International Journal of Geo-Information

MDPI

Article

A Hybrid Approach for Geo-Referencing Tweets: Transformer Language Model Regression and Gazetteer Disambiguation

Thomas Edwards, Padraig Corcoran and Christopher B. Jones *

School of Computer Science and Informatics, Cardiff University, Senghennydd Rd, Cardiff CF24 4AX, UK; edwardstj1@cardiff.ac.uk (T.E.); corcoranp@cardiff.ac.uk (P.C.)
* Correspondence: jonescb2@cardiff.ac.uk

Abstract

Recent approaches to geo-referencing X posts have focused on the use of language modelling techniques that learn geographic region-specific language and use this to infer geographic coordinates from text. These approaches rely on large amounts of labelled data to build accurate predictive models. However, obtaining significant volumes of geo-referenced data from Twitter, recently renamed X, can be difficult. Further, existing language modelling approaches can require the division of a given area into a grid or set of clusters, which can be dataset-specific and challenging for location prediction at a fine-grained level. Regression-based approaches in combination with deep learning address some of these challenges as they can assign coordinates directly without the need for clustering or grid-based methods. However, such approaches have received only limited attention for the geo-referencing task. In this paper, we adapt state-of-the-art neural network models for the regression task, focusing on geo-referencing wildlife Tweets where there is a limited amount of data. We experiment with different transfer learning techniques for improving the performance of the regression models, and we also compare our approach to recently developed Large Language Models and prompting techniques. We show that using a location names extraction method in combination with regression-based disambiguation, and purely regression when names are absent, leads to significant improvements in locational accuracy over using only regression.

Keywords: regression; neural networks; social networks; geo-referencing; geo-coding; wildlife; ecology

Academic Editors: Wei Huang and Wolfgang Kainz

Received: 14 May 2025
Revised: 18 July 2025
Accepted: 15 August 2025
Published: 22 August 2025

Citation: Edwards, T.; Corcoran, P.; Jones, C.B. A Hybrid Approach for Geo-Referencing Tweets: Transformer Language Model Regression and Gazetteer Disambiguation. *ISPRS Int. J. Geo-Inf.* **2025**, *14*, 321. https://doi.org/10.3390/ijgi14090321

1. Introduction

Social media platforms offer a rich source of information for diverse applications, including health, disaster response, marketing, and environmental monitoring [1–3]. Geo-referenced microblogs are particularly valuable in ecology, where they support research on wildlife distribution, climate change, disease spread, and invasive species [4,5]. Despite recent restrictions following Twitter's transition to X, historical Twitter data remains a key resource—especially given its global reach, real-time nature, and relevance to event-based and spatially grounded phenomena. The increasing difficulty of collecting new data under current platform constraints further heightens the value of existing collections. Moreover, methods developed on Twitter data can generalise to similar platforms such as Mastodon and Bluesky.

However, most Twitter/X posts lack explicit geo-coordinates [1,6,7]. This presents challenges, particularly for domain-specific queries such as locating wildlife observations

within specific regions. As a result, geo-referencing—inferring coordinates from post content—has become an important research task [3,8,9].

Early methods rely on gazetteers to detect and disambiguate place names [10–12] but suffer when toponyms are ambiguous or absent [3].

More recent work uses language models to exploit broader linguistic cues, dividing training data into spatial regions and learning region-specific language patterns [9,13]. These methods often require large labelled datasets and careful region granularity tuning [14].

While many approaches focus on predicting user-level locations using Tweet text, metadata, and network structure [15–17], fewer address the finer-grained task of predicting coordinates for individual posts. Regression-based models, particularly those using transformer architectures, offer an alternative by directly predicting latitude and longitude without partitioning space into discrete regions [18,19]. These models avoid grid-size tuning and can operate more flexibly, but their application in geospatial contexts remains underexplored.

Transfer learning offers a promising solution to the challenge of limited labelled data. Prior work has shown that models trained on one platform (e.g., Flickr) can improve geo-inference on another (e.g., Twitter/X) [20]. However, most studies use classical machine learning approaches rather than transformer-based models. Likewise, hybrid methods combining language models with gazetteers have proven effective in toponym disambiguation [21], but their integration with regression-based models has not been investigated.

