shown in Table 1, to further determine the drivers of proximity. They can be related to different parts of a sentence structure, namely 'subject, relation, object and object features'; these distinct arguments of a sentence are equivalent to the various layered categories in earlier defined model (Section 2.3.3) proposed by Stock and Hall (2018). They can also be related to other roles in a place-human-narrative trio (Wolter and Yousaf 2018), such as 'environment, observer or audience' (Stock and Hall 2018).

Novel et al. (2020) explicate the distinct semantic properties of textual descriptions about places, these properties provide a sound outline to frame the level of information extraction applicable here. The first semantic property classifies these textual accounts based on whether they are actually describing a place or identifying it. As we previously discussed in Section 2.3.2, expressions involving nearness may present in two contrasting forms: the first form is the description that addresses the question of 'where'; conversely the second form is the description about identifying a reference location or addressing the question of 'which'. In our case, the second form of description about places is primarily important to resolve linguistic references of 'near'.

Given the knowledge about the type of textual description to use, the second semantic property elucidates the different types of spatial entities that are used in conjunction with 'near' to collectively form an expression of nearness in a textual description about a place. These spatial entities usually denote specific toponyms. They can also be any geographic nouns which will be counted as spatial references when used as noun phrases (NPs). In a near-specific text, these toponyms and geographic NPs can be accompanied with other prepositions informing about their relative location. Our task consists of identifying the term 'near' and the arguments of the nearness relation (LO, RO and LO/RO attributes) that collectively form a spatial relation triple. The approach builds on existing NER and relation extraction frameworks for CLDW (Ezeani et al. 2023, Haris et al. 2024). We extend the triples extraction method to determine the context factors of Table 1 and incorporate associated features and activities as identified from the CLDW annotated corpus in the NER system.

Finally, the third semantic property highlights a subtle yet significant difference between textual descriptions that are from a particular location or about a particular location. In the first case, the location or reference object can be identified through metadata associated with the text, whereas in the latter case, thematic analysis is usually needed to deduce knowledge about the reference object. For the CLDW, the available meta-data and customized gazetteer entries provide information on the source location of 80 narrative files and contained toponym entries respectively (Rayson et al. 2017), which facilitate in defining or resolving the geographic scope of spatial entities and discerning the origin of text file from reference object.

3.2.3. Comparing narrative distances

A shift from the linguistic to quantitative realization of proximity relations is needed for a comparative analysis of perceived vs. actual distance. Quantitative spatial relationships require different mathematical metrics to be computed so one can depict them on measurable scales, such as distance or time required to reach destination.

Usually, these metrics are absolute, such as Euclidean distance; at other times it can be travel time (Novel et al. 2020). The representation of spatial objects involved in a relation also requires transformation, generally into a geometric object that can be described on a Cartesian plane with precise spatial measurements. Generally, a point, a line or a polygon is used for geometric representations and several metrics based on absolute or relative aspects, can be used to compute the distance between these objects, with either center or boundary as reference. This takes us to define the underlying process to achieve the transformation, called geocoding, which translates toponyms into precise geographic coordinates; this translation from linguistic elements to geographic references is essentially represented as a geometric object.

However, geocoding itself depends on the availability of gazetteers that contain location information, such as the name of location along with its semantic category or type and geographic coordinates. Several free and collaborative platforms exist for this purpose providing access to the global geographic data. OpenStreetMap (OSM) is one of the most commonly used crowdsourced platforms for accessing geographic data. It provides an API, called Overpass API with a custom query language called, OSM QL (Olbricht 2015) that lets us query the OSM database and retrieve the required information. The OSM schema represents geographic features through distinct elements: *nodes* represent point objects in space, *ways* represent linear features, *areas* represent boundaries as polygon objects, and *relations* define the association among these objects. These elements carry attributes, known as *tags*, in key-value pairs. For example, '*boundary = administrative*' marks an administrative boundary; this demonstrates OSM's adabtable approach to capturing and querying geographic data. Figure 4 shows an example of a query to extract lake Windermere's boundary, a well-known point of interest in the Lake region.

Here, the task is to determine both absolute distance (AD) and route distance (RD) for one of the two types of relative references (as pointed in Section 3.2.2) defined below:

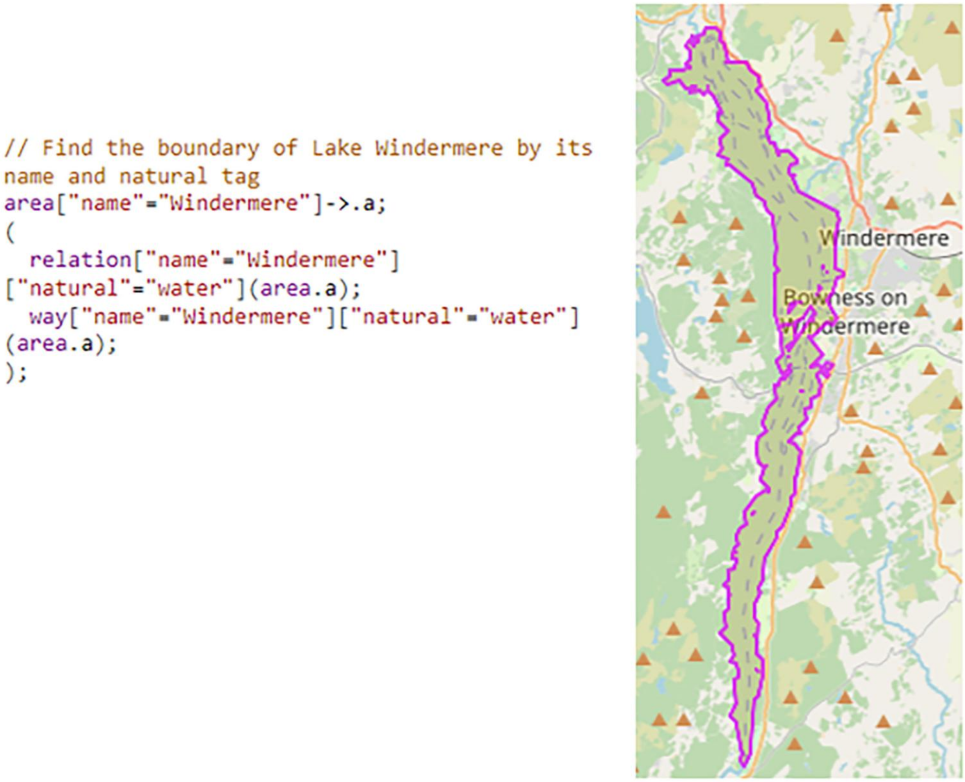
- spatially relative references with measurable AD and RD - definite toponyms locatable on a map.
- spatially relative references with non-measurable AD and RD - provides information on other aspects related to LO and RO at feature level (such as lake features, for example: '*Windermere*' is a lake with '*a curve near its centre*').

For AD, the focus is on determining distances from source to destination using Euclidean distance between two objects (LO and RO). For RD, the selected travel metric is the walking distance between two objects (LO and RO), as we are interpreting historical writings with fewer references to old modes of travel, comparing narrative perception of nearness in terms of walking distance using contemporary maps and distance tools seems a reasonable choice.

4. Experiments and results

4.1. Contextual word embeddings, clustering and visualization

We used three different BERT architectures for computing contextual embeddings, *bert-base-uncased*, *roberta-base* and *distilbert-base-uncased*. RoBERTa is an optimized version



```
// Find the boundary of Lake Windermere by its
name and natural tag
area["name"="Windermere"]->.a;
(
  relation["name"="Windermere"]
["natural"="water"](area.a);
  way["name"="Windermere"]["natural"="water"]
(area.a);
);
```

Figure 4. Overpass query to extract lake Windermere boundary.