Our work addresses these gaps by adapting transformer models for coordinate regression and enhancing them using domain-specific training, transfer learning, and hybrid disambiguation strategies. We focus on geo-referencing wildlife-related Tweets, a domain that presents unique linguistic challenges and offers high ecological value. We also conduct comparative evaluations with large language models (LLMs) and graph-based baselines to provide a comprehensive view of the trade-offs among existing methods.

1.1. Machine Learning Approaches

1.1.1. Statistical Machine Learning

Traditional machine learning models such as Support Vector Machines (SVM) and Support Vector Regression (SVR) have been widely used for text classification and spatial prediction tasks, including coordinate estimation [18,22,23]. These models typically rely on word frequency-based feature vectors and perform well in structured settings. However, they are limited in handling out-of-vocabulary (OOV) words and can struggle with fine-grained multi-class problems due to their reliance on sparse, context-independent representations.

1.1.2. Neural Network and Transformer-Based Models

Neural networks, particularly those using word embeddings, address some of these limitations by capturing semantic relationships and contextual meaning. Transformer-based models, such as BERT [24] and its improved variant RoBERTa [25], use attention mechanisms to model word context more effectively. These models are pre-trained on large unlabelled corpora and then fine-tuned for specific tasks, enabling strong performance even with limited labelled data. In this work, we adopt RoBERTa in regression mode due to its consistent performance gains over BERT in downstream NLP tasks.

1.1.3. Generative Language Models

Large generative models such as GPT [26] and LLaMA [27] have introduced in-context learning, enabling them to perform tasks with minimal or no additional training by interpreting plain-language instructions or few-shot examples [28,29]. While promising, these models are rarely evaluated in ecology-focused applications or geospatial tasks. Despite the

advances in transformer-based models, their application to geo-referencing social media—especially wildlife-related posts—remains limited. Our work explores this under-studied area by applying domain-adapted transformer models and transfer learning strategies to improve coordinate prediction.

1.2. Related Work: Geo-Referencing Social Media Data

Much of the existing research on geo-referencing social media focuses on predicting user locations rather than the locations of individual posts. These approaches often integrate Tweet content, metadata, and user networks [15,17,30]. For instance, Rahimi et al. [15] proposed a Graph Convolutional Network (GCN) model combining textual and network features for user geolocation. While effective, these methods typically aggregate multiple Tweets per user and are thus less suited to inferring the location of individual posts—especially those tied to specific events or observations.

Language modelling is a dominant approach for post-level geo-referencing, where spatially labelled data is divided into regions such as grid cells or clusters [1,13,31,32]. Models then learn regional language profiles to predict the most likely region for a given post [33,34]. Variants include fixed or adaptive grid systems [35,36] and region-based classification [37,38]. However, these methods require large labelled datasets, particularly for fine-grained predictions, and performance is sensitive to the region partitioning [14].

Some models incorporate gazetteers to match and disambiguate place names [3], but these rely on the presence of explicit toponyms, limiting their applicability. For example, Sherloc [1] uses an embedding space built from gazetteers to infer locations but fails when toponyms are missing or a reference area is not predefined. Similarly, Masis and O'Connor [39] match user-entered location strings to GeoNames embeddings, though their method targets user location and omits Tweet content.

Transfer learning approaches address data scarcity by leveraging cross-platform datasets. Van Laere et al. [20] showed that Flickr and Twitter data could be used to improve geo-referencing of Wikipedia articles. However, their models used traditional machine learning and required extensive feature engineering. Our work extends this idea by applying transfer learning to transformer-based models using Flickr data to improve Twitter/X post-level geo-inference.

Toponym disambiguation has also been explored using neural models. For example, LGGeoCoder [40] and CamCoder [41] use CNNs and embeddings to resolve place names in long texts. Other work combines ELMo or LSTM embeddings [42,43] or character-level n-grams [44] to distinguish regional naming patterns. However, these models are designed for longer documents and are less applicable to short, noisy Tweets.

Transformer models such as BERT have recently been applied to Twitter/X classification tasks [45,46], but few studies have adapted them for coordinate regression. Scherrer et al. [18,19] were among the first to use BERT for predicting coordinates directly, showing that regression models can outperform classification-based methods, particularly on small datasets. Our work builds on this by comparing regression-based models (BERT and RoBERTa) and demonstrating the superior performance of RoBERTa in our wildlife domain.

Further, Born and Manica [47] found BERT-based regressors outperform XLNet and traditional regressors (e.g., Random Forest, XGBoost). However, these models were not tailored for domain-specific content or augmented through transfer learning.

Other studies [48,49] use transformers for geolocation but treat it as a classification task tied to pre-defined locations, such as points of interest or user home locations—assumptions that do not hold in our wildlife domain.

Recently, prompting-based approaches using large language models (LLMs) such as GPT have shown promise in spatial reasoning [50,51]. However, they lack systematic comparison with transformer regression baselines, and most assume that sufficient location cues are present in the text. Reviews such as Tucker [52] also overlook detailed benchmarking across traditional and modern approaches.

In summary, there is a lack of prior work that (i) applies transformer models directly for coordinate regression in geo-referencing tasks; (ii) explores transfer learning using domain-relevant data sources such as Flickr; and (iii) integrates regression models with gazetteer-based disambiguation to improve precision.

We address these gaps through the following contributions:

- We fine-tune a transformer-based model (RoBERTa) for multivariate coordinate regression on wildlife-related Twitter/X posts, outperforming traditional statistical regressors.
- We demonstrate that domain-specific training on wildlife Tweets yields better geo-referencing performance than general-purpose models, including generative LLMs.
- We introduce a transfer learning approach that augments training data with geo-tagged Flickr posts, enhancing model accuracy under limited Twitter/X data availability.
- We propose a hybrid strategy that combines the regression model with toponym disambiguation, improving precision when place names are present.
- We release what we believe to be the largest geo-referenced dataset of wildlife-related Tweets—a valuable resource for ecological and geospatial research.
- We evaluate our approach against strong baselines, including a second hybrid approach that combines the regression model with semantic similarity; generative LLMs with prompting; BERT-based regression models; and GCN-based user location models.

2. Materials and Methods

This section outlines our methodology for geo-referencing wildlife-related social media posts in the UK by predicting geographic coordinates from Tweet text. Building on the BERT-based regression approach introduced by Scherrer et al. [19], we enhance performance through three key innovations. First, we adopt the more advanced RoBERTa model [25], which offers improved language understanding and representation. Second, we incorporate domain adaptation and transfer learning techniques, inspired by Van Laere et al. [20], by leveraging data from other social media platforms to better align our models with the target domain of Twitter posts. Third, we integrate traditional place name disambiguation strategies—based on gazetteers and semantic similarity measures—with our transformer-based regression models, extending prior work on location resolution in textual data [3,21,39]. We evaluate our approach against a range of baselines, including statistical machine learning models, neural networks, and generative language models. Our methodology comprises three main components: (1) pre-training and fine-tuning of language models (Section 2.2); (2) coordinate prediction using regression (Section 2.3); and (3) hybrid improvements via place name resolution and semantic similarity (Sections 2.4 and 2.5).

2.1. Problem Formulation

Given a Tweet T, our objective is to predict a pair of coordinates $(lat, lon) \in \mathbb{R}^2$ corresponding to the location of the wildlife-related observation described in the Tweet. Unlike classification-based approaches that map inputs to predefined regions or grids [53,54], we adopt a regression-based method to directly predict continuous-valued geographic coordinates. Further, we investigate the effectiveness of improving the accuracy of the result by using the regression coordinates to disambiguate and geocode place names, where present in T, the coordinates of which are then substituted as the prediction.

2.2. Language Model Pre-Training

We adopt the RoBERTa architecture [25], leveraging its strong contextual representation capabilities. RoBERTa model has a similar architecture to BERT (see Section 1.1.2) which makes it well-suited for tasks involving the prediction of token-level information. However, RoBERTa improves upon BERT by being trained on a significantly larger corpus and using longer input sequences, enabling it to capture broader contextual dependencies and achieve stronger performance across various prediction tasks. We use two model variants:

- Generic RoBERTa: Pre-trained on large-scale general English corpora.
- Domain-specific RoBERTa: We have fine-tuned the base RoBERTa model to our domain, i.e., wildlife Tweets. For these purposes, we used the wildlife-related Tweets, described in Section 2.6, that are not associated with coordinates. We used the masked language modelling technique to fine-tune RoBERTa, in which the model is trained to predict a subset of words that have been masked out [25]. This technique enables learning more contextually rich sentence representations, compared to earlier neural network models (see Section 1.1.2). It may be noted that the MLM technique is used for pre-training the base RobERTa model. The model was fine-tuned here for three epochs using the Hugging Face library [55] implementation for MLM.