Table 2. Experiment attributes summary for generating contextual word embeddings.

Attribute	Value
No. of words in selected text	90,236
Pre-trained models used	*bert-base-uncased*, *roberta-base*, *distilbert-base-uncased*
Heuristics for hidden layer	second-to-last, sum of all, sum of last four, concatenation of last four
Post-processing	k-means clustering, t-SNE visualization

of BERT which has been trained on more data with larger batches for improved accuracy (Liu et al. 2019), whereas DistilBERT is a faster and efficient version of BERT trained via knowledge distillation while maintaining BERT's performance (Sanh et al. 2020). Table 2 provides information on the dataset size, BERT variants and heuristics used for the selection of hidden layers. A hidden state h^l, where $l = 1,2 \ldots ,L$ is a representation of each token, where each layer's output integrates syntactic or semantic information in a hierarchy. Lower layers capture syntactic features (such as part-of-speech (POS) information), while upper layers encode more complex, semantic relationships. The process initiates with tokenization where the tokenizer converts sentences into token IDs, which are processed by BERT to generate embeddings for each token. The self-attention mechanism in each transformer layer uses attention heads, which dynamically adjust the significance of each token in relation to others. This process results in unique embeddings at each layer that encode the evolving context (of 'near') as the input passes through the BERT layers. The contextual embedding for 'near' are then extracted by identifying its position in the tokenized sentence. If 'near' is not present, it uses the embedding for [CLS] token as a fallback. For each token in a sentence, h^l is represented as a matrix of size [n,d] which is a token embedding matrix for layer l where n is the number of tokens in the sentence (including padding), and d is the embedding dimension.

We experimented with different combinations of hidden layers for the final embeddings extraction, noted in the literature (Devlin et al. 2019, Wiedemann et al. 2019) namely *second-to-last hidden layer*, *sum of last four hidden layers*, *sum of all hidden layers* and *concatenation of last four hidden layers*. For instance, the last approach sums the hidden state h^l across all L layers for the token of interest. This embedding can be useful for capturing both lower-level syntactic and higher-level semantic information, given by Eq. (1):

$$E_{\text{near}} = \sum_{l=1}^{L} h_{near}^{(l)} \tag{1}$$

where $h_{near}^{(l)}$ denote the embedding of token 'near' at hidden state l.

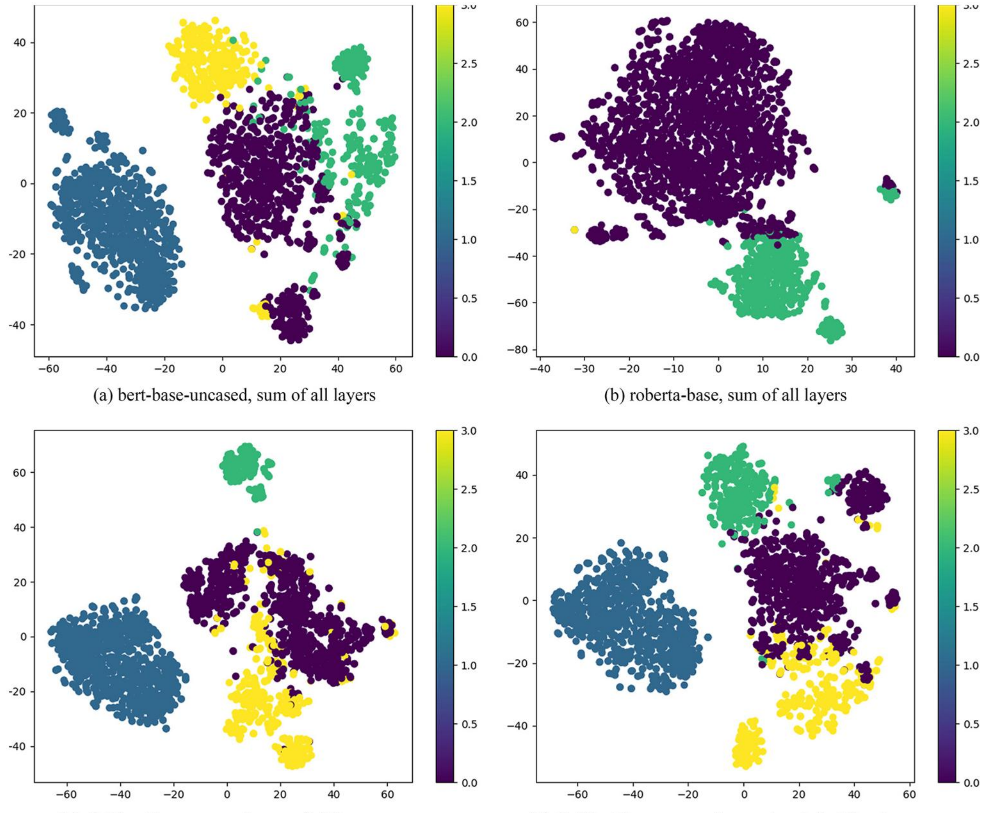
The final contextual embeddings are clustered using k-means, grouping sentences based on the similarity of their near-related embeddings. The k-means clustering minimizes the within-cluster sum of squared distances between embeddings, which facilitates the grouping of similar contexts. To ensure reproducibility, a fixed random seed `random_state = 42` was used. For visualization, t-SNE (t-distributed stochastic neighbor embedding) helps map these high-dimensional embeddings into two dimensions while preserving local similarity patterns, enhancing interpretability of clusters. Moreover, to assess the cluster cohesion along with separation, we computed the average Silhouette score for which the intra-cluster distance and mean nearest-cluster distance are measured per sample. Figure 5 illustrates contextual embedding clusters for the four embedding outputs based on the average Silhouette scores in Table 3 that lists the best performing model-layer heuristic configuration and cluster counts, as increasing the number of clusters started reducing the average Silhouette score.

4.2. Selection and evaluation of contextual embedding clusters

After performing a careful qualitative analyses of resultant clusters, both manually and with the help of various methods including frequency of named-entities and geonouns, keywords analysis, thematic analysis and word cloud visualization, we selected one of the reasonable set of embeddings, *distilbert-base-uncased, sum of all layers* for the next stage. Among such analyses, we have reported the results of adopting the k-NN in an unsupervised setting to determine similarity of the corpus sentences with the generated contextual embeddings. Appendix A, Table A1 displays a subset of result for selected embeddings, three sentences with k = 10 nearest neighbors. The observations on embeddings, selection of clustering output, k-NN results and other related arguments have been described in Section 5.

4.2.1. Annotation-based evaluation

To precisely report the quality and interpretability of BERT-derived semantic clusters for nearness sense classification, we conducted quantitative evaluation of the selected



(a) bert-base-uncased, sum of all layers

(b) roberta-base, sum of all layers

(c) distilbert-base-uncased, sum of all layers

(d) distilbert-base-uncased, concatenate last four layers

Figure 5. t-SNE plots of contextual embeddings for 'near' across models and heuristics, showing four k-means clusters $c = 4$.

Table 3. Average Silhouette scores of embedding clusters.