2.3. Coordinate Prediction via Regression

We fine-tune RoBERTa to jointly predict latitude and longitude as a multivariate regression task. Given an input Tweet x, the model outputs a coordinate vector $\hat{\mathbf{y}} = (\hat{y}_{\text{lat}}, \hat{y}_{\text{lon}})$, where $\hat{y}_{\text{lat}}$ and $\hat{y}_{\text{lon}}$ are the predicted latitude and longitude, respectively. The model is trained using the Mean Squared Error (MSE) loss, consistent with [19]:

$$\mathcal{L}_{\text{MSE}} = \frac{1}{N}\sum_{i=1}^{N}\left\|\hat{\mathbf{y}}^{(i)} - \mathbf{y}^{(i)}\right\|^2 = \frac{1}{N}\sum_{i=1}^{N}\left[(\hat{y}_{\text{lat}}^{(i)} - y_{\text{lat}}^{(i)})^2 + (\hat{y}_{\text{lon}}^{(i)} - y_{\text{lon}}^{(i)})^2\right] \quad (1)$$

where N is the number of training examples, $\hat{\mathbf{y}}^{(i)}$ is the predicted coordinate vector, and $\mathbf{y}^{(i)}$ is the ground-truth coordinate vector for the i-th Tweet. This loss encourages the model to minimize the squared Euclidean distance between predicted and actual coordinates.

Figure 1 shows an overview of the regression-based method for coordinate prediction.

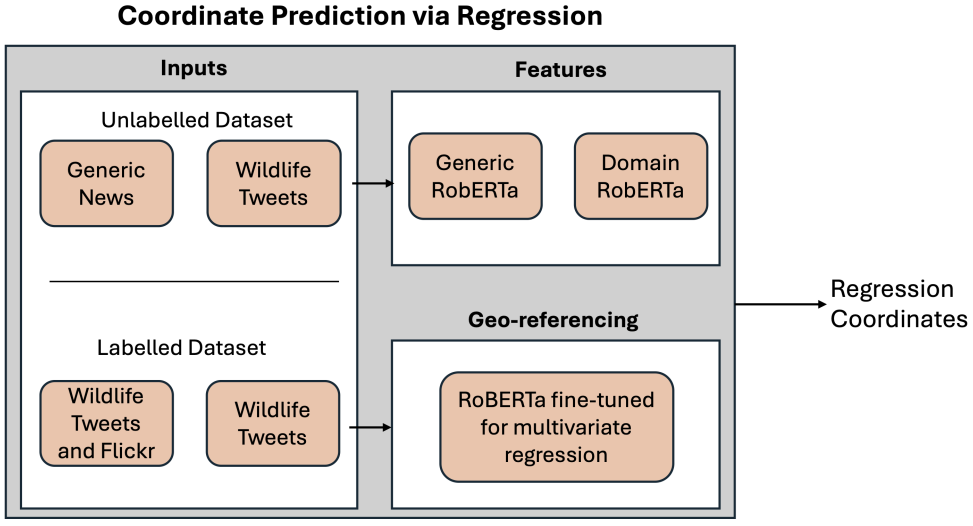

Figure 1. Regression-based approach for coordinate prediction.

Coordinate Normalization

Accurate coordinate prediction with neural models requires effective handling of geographic label distributions, which are often uneven and skewed. Without normalization,

neural networks tend to perform poorly on such data [18]. To mitigate this, we apply the coordinate normalization method proposed by Scherrer et al. [18], which improves regression performance on spatial data. Specifically, we normalize the latitude and longitude values by (1) Subtracting the mean of each coordinate dimension (latitude and longitude); (2) Applying joint scaling using the standard deviation computed over both dimensions combined.

Joint standardization helps maintain the relative scale and avoids introducing distortions that can result from normalizing each axis independently. This strategy has been shown to yield stable convergence and improved accuracy in coordinate prediction models [18]. Although the original work applied this method to BERT, we adopt the same procedure for RoBERTa, given their architectural similarities and shared regression interface.

2.4. Hybrid Enhancement I: Location Name Resolution

This hybrid approach combines NER-based location name detection and geocoding with the RoBERTa regression model (see Figure 2). The location names approach detects the presence of place names within the Tweets with NER, before mapping them to coordinates with a gazetteer. If the detected place name is ambiguous, and hence has more than one candidate location, disambiguation is performed by selecting the candidate closest to the coordinates returned by the RoBERTa regression method. When no place names are found in a Tweet, we use the RoBERTa regression model coordinates. However, our work is the first to incorporate recent neural network models for coordinates prediction and more traditional dictionary-based methods for performing location disambiguation.

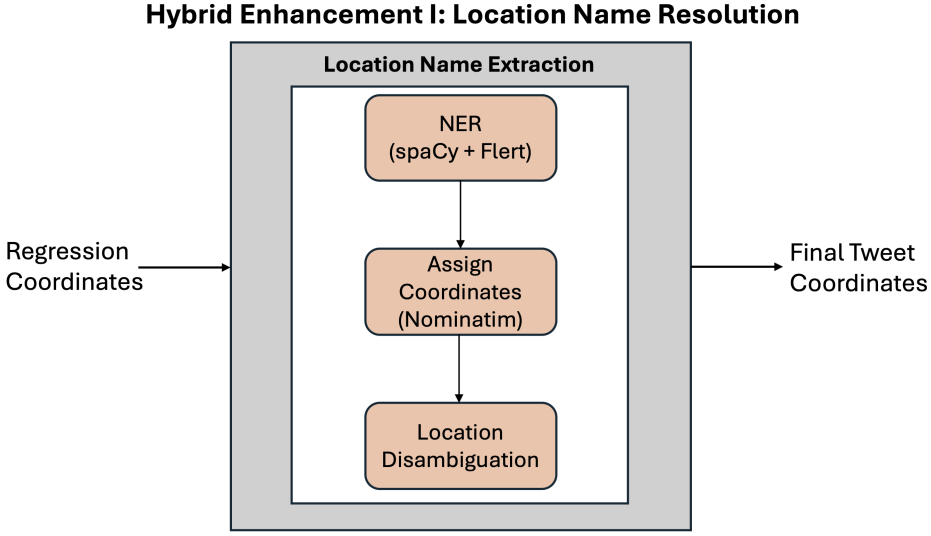

Figure 2. Hybrid enhancement I: location name resolution.

2.4.1. NER-Based Location Detection

We identify location names within the Tweets using two named entity recognition (NER) methods. An entity is regarded as a potential place name if it has one of the following NER labels: 'GPE', 'FAC', 'LOC', or 'ORG'. To improve the precision of the NER process, we apply voting between the methods, where a place name is considered genuine if both methods have identified it as a location using one of the above labels. Our first NER method uses the spaCy library [56], which has been successfully used for NER for short texts in previous research. We use a spaCy pre-trained transformer NER model, which is part of the library. We also use the Flert NER model [57] trained on a large English language corpus, available from https://huggingface.co/flair/ner-english-large (accessed on 13 May 2025). In an initial analysis, we experimented with the off-the-shelf Named Entity Recognition (NER) model based on BERT [24], trained on the CoNLL-2003 English news

articles dataset [58] (BERT NER model available at: https://huggingface.co/dslim/bert-base-NER, accessed on 13 May 2025). Our results showed that the latter model does not perform well for the given dataset. A possible reason for this is that the model has been adapted for NER with longer text sequences rather than social media posts.

2.4.2. Geocoding

We obtain the coordinates for each identified location name using the geocoding library Nominatim (Nominatim: https://nominatim.org, accessed on 13 May 2025). Nominatim uses OpenStreetMap data to find the coordinates for given location names.