	Average Silhouette Score		
Pre-trained model with heuristic	c = 3	c = 4	c = 5
bert-base-uncased			
sum of all layers	0.3620	0.3233	0.3271
concatenate last four layers	0.2668	0.2292	0.2209
roberta-base			
sum of all layers	0.4370	0.4368	0.1988
concatenate last four layers	0.3988	0.3225	0.1301
distilbert-base-uncased			
sum of all layers	0.3173	0.3479	0.3484
concatenate last four layers	0.2743	0.2759	0.2786

set of embedding *distilbert-base-uncased, sum of all layers* by carefully annotating a small sample for gold-standard dataset. We selected clusters 0, 2 and 3 out of four resultant clusters, the reason for excluding cluster 1 is described in Section 5. The annotation set comprised 130 sentences, balanced across clusters' relevant sizes, each sentence labeled by a domain expert with one of several fine-grained categories: Spatial–Toponymic, Spatial–Non-Toponymic, Non-Spatial (Interaction), Non-Spatial (Approaching), Non-Spatial (Approximating), and Non-Spatial (Temporal). Table 4 summarizes the annotation statistics.

Table 4. Distribution of categories across clusters.

Cluster	Spatial–toponym	Spatial–non-toponym	Non-spatial: Interact	Non-spatial: approach	Non-spatial: approx	Non-spatial: temporal	Total
0	24	35	1	0	0	0	60
2	33	1	1	0	0	0	35
3	1	18	7	4	4	1	35
Total	58	54	9	4	4	1	130

Table 5. Performance metric results.

Class	Precision	Recall	F-1
Spatial–toponymic	0.943	0.569	0.711
Spatial–non-toponymic	0.583	0.648	0.614
Non-spatial	0.457	0.889	0.604

Toponymic judgments (e.g. 'near Keswick', 'near Penrith') were made by cross-checking all named entities against a custom gazetteer (Rayson et al. 2017). If 'near' referred to an object, part, or entity not in the gazetteer but still spatial (e.g. 'near the summit', 'near the house'), it was labeled Spatial–Non-Toponymic. Non-spatial and metaphorical uses (e.g. 'near five years', 'near the gates of death') were identified through close reading and attention to figurative, emotional, or related senses. Borderline and ambiguous cases were discussed in detail and marked as 'Ambiguous' if no clear label emerged. Where sentences contained multiple senses (e.g. two 'near' expressions), each instance was considered in context, and the dominant or most relevant sense was annotated.

4.2.2. Evaluation metrics

For evaluation, each cluster was assigned a predicted label corresponding to the majority gold annotation within that cluster. Thus, clusters acted as the system's predictions, while the expert annotation served as ground truth. For reporting and metric aggregation, non-spatial subtypes were also collapsed into a single Non-Spatial supercategory. Hence, based on statistics defined in Table 4, the mapping resulted as follows: *Cluster 2* → *Spatial–Toponymic*, *Cluster 0* → *Spatial–Non-Toponymic* and *Cluster 3* → *Non-Spatial*. Here, it is important to mention that we assigned Cluster 3 as Non-Spatial despite it having the similar number of Spatial–Non-Toponymic cases, since the Spatial–Non-Toponymic label was assigned to Cluster 0, requiring each cluster to have a distinct label.

We constructed a confusion matrix, illustrated in Figure 6 by cross-tabulating clusters (predicted labels) with gold annotation (true labels). Finally, precision, recall and F-1 scores are computed and summarized in Table 5, whereas macro/micro averaged scores are also depicted in Figure 6. The visualizations facilitate direct comparison of cluster-label correspondence and highlight the strengths and limitations of unsupervised BERT-based sense separation, particularly in distinguishing spatial versus non-spatial uses.

4.3. Context factors identification and quantitative distance computation

Based on the aspects of information or relation extraction described in Section 3.2.2 and applying the extraction technique on the set of embeddings or sentences with

Table 6. Identified context factors of proximity in CLDW.

Context dactor	Examples
Type of object	Druid's Circle, Keswick
Scale of object	Cumberland, England
Size of object	door, house, island
Feature/sub-part of object	head of the lake, foot of the lake
Geography/demography of object	village, town, parish
Means of travel	walk, boat, carriage
Activity	ascend, view, excursion

Table 7. Comparing quantitative distances in kilometers.

LO	RO	Context factor of proximity	AD (feature-center)	RD (walking)
Castlerigg (POI)	Keswick (Settlement)	Object type	1.65	2.30
Greta Hall (POI)	Keswick (Settlement)	Object type	0.77	1.00
Newlands (Region)	Keswick (Settlement)	Scale, Demography	5.18	6.90
Bowness (Settlement)	Windermere (Lake)	Object size, feature	0.31	0.50
Long Meg and Her Daughters (POI)	Penrith (Settlement)	Object type	10.41	12.2
King Arthur's Round Table (POI)	Eamont Bridge (Settlement)	Object type	0.35	0.50
Ullswater (Lake)	Pooley Bridge (Settlement)	Mode of travel	0.13	0.22
Lowdore (Waterfall)	Derwent Water (Lake)	Object type, feature	0.61	0.70
Ullswater (Lake)	Patterdale (Settlement)	Object size	0.62	0.80
Dunmallard (Hill)	Pooley Bridge (Settlement)	Object type	0.38	0.90

spatial proximity information, we obtained spatial relation triples along with the associated contextual information. Table 6 highlights the identified context factors with examples to illustrate their presence in the text. The custom gazetteer (Rayson et al. 2017) contains 19 distinct categories that provide information on the type of an object, and another attribute in the gazetteer labels each spatial entity as internal or external to the Lake District area that further helps in identifying the scale of an object. Usually, the labeled categories are indicators of the size of an object as well, such as an island, a lake, a spatial entity with height, a specific house or farm or street and so on. Other categories provide information on geographic/demographic context of an object, such as region, settlement and vale. Hence, collectively, categories give information on the type, size, scale and geographic/demographic context of objects in cases where the object is a particular named-entity. Non-toponymic values for any of these context factors, feature/sub-part of an object and factors, such as words giving information on means of travel and activities do not require gazetteer-based matching but only need to be extracted in connection to the primary object in a sentence as described in Section 3.2.2.

Having extracted the proximity triples, we present the results for context factor identification and quantitative distance computation for selected LO-RO pairs in Table 7, the source snippets are present in Appendix B, Table B1. Let us focus on the first three columns in Table 7. We defined some rules to apply reasoning and derive the underlying context factor of proximity. For example, if LO is a point of interest (POI) whose location is defined with respect to a RO, then the RO has to be generally an equally or more well-known spatial entity qualified to be referenced to, hence, the primary factor of proximity is the type of objects. In such a case, the size of the RO can also be the driver of proximity as mentioned earlier that larger objects are generally better reference objects. Although the labels mentioned against LO and RO in Table 7 are based on the custom gazetteer, the nature of a spatial entity can be overlapping, such as a settlement can be a popular tourist hub. The interpretation of contextual proximity is driven in relation to identifying the category of LO and other extracted aspects such as feature information as well. Next, a relation between LO of large area or extent as of RO suggests the proximity factor to be the scale of objects. Similarly, when geographical objects of similar nature or comparable proportions are related, the underlying proximity context is the size of objects.

For quantitative distances, we utilized QGIS[4] Proximity analysis tool for Cartesian computation and map visualization. The AD between the extracted entities has been calculated in two ways. The first approach is the *distance to the nearest hub* defined as the point to line or polygon distance based on measure of the distance from the centroid of the point source to the centroid of the line or polygon feature which is the destination. The second approach is the *shortest line between features* that determines the point to line or point to polygon distance from the source to the nearest feature of the destination. We first retrieved the point, line and polygon geometries of the spatial entities using Overpass Turbo[5]. The resultant GeoJSON file containing geographical coordinate data of the extracted objects is passed to QGIS for distance calculation. It should be noted that any pair containing water bodies such as lakes, the distance is computed for the nearest accessible shoreline of the named lake, rather than to the lake's geometric centroid.