2.4.3. Disambiguation

We perform location names disambiguation at two stages of the approach, also described in Algorithm 1:

- If a Tweet contains more than one place name, we select the name that refers to the fine-grained geographic object. We identify the fine-grained geographic location by selecting the most specific and complete location string. Through initial analysis of our dataset, we observed that finer-grained locations are often expressed as longer phrases with multiple place names separated by commas, effectively representing a more precise address—for example, "London, UK" rather than just "London" or "UK." Accordingly, we treat such multi-part location strings as the fine-grained geographic reference. For instance, in the Tweet "Today's photo of the day 'Goldie', Cold Ashby, Northamptonshire #goldfinch...," we select "Cold Ashby, Northamptonshire" as the fine-grained location rather than just "Cold Ashby" or "Northamptonshire" individually, as this combined name provides a more detailed and accurate geographic reference. This approach helps improve location name extraction and disambiguation by prioritizing the most specific location information available within the Tweet.
- If the selected place name is ambiguous, i.e., Nominatim returns multiple pairs of coordinates for the given place name, we use the Tweet coordinates obtained with the RoBERTa-based regression model to disambiguate the location. We calculate the distance between each pair of coordinates returned by Nominatim and the coordinates returned by the regression model. We then select the Nominatim-based coordinates that are closest to the regression-based coordinates. The distance is calculated using the Haversine formula.
- If a Tweet does not contain location names, then we use the coordinates returned by the regression model.

Algorithm 1 Location Name Disambiguation Heuristic

Input: Tweet
Output: Tweet(lat,lon)
if $Tweet.contains(loc\_name)$ **then**
  if $multipleloca\_name$ **then**
    return finest grain $loca\_name$
  end if
  if Nominatim returns multiple $loc\_name(lat, lon)$ **then**
    if $min((dist(loc\_name(lat, lon), regression(lat, lon)))$ **then**
      return loc_name(lat,lon)
    end if
  end if
else
  return regression(lat, lon)
end if

2.5. Hybrid Enhancement II: Semantic Similarity Matching

Semantic similarity-based methods are commonly used in combination with language modelling approaches in which having selected a predicted region, usually a grid cell or a spatial cluster, the aim is to find the training media item in the target cell or cluster that is most similar to the item to be georeferenced and use the coordinate of the training item as the prediction. We adapt this approach by using radial distances from the regression prediction coordinates to represent the predicted region. The steps of the approach, illustrated in Figure 3, are as follows:

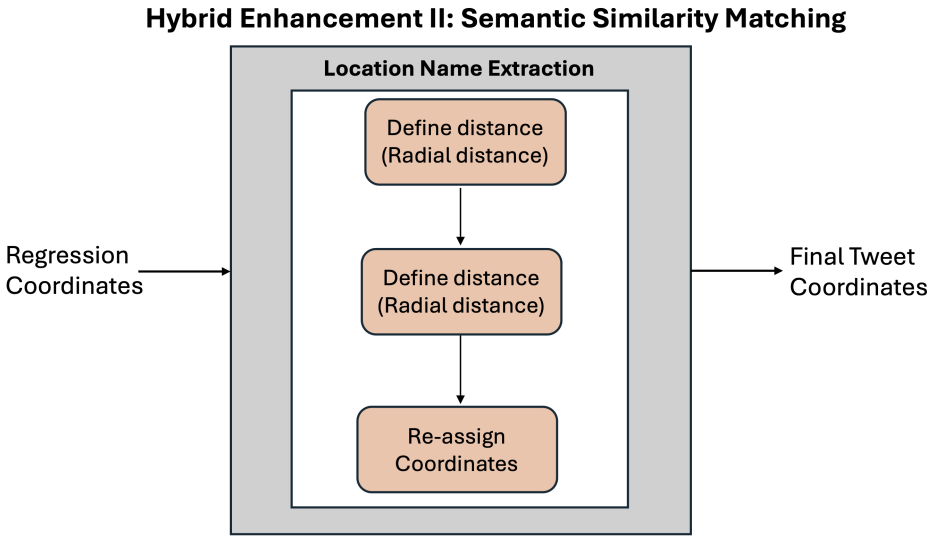

Figure 3. Hybrid enhancement II: semantic similarity matching.