For the RD computation and visualization, we utilized the Google Maps JavaScript API[6] and developed script to calculate the walking distance along mapped pedestrian routes. The Directions API service allows the calculation of route distance with practical considerations such as turns and the local street and footpath network, providing an estimate of the walking distance between the two features. The values for AD and RD are listed in Table 7 for selected LO-RO pairs. Figure 7 visualizes the AD between a point object (Castlerigg) and a polygon object (Keswick), where the distance based on feature center is shown by the dotted blue line and the distance to the closest boundary point of the destination feature by a dotted red line.

5. Discussion

Referring to the research questions defined in Section 1, the discussion below will be on understanding the results generated at various stages and how they justify.

5.1. Contextual embedding and clustering

The first research question is concerned with the focus of BERT contextual embeddings whether it is on syntactics or semantics of 'near'. We begin with the reason for choosing the embeddings generated with *distilbert-base-uncased, sum of all layers*

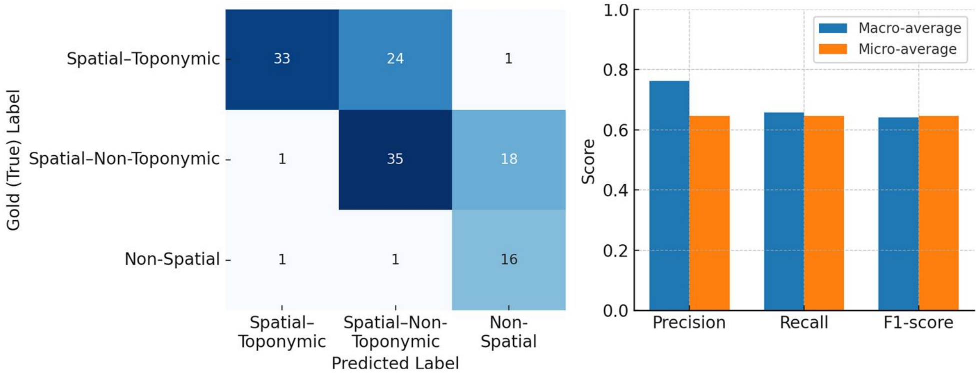


Figure 6. Confusion matrix (left) and macro/micro averaged scores(right).

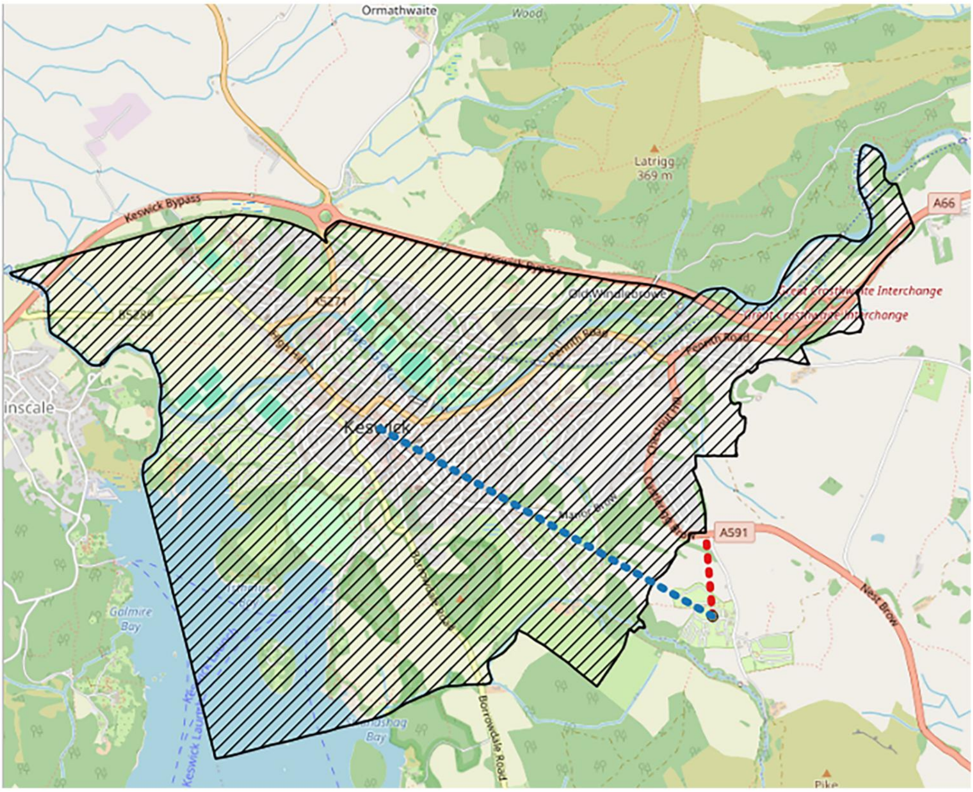
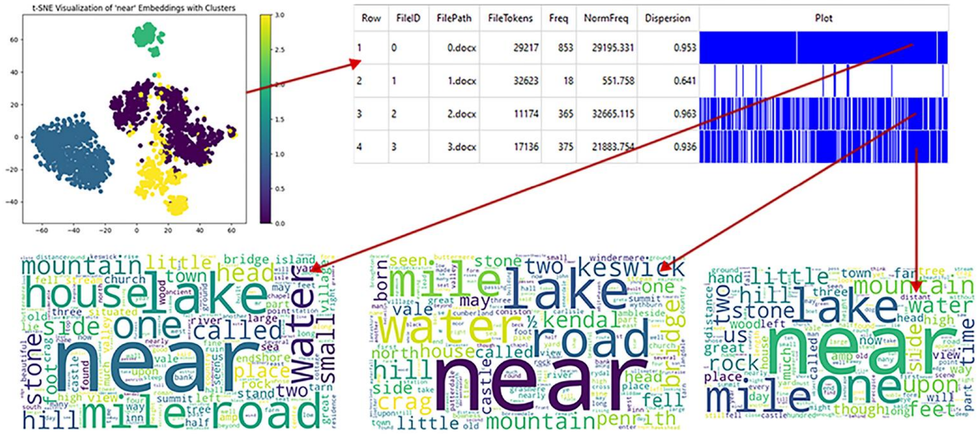


Figure 7. Absolute distances (AD) from Castlerigg (point) to Keswick (polygon): centre-to-centre = 1.65 km (blue), centre-to-nearest boundary = 0.0056 km (red).

configuration and number of clusters c = 4. Although Table 3 reflects the highest scoring setting as *roberta-base* with *sum of all layers* and number of cluster c = 3, the score drops drastically from c = 4 to 5, whereas *bert-base-uncased* and *distil-bert-base-uncased* remain stable across cluster counts. A careful analysis of the t-SNE visualization in Figure 5 shows uneven clustering where *roberta-base* forms one dominant cluster and tiny singletons, indicating weaker separation in its embedding space for



Row	FileID	FilePath	FileTokens	Freq	NormFreq	Dispersion	Plot
1	0	0.docx	29217	853	29195.331	0.953	
2	1	1.docx	32623	18	551.758	0.641	
3	2	2.docx	11174	365	32665.115	0.963	
4	3	3.docx	17136	375	21883.754	0.936	

Figure 8. Word clouds for the plot of 'near', generated after stopword removal, illustrating the raw distribution of frequently used terms.

this dataset; k-means then collapses most points into a single cluster and pushes outliers into the rest. This effect was consistent across different layer combination strategies and cluster counts. Hence, both k-means with t-SNE visualization and average Silhouette scores have assisted in selecting a better embedding output, further verified with qualitative analysis as mentioned in Section ??. The finding highlights the importance of model and parameter selection in unsupervised sense clustering tasks.

Moving to cluster interpretation, Figure 8 shows the frequency, normalized frequency and dispersion of 'near', developed with AntConc (Anthony 2023), for the selected embeddings. It can be seen that cluster 1 has surprisingly very few occurrences of 'near' compared to other three clusters. This happened because the BERT model considered the contextual difference between 'near' and 'nearly' and assigned maximum 'nearly' instances to a separate cluster i.e., Cluster 1. This is the reason that Cluster 1 was excluded from the annotation-based evaluation process since the focus of the study is on 'near'.

Figure 8 also visualizes textual details of the frequency chart with stop words-filtered word clouds. The thematic distinctions between clusters have become more apparent and interpretable:

- Cluster 0 is dominated by references to the physical geography of the region, with frequent terms such as 'lake', 'mountains', 'village', 'castle', 'hall', and 'river'. This cluster primarily represents sentences describing natural features and notable sites within the landscape.
- Cluster 2 is centered on towns, settlements, and routes, as reflected by prominent words like 'keswick', 'penrith', 'bowness', 'ambleside', 'church', 'road', 'bridge', and 'castle'. The vocabulary indicates a focus on inhabited places and the infrastructure connecting them, including roads, bridges, and historical landmarks.
- Cluster 3 features terms associated with sensory experience, scene-setting, and temporal reference, such as 'cloud', 'mountain', 'view', 'water', 'appearance', 'sun', 'light', 'distance', and 'scene'. This suggests a cluster of sentences conveying

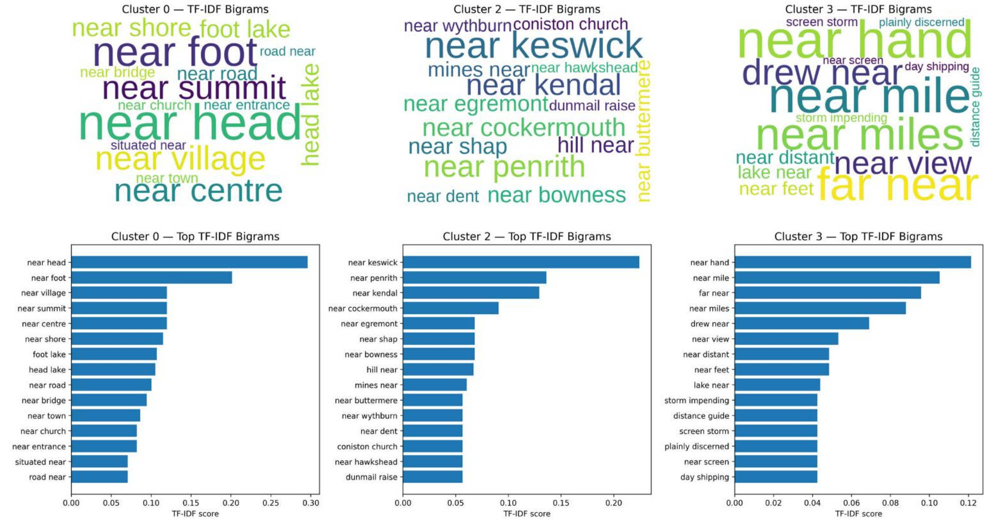


Figure 9. Enhanced visualization using TF-IDF weighted bigram word clouds (top row) and corresponding bar charts (bottom row), highlighting distinctive cluster-specific vocabulary.

subjective impressions, landscape descriptions, and changing atmospheric or temporal conditions.

The distribution of the word 'near' across Clusters 2 and 3 indicates that both spatial and non-spatial uses are present, highlighting the inherent ambiguity of certain spatial language in context. To further refine these distinctions, Figure 9 presents TF-IDF weighted bigram word clouds with supporting bar charts, which surface more cluster-specific collocations. Because stopwords are removed prior to bigram generation, some phrases appear in shortened form (e.g. 'near at hand' becomes 'near hand'), but they still capture the intended associations. In summary, the enhanced analysis clarifies the thematic coherence of each cluster and reinforces the meaningfulness of the k-means groupings, even where some overlap remains.

5.2. Observations on evaluation outcomes

While k-Means clustering effectively groups sentences by topic and theme, the distinction between spatial and non-spatial senses, as well as further subclassification based on specific geographic entities, remains challenging and requires additional analysis. Hence, we adopted k-NN in an unsupervised setting to determine similarity of the corpus sentences with the generated contextual embeddings. The sentence by sentence inspection revealed the inaccuracies of embeddings organized as contextually similar. Appendix A, Table A1 contains k-NN results for three cases. The first sentence's triple pattern is $\ll near, LO, RO \gg$ and k-NN returns sentences similar to this syntactic pattern. For the second sentence, the results show relevance to the theme of first person traveling and admiration of nature. Though we have not trained a classifier here, the argument of Wiedemann et al. (2019) is valid and relevant to the outcomes obtained that the similarity of contextual embeddings depends heavily on the semantic and structural similarity of sentences for polysemous target words (such as the case of 'near'), the greater the number of example sentences that are available for a particular sense, the greater the likelihood that a nearest neighbor will convey that same meaning. This has been observed for more complex senses, such as for the third sentence, k-NN could not accurately determine similar embeddings as there were not enough samples for a non-spatial interaction sense, for instance nearness to danger in this case, produced a group of heterogeneous sentences with different non-spatial senses.

The other approach to assess effectiveness of the method is the annotation-based evaluation. The results presented in Table 5 and Figure 6 show that Spatial–Toponymic predictions are very precise (94.3%) but have moderate recall (56.9%), meaning many true instances are missed. Spatial–Non-Toponymic shows balanced precision (58.3%) and recall (64.8%) but is prone to confusion with both other classes. Non-Spatial achieves high recall (88.9%) but low precision (45.7%) due to substantial miss-classification from Spatial–Non-Toponymic. Macro-averages indicate moderate overall performance, while slightly lower micro-averages reflect the impact of errors in larger classes. With an annotation set of only 130 samples, results should be viewed as indicative rather than definitive.

We experimented with different hidden layer representations, the best performing is the sum of all layers heuristic which indicates that both syntactic and semantic information have been incorporated in the transformer layers for generating contextual embeddings. However, due to the complexity of sentences in the corpus, it has been observed that structural similarity, syntactic information such as POS tags, high occurring geographic nouns and toponyms have influenced the embedding computation.

5.3. Context factors and quantitative distances

Having obtained the set of sentences with spatial sense of 'near', the second research question is concerned with the fine-grained context factors of proximity in a relation triple. The Lake District, as its name says, is known for the lakes in the region and writers associated the sense of proximity with different parts of the lakes, and so, evident from the computations at both course and fine-grained level. For the CLDW, the most important reference objects for proximity have been the large and well-known settlements among gazetteer entries/toponyms and features of lakes among non-gazetteer entries, such as *'head/foot/centre/margin of the lake'*. It should be noted that some toponyms correspond to both settlements and lakes, such as *'Windermere'* is a settlement and a lake as well.