- **Region Restriction:** For each unlabelled Tweet, we identify training samples within a 5 km, 10 km, or 20 km radius of the regression prediction. We selected these distances to balance spatial precision with the availability of sufficient nearby training samples for reliable semantic matching. These distances align with common geospatial scales that effectively capture local context in social media geo-referencing.
- **Semantic Matching:** Using Sentence-BERT [59], we compute cosine similarity between test and training samples. The SBERT model, trained using more than 1 billion training instances, is available from the Hugging Face library at https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2 (date of access: 31 March 2023). Sentence-BERT was chosen because its architecture is specifically designed to produce high-quality sentence embeddings optimized for semantic similarity tasks, making it more suitable than RoBERTa for this purpose. We also performed experiments with corpus-trained embeddings obtained with the fastText architecture [60]. However, the latter results were unsatisfactory.
- **Coordinate Assignment:** The coordinates of the most similar sample in the region are assigned to the test Tweet.
- **Optional Averaging:** We average the coordinates obtained using the two methods, i.e., regression and semantic similarity approach. We perform experiments with and without this final step.

2.6. Dataset Description

We collected Twitter/X and Flickr datasets limited to UK boundaries and related to wildlife observations. For these purposes, we used search phrases relevant to common and scientific names of various species within the UK (see Appendix A). The Tweets, which relate to the period 2007–2019, were gathered irrespective of whether they had geo-tags.

We retrieved the post, and any hashtags, mentions, and links for each Tweet. We used the labelled instances (i.e., instances with coordinates) for training prediction models, while the unlabelled instances (instances with no coordinates) were used to pre-train the RoBERTa model (see Table 1).

We downloaded Flickr data using the Flickr API interface for the period 2007–2019. We limited the search to geo-referenced Flickr posts because we use Flickr data only as a supplement to the fine-tuning stage where labelled data is required. Additionally, we downloaded only the text-related data (title, description, tags) with no exploitation of associated images.

Table 1. Overview of the social media datasets: Average number of tokens per instance (*Avg Length*); Number of instances with associated coordinates used for training prediction model (*#Instances (labelled)*); Number of instances used to train language model without associated coordinates (*#Instances (unlabelled)*).

	Twitter Dataset			Flickr Dataset	
	#Instances (Labelled)	**#Instances (Unlabelled)**	**Avg Length**	**#Instances (Labelled)**	**Avg Length**
Train	118,786	1,582,928	-	14,658	-
Dev	13,199	19,063	-	-	-
Test	14,666	-	-	-	-
Total	146,651	1,601,991	16	14,658	23

2.6.1. Training and Testing Data

Statistics about the datasets used for training language models and regression models are presented in Table 1, where the 'labelled' instances are instances associated with coordinates and are used as a training set for the regression models, while the 'unlabelled' instances are those used for pre-training the RoBERTa domain-specific language model. For evaluation purposes, we used a train, development, test split of approximately 80/10/10 for the Twitter dataset. We have obtained in total 146,651 labelled wildlife-related Tweets, which is, so far as we know, the largest collection of geo-referenced wildlife-related Twitter data available.

2.7. Evaluation Metrics

We use standard evaluation measures employed in previous related research on predicting coordinates for social media posts [2,61]. These are median error distance (MedianED) and mean error distance (MeanED). The measures are expressed with respect to the Distance Error $DE(t)$. For a given Tweet t, at location $loc_r(t)$, $DE(t)$ corresponds to either the Haversine distance (as used here) or Euclidean distance d between $loc_r(t)$ and an estimated location, $loc(t)$: $DE(m) = d(loc(t), loc_r(t))$. The MeanED refers to the average DE for each Tweet, while the MedianED is the median of DE for each Tweet.

2.8. Baseline Methods

We compare a regression model based on the RoBERTa language model with two other regression models employed in previous research on geo-referencing social media data. These are the Support Vector Regression (SVR) statistical machine learning algorithm and a BERT-based regression model. Additionally, we compare our approach to the work by [15], which has been shown to outperform a number of text -and network-based geolocation approaches and to perform well for small training sets. Finally, we compare our method to recently developed LLMs combined with prompting techniques. In summary, the baselines are

- Linear SVR: The SVR classifier is based on TF-IDF frequencies of character grams of length 3–10.