Table 7 provides a comparative list of LO and RO with their categories derived from the custom gazetteer and Section 4.3 describes the reasoning applied to determine the context factor of proximity for each LO-RO pair. The nature of the LO influences the interpretation of the nature of RO in deciding a context factor. For example, *'Keswick'* is associated with a POI *'Castlerigg'* and with a valley *'Newlands'*. This association with different types of LO defines the different nature of spatial entity *'Keswick'* in different relations. In the first triple, it is used as a reference object for a tourist place, whereas in the second triple, it is used as a reference object for a region in demographic comparisons. Appendix B, Table B1 provides the source texts for these relation triples, and certain factors, such as object feature and mode of travel can be understood by reading the text. Lastly, the factor of object size has been assigned when entities of similar magnitude are referenced.

The other columns provide metrical distances between each pair. For most of the cases, the textual proximity is validated by the quantitative distances. It can be observed that large quantitative distances have also been obtained for certain entries which is logical as the corpus contains entries from the 16th century onwards to the 19th century, indicating difference of sensing nearness compared to contemporary perceptions. Finally, the nearest shoreline distance approach captures the real-world immediacy of shoreline proximity but produces substantially smaller values than centroid-to-centroid measurements, particularly for long, narrow lakes such as *Ullswater*, where the centroid lies several kilometers from either end.

6. Conclusion

Digital humanities has recently witnessed significant work in analyzing the narratives in historical archives using GIS. However, the studies have largely focused on GTA-based or standard NLP approaches. We address this gap in connection to with CLDW and conducted an interdisciplinary study for the analysis of spatial proximity. The framework utilized the pre-trained BERT model with its variants to generate contextual word embeddings for a proximity indicator 'near'. We have not fine-tuned a supervised model here as the proposed work, to the best of our knowledge, is the first experimental initiative on the semantic interpretation of 'near' focusing on identifying coarse and fine-grained aspects using advanced NLP methods and we chose to explore the utility of existing models to generate a baseline for future work. Owing to the impoverished linguistic polysemy of 'near', we address the problem as an open-ended task to assess the focus of BERT contextual embeddings in this scenario, thereby analyzing its potential for usage and limitations. Hence, we resorted to analysis of the embeddings directly, avoiding fine-tuned probes. Our approach can serve as a foundational study for using pre-trained models for syntactic and semantic analysis of spatial proximity indicators, particularly in narrative texts where the complexity of writing is profoundly distinct from generic datasets.

The BERT-based contextual embedding approach has been able to produce clusters based on spatial category of geographic entities, that is to say, majority of the embeddings associated with lakes and those with settlements or towns have been clustered separately, and non-spatial senses are also largely collected in a distinct cluster. Moreover, the method considered semantic differences of 'near' and 'nearly' as well. Nevertheless, Section 5 brings up some limitations of experimental findings. In general, the available number of instances of 'near' and the intricacy of narrative writing in CLDW have influenced the embedding results, particularly with respect to discriminating the various non-spatial senses of 'near' as we have discussed specific examples in Section 5.2 where lack of sentences for a particular type of proximity sense has resulted in a noisy K-NN group. This can also be explained by the annotation-based evaluation where the Non-Spatial majority cluster is cross-contaminated by Spatial–

Non-Toponymic items because sentence complexity makes it difficult to determine the primary sense in toponym-free sentences. From a contextual factor identification perspective, our approach is limited to identifying factors pertaining to the environmental/physical and functional context tied to proximity while specifically focusing on the term 'near', whereas the narrative descriptions of the CLDW abundantly contain implied or evident references to sense of spatial nearness in idiomatic, poetic, metaphorical and other forms of constructs.

In the future, we aim to improve the performance of contextual word embeddings to better disambiguate semantic sense through training on sense-annotated data, incorporation of wider textual context and integration of external knowledge from structured resources, such as gazetteers and ontologies. Moreover, though the present study is focused on the preposition 'near' in a single historical corpus, which allowed us to conduct an in-depth analysis of spatial and non-spatial sense discrimination in a rich, domain-specific context, our methodology using contextual embeddings and unsupervised clustering is, in principle, applicable to other spatial prepositions (such as 'across' or 'within') and to other corpora, both historical and contemporary. Exploring these extensions remains an important avenue for future research, and would provide further evidence for the robustness and adaptability of our approach. Along this line, we also seek to experiment with the identification of other linguistic relation expressions, such as those defined in Section 3.1.2 which describe the different impressions of proximity writers use in their narratives. This will allow the assessment of memory, emotions, or other related factors' influence on the descriptions of proximity. An important aspect is to consider the timeline of the corpus (the Age of Sensibility, the Romantic and the Victorian era) to analyze the shift, if any, in the perception of nearness, for instance, tracing the advent of railways and its effect on the interpretation of distances. With regards to tools and techniques, the use of OSM and the Overpass API for spatial data extraction demonstrates both the strengths and limitations of current digital gazetteers, particularly when applied to historical sources. Addressing issues of incomplete, changed or evolving place representation is an important avenue for future research, both in digital humanities and geospatial NLP.

Notes

1. CLDW document IDs for example sentences: Anon-cqp-66-1857-b.txt, Gilpin-cqp-21-1786-a.txt, Waugh-cqp-69-1861-b.txt, Budworth-cqp-26-1792-b.txt, Black-cqp-64-1853-a.txt
2. The English Lake District: https://whc.unesco.org/en/list/422/
3. Lake District National Parkhttps://www.lakedistrict.gov.uk/
4. QGIS: https://www.qgis.org/
5. Overpass Turbo: https://overpass-turbo.eu/index.html
6. Google Maps API: https://developers.google.com/maps/

Acknowledgements

The authors thank the entire team of the project 'Understanding imprecise space and time in narratives through qualitative representations, reasoning, and visualization' (https://spacetime-narratives.github.io/). Valuable feedback was provided by the anonymous reviewers and is gratefully acknowledged.

Disclosure statement

No potential competing interest was reported by the author(s).

Funding

The research is supported by the UK's Economic and Social Research Council (ES/W003473/1) and the US National Science Foundation (NSF#2225179). This research was also partially supported by the EPSRC under grant (EP/Z003512/1). The support of Microsoft under their Accelerating Foundation Models Research initiative in providing Azure credits to Anthony G. Cohn is also acknowledged.

Notes on contributors

Erum Haris is a Research Fellow in the School of Computer Science at the University of Leeds. Her research focuses on applying geospatial AI to text corpora. She contributed to the conceptualization, data analysis, experiments, and manuscript preparation.

Anthony G. Cohn is Professor of Automated Reasoning at the University of Leeds and Foundation Models Lead at the Alan Turing Institute, specializing in knowledge representation and qualitative spatio-temporal reasoning. He supervised the research, contributed to the theoretical framework, and manuscript review.

John G. Stell is a Senior Lecturer in the School of Computer Science at the University of Leeds, specializing in qualitative spatial reasoning. He supervised the research, contributed to conceptualization, manuscript review, and secured project funding.

Data and codes availability statement

The data, codes, and instructions that support the findings of this study are available at: https://doi.org/10.6084/m9.figshare.28616981.v1

CLDW is available at GitHub: https://github.com/UCREL/LakeDistrictCorpus and archived at: swh:1:dir:221aa46bd3eff0783f0a471d7b8031a2910b6739.

References

Anthony, L., 2023. Antconc (version 4.2.4) [computer software]. Available from: https://www.laurenceanthony.net/software.

Barker, E., Isaksen, L., and Ogden, J., 2015. Telling stories with maps. In: *New worlds from old texts: Revisiting ancient space and place*. Oxford, UK: Oxford University Press, 181.

Brenda, M., 2017. A cognitive perspective on the semantics of near. *Review of Cognitive Linguistics*, 15 (1), 121–153.

Brennan, J., and Martin, E., 2012. Spatial proximity is more than just a distance measure. *International Journal of Human-Computer Studies*, 70 (1), 88–106.

Brown, T.B., *et al.*, 2020. Language models are few-shot learners. In: *Proceedings of the 34th International Conference on Neural Information Processing Systems (NeurIPS 2020),* Red Hook, NY. Curran Associates, Inc.

Burigo, M., and Coventry, K.R., 2010. Context affects scale selection for proximity terms. *Spatial Cognition & Computation*, 10 (4), 292–312.

Butler, J.O., *et al.*, 2017. Alts, abbreviations, and AKAs: historical onomastic variation and automated named entity recognition. *Journal of Map & Geography Libraries*, 13 (1), 58–81.

Candela, G., *et al.*, 2023. An ontological approach for unlocking the colonial archive. *Journal on Computing and Cultural Heritage*, 16 (4), 1–18.

Chesnokova, O., *et al.*, 2019. Hearing the silence: finding the middle ground in the spatial humanities? extracting and comparing perceived silence and tranquillity in the English Lake District. *International Journal of Geographical Information Science*, 33 (12), 2430–2454.

Coventry, K.R., and Garrod, S.C., 2004. *Saying, seeing, and acting: The psychological semantics of spatial prepositions. Essays in Cognitive Psychology*. Hove, UK: Psychology Press.

Devlin, J., *et al.*, 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4171–4186.

Dey, A.K., 2001. Understanding and using context. *Personal and Ubiquitous Computing*, 5 (1), 4–7.

Donaldson, C., Gregory, I.N., and Taylor, J.E., 2017. Locating the beautiful, picturesque, sublime and majestic: spatially analysing the application of aesthetic terminology in descriptions of the English Lake District. *Journal of Historical Geography*, 56 (1), 43–60.

Ezeani, I., *et al.*, 2023. Towards an extensible framework for understanding spatial narratives. In: *Proceedings of the 7th ACM SIGSPATIAL International Workshop on Geospatial Humanities (GeoHumanities '23)*, New York, NY. ACM Press, 1–10.

Fabregat, F.N., 2022. The polysemy of prepositions at, beside, by, near and next to: the horizontal axis of spatial relations. Unpublished manuscript.

Foka, A., *et al.*, 2020. Semantically geo-annotating an ancient Greek "travel guide": itineraries, chronotopes, networks, and linked data. In: *Proceedings of the 4th ACM SIGSPATIAL Workshop on Geospatial Humanities (GeoHumanities '20)*, New York, NY, USA, 1–9.

Gahegan, M., 1995. Proximity operators for qualitative spatial reasoning. *In*: A.U. Frank and W. Kuhn, eds. *Spatial information theory: A theoretical basis for GIS. COSIT 1995. Lecture Notes in Computer Science*, vol. 4128. Berlin, Heidelberg: Springer, 31–44.

Gregory, IN., *et al.*, 2024. Exploring qualitative geographies in large volumes of digital text: Placing tourists, travelers, and inhabitants in the English Lake District. *Annals of the American Association of Geographers*, 114 (9), 1985–2009.

Hadiwinoto, C., Ng, H.T., and Gan, W.C., 2019. Improved word sense disambiguation using pre-trained contextualized word representations. ArXiv:1910.00194.

Hahn, J., *et al.*, 2016. A computational model for context and spatial concepts. *In*: T. Sarjakoski, Y.S. Maribel and L.T. Sarjakoski, eds. *Geospatial data in a changing world: Selected papers of the 19th AGILE conference on geographic information science. Lecture Notes in Geoinformation and Cartography*. Switzerland: Springer International Publishing AG, 3–19.

Hamzei, E., Winter, S., and Tomko, M., 2020. Place facets: a systematic literature review. *Spatial Cognition & Computation*, 20 (1), 33–81.

Haris, E., Cohn, A.G., and Stell, J.G., 2023. Understanding the spatial complexity in landscape narratives through qualitative representation of space. *In*: R. Beecham, J.A. Long, D. Smith, Q. Zhao and S. Wise, eds. *GIScience 2023. Schloss Dagstuhl–Leibniz-Zentrum für Informatik, Leibniz International Proceedings in Informatics (LIPIcs)*. Saarbrücken/Wadern, Germany: Schloss Dagstuhl – Leibniz-Zentrum für Informatik, vol. 277, 37:1–37:6.

Haris, E., Cohn, A.G., and Stell, J.G., 2024. Semantic perspectives on the lake district writing: Spatial ontology modeling and relation extraction for deeper insights. In: *16th International Conference on Spatial Information Theory (COSIT 2024)*. Schloss Dagstuhl–Leibniz-Zentrum für Informatik.

Langacker, R.W., 2008. *Cognitive grammar: A basic introduction*. Oxford: Oxford University Press.

Liu, Y., *et al.*, 2019. RoBERTa: a robustly optimized BERT pretraining approach. ArXiv:1907.11692 [cs].

Mayfield, E., and Black, A.W., 2020. Should you fine-tune BERT for automated essay scoring? In: *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, 151–162.

McDonough, K., Moncla, L., and van de Camp, M., 2019. Named entity recognition goes to old regime France: geographic text analysis for early modern French corpora. *International Journal of Geographical Information Science*, 33 (12), 2498–2522.

McKenzie, G., and Hu, Y., 2017. The "nearby" exaggeration in real estate. In: *Proceedings of the Workshop on Cognitive Scales of Spatial Information.*

Mikolov, T., *et al.*, 2013. Efficient estimation of word representations in vector space. ArXiv: 1301.3781.

Miller, G.A., and Johnson-Laird, P.N., 1976. *Language and perception*. Cambridge: Cambridge University Press.

Minock, M., and Mollevik, J., 2013. Context-dependent near and far in spatial databases via supervaluation. *Data & Knowledge Engineering*, 86, 295–305.

Murrieta-Flores, P., and Martins, B., 2019. The geospatial humanities: past, present and future. *International Journal of Geographical Information Science*, 33 (12), 2424–2429.

Murrieta-Flores, P., Favila-Vázquez, M., and Flores-Morán, A., 2019. Spatial humanities 3.0: Qualitative spatial representation and semantic triples as new means of exploration of complex indigenous spatial representations in sixteenth century early colonial Mexican maps. *International Journal of Humanities and Arts Computing*, 13 (1–2), 53–68.

Nadeem, F., *et al.*, 2019. Automated essay scoring with discourse-aware neural models. In: *Proceedings of the ACL Workshop on Innovative Use of NLP for Building Educational Applications*, 484–493.

Novel, M., *et al.*, 2020. Nearness as context-dependent expression: an integrative review of modeling, measurement and contextual properties. *Spatial Cognition & Computation*, 20 (3), 161–233.

Olbricht, R.M., 2015. Data retrieval for small spatial regions in OpenStreetMap. *In*: J. Jokar Arsanjani, A. Zipf, P. Mooney and M. Helbich, eds. *OpenStreetMap in GIScience. Lecture Notes in Geoinformation and Cartography*. Cham: Springer.

Pennington, J., Socher, R., and Manning, C.D., 2014. GloVe: Global vectors for word representation. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 1532–1543.

Peters, M.E., Ruder, S., and Smith, N.A., 2019. To tune or not to tune? adapting pretrained representations to diverse tasks. In: *Proceedings of the Workshop on Representation Learning for NLP (RepL4NLP*-2019), 7–14.

Rayson, P., *et al.*, 2017. A deeply annotated testbed for geographical text analysis: The corpus of Lake District writing. In: *Proceedings of the 1st ACM SIGSPATIAL Workshop on Geospatial Humanities (GeoHumanities '17)*, New York, NY. ACM Press, 9–15.

Sanh, V., *et al.*, 2020. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. ArXiv:1910.01108 [cs].

Steiner, E., *et al.*, 2023. Spatiotextual regions: Extracting sense of place from spatial narratives. *In*: R. Westerholt and F. Mocnik, eds. *Proceedings of the Fourth International Symposium on Platial Information Science (PLATIAL'23),* Dortmund, Germany. PLATIAL'X, 1–8.

Stock, K., and Hall, M., 2018. The role of context in the interpretation of natural language location descriptions. In: *Proceedings of Workshops and Posters at the 13th International Conference on Spatial Information Theory (COSIT 2017) 13*. Springer International Publishing, 245–254.

Tyler, A., and Evans, V., 2003. *The semantics of English prepositions: Spatial senses, embodied meaning and cognition*. Cambridge: Cambridge University Press.

Wiedemann, G., *et al.*, 2019. Does BERT make any sense? interpretable word sense disambiguation with contextualized embeddings. ArXiv:1909.10430.

Wolter, D., and Yousaf, M., 2018. Context and vagueness in automated interpretation of place description: A computational model. In: *Proceedings of Workshops and Posters at the 13th International Conference on Spatial Information Theory (COSIT 2017) 13*. Springer International Publishing, 137–142.

Xu, S., and Klippel, A., 2013. Linking context and proximity through web corpus. In: *Proceedings of the 7th Workshop on Geographic Information Retrieval (GIR '13),* New York, NY. ACM Press, 45–46.

Yao, X., and Thill, J., 2005. How far is too far? a statistical approach to context-contingent proximity modeling. *Transactions in GIS*, 9 (2), 157–178.

Zwarts, J., 2017. Spatial semantics: Modeling the meaning of prepositions. *Language and Linguistics Compass*, 11 (5), 1–20.

Appendix A. k-NN results for selected sentences

Table A1. An unsupervised application of k-NN on DistilBERT embeddings.

Sentence	Nearest neighbors (k = 10)
Near the library was the room in which he died after years of mental darkness.	- Near the church are the Parsonage House and the School. - Near the altar is a marble monument by Flaxman, erected in memory of Dr. - Near the cathedral there is a very modern church, which looks on the outside more like a ball-room than a place of worship. - Near the church is a public house. - Near the church, as usual, there is a comfortable old inn. - Near the bridge are the works of Elterwater Gunpowder Company. - Near this chapel is the Thrang Slatequarry, where the stranger should look in and see what a mighty excavation has been caused by the demand for this fine slate. - Near the church, as usual, there is an old inn. - Near the head is Woodhouse, Mr.
The sun was setting, and as we were drawing near our destination I almost forgot my fatigue.	- Leaving Borrodale, you again find yourself near the head of Derwentwater. - So when we arrived near the summit I alighted. - I began to think myself near my destination, for smoke was curling up here and there among the trees below. - For three or four miles the road lies more or less near the stream, and we trailed by through its enchanting sights and sounds. - He came in, and being directed to sit near the fire, was asked a number of questions about his farm. - A good turnpike-road, on which we entered near the village of Lorton, and a knowledge of the country, set at naught all such ideas with us. - We are drawing near the foot of Wast Water. - The stranger must not omit to observe near the head of the pass, the fallen rock, ridged like a roof, whose form(like that of a miniature church) has given its name to its precincts. - We proceeded into the entrance of Barrowdale, and came near the unadorned, but picturesque, village of Grange. the foam would have nearly covered them.
Did hapless mortals view The dangers near, what misery would ensue.	- Or could he think, his homely garments near, His lord would rising from the heath appear; And dressed again, over mountains take his way, Marking with kind encouragement his play. - We meet and pass; but still I have thee near, Tuning thy murmurs in my partial ear. - Day's adventures talked over and bedtime near. - His faithful steed made signs of inward fear, Filled with a sense of hidden danger near; Yet still he rode, and rode, and rode along, With silent speed the snowy hills among. - What, do not you think we the summit are near. - Here the outline of the hills both near and distant is wonderfully varied. - We began to think that Cockley Beck ought to be near, and yet there was no visible habitation in the silent glen. - The dog being yet too far off, and no help near, and no time to be lost, he had presence of mind to place the end of the fork-shaft against the base of the stiffest bull-toppin within reach, and pointing the grains of the fork in the direction of the coming attack, stood with it in his hand poised between his knees so that he could raise the points or lower them to where they were most likely to take effect. - Then perceiving a hill that was rising quite near, We climbed it, and found it overhung Windermere.

Appendix B. Sentence snippets for LO-RO pairs in Table 7

Table B1. LO-RO pairs with sentence snippets from CLDW.

LO	RO	Sentence snippet
Castlerigg (POI)	Keswick (Settlement)	Having climbed for nearly a mile, please to halt and look back, and you have a view well worth all your toil, embracing Little Langdale, Colwith, Skelwith, Loughrigg, the bright waters of Windermere, and the groves and mountains beyond, altogether making up a picture approaching in beauty, though inferior in richness and variety,(as all other prospects are) to that seen from the Castlerigg, near Keswick.
Greta Hall (POI)	Keswick (Settlement)	Amongst the villas near Keswick is Greta Hall, the seat of Robert Southey, Esq.
Newlands (Region)	Keswick (Settlement)	The person, for instance, who held the curacy in the Vale of Newlands, near Keswick, at that period, exercised the various trades of tailor, clogger, and butter-print maker.
Bowness (Settlement)	Windermere (Lake)	Bowness is charmingly situated near the centre of Windermere, on its eastern bank, and at the bottom of a small bay.
Long Meg and Her Daughters (POI)	Penrith (Settlement)	We ourselves counted Long Meg and her daughters, near Penrith, many times before making out the prescribed sixty - seven, with any certainty.
King Arthur's Round Table (POI)	Eamont Bridge (Settlement)	In the same neighbourhood, near Eamont bridge, is another antiquity, called Arthur's round table, of an exact circular figure, rising above the plain on which it stands, and surrounded by a trench, about ten paces wide, from which the earth by which it is formed has been taken..
Ullswater (Lake)	Pooley Bridge (Settlement)	Approaching from Penrith, travellers may go by Eamont Bridge, Yanwath, and Tirrel, and past the old tower of Yanwath Hall, in five miles to Pooley Bridge; or they may turn off the Keswick road beyond the second mile-stone, and pass through the beautiful grounds of Dalemain, coming to Ullswater near Pooley Bridge, in 6 miles, and to Patterdale in 15 m.
Lowdore (Waterfall)	Derwent Water (Lake)	Three miles from Keswick we arrive at the water-fall of Lowdore, called the niagara of Derwent Water, situated near the head of the lake.
Ullswater (Lake)	Patterdale (Settlement)	Ullswater is in general very deep, and particularly near to Patterdale.
Dunmallard (Hill)	Pooley Bridge (Settlement)	Near Pooley Bridge, a remarkable hill called Dunmallard stands at the foot of Ullswater; it appears to be formed of a conglomerated mass, which may be seen on the sides of the road, and by which the lake seems to be embanked.