- BERT-based regression model: This is based on the work of [19] and uses a pre-trained BERT language model, trained on the generic dataset and then adapted for the regression task. This is the only work of which we are aware that uses transformer-based models in regression mode for coordinate prediction.
- GCN baseline [15] (described in Section 1.2): To enable comparison, we prepossessed our data to include the user network information. This consisted of retrieving additional metadata, specifically user mentions for the Tweets we present in Table 1.
- OpenAI GPT-4o model combined with zero- and five- shot prompting: The GPT-4o model by OpenAI is among the most advanced in natural language processing and is widely recognized for its strong performance in zero-shot and few-shot learning scenarios [62,63]. We combine GPT-4o with prompting techniques using only an instruction describing the task (zero-shot) and also by providing five randomly selected examples to the model along with the instruction (five-shot) (prompts available in Appendix B).
- LLaMA 3 model [64] combined with zero- and five- shot prompting: The LLaMa 3 model is known to be one of the most advanced open source language models [64]. We use the LLaMA 3 model with 8 billion parameters, pre-trained with instructions, downloaded from HuggingFace [55]. Similarly to the GPT-4o model, we perform experiments in zero- and five-shot settings. We use the same instruction and examples for both models (prompts available in Appendix B).

3. Results

3.1. Evaluation Experiments

We performed experiments with the RoBERTa model, pre-trained on a generic dataset, and a RoBERTa model that has been fine-tuned to the Twitter/X domain using the Tweets we have collected related to wildlife observations (see Table 1).

Currently, the RoBERTa architecture supports the regression task for single values using the Mean Square Loss function. We adapt RoBERTa to multivariate regression (for both latitude and longitude prediction), calculating the Mean Square Loss function per label, using the Huggingface implementation of RoBERTa for multi-label classification. The latitude and longitude values are predicted jointly. In order to train our regression model, we used 10 epochs, a batch size of 32 and we also saved only the model which performed the best for the development set. We present two approaches for improving the accuracy of geo-referencing models based on location name extraction and on semantic similarity between training and unlabelled instances. The development ('dev') set is used for identifying and saving the best-performing model, which is then used for assigning coordinates to the test instances.

3.2. Regression Results

Results from the performance of the regression models (see Table 2) show that the BERT model has a significant advantage over traditional machine learning models for geo-referencing social media content. Both the median and the mean error distances are much lower even for the baseline BERT-based regression model when compared to the Linear SVR model, with a margin exceeding 50 km on the test set (MedianED (BERT) = 94.90 km versus MedianED (Linear SVR) = 156.54 km and the MeanED (BERT) = 121.37 km versus MeanED (SVR) = 181.32 km). Notably, the BERT baseline also outperforms the GCN baseline for both median and mean error distance. Further, the transformer-based regression models perform very similarly for both the dev and test sets. Thus the models generalise well for unseen datasets, while the performance of the Linear SVR model declines significantly on the test set. For example, the best-performing Linear SVR shows a MedianED increase from 98.60 km on the dev set to 156.82 km on the test set, a difference of approximately 58 km.

Table 2. Results from regression model performance: 'generic' refers to a pre-trained publicly available language model trained using generic online datasets; 'wildlife Tweets' refers to a language model fine-tuned to the domain (wildlife-related Tweets); 'wildlife Tweets+combined training set' refers to a regression model that is using a RoBERTa model, fine-tuned to the domain and a training set, consisting of Twitter and Flickr data; 'NER + RoBERTa-based regression (Hybrid I: Location Name Resolution)' refers to the hybrid approach consisting of location name extraction and regression; 'semantic similarity + RoBERTa-based regression (Hybrid II: Semantic Similarity Matching)' refers to the hybrid approach consisting of semantic similarities and regression; 'best single NER model' refers to using a single best performing NER model (spaCy library) for location extraction as part of the hybrid approach; 'voting mechanism' refers to the voting approach where we perform voting between results obtained with both the spaCy NER library and Flert NER model. The best results are denoted in bold.

Model	Method	Dev Set		Test Set		Unlocated (%)
		MedianED	MeanED	MedianED	MeanED	
Linear SVR baseline	TF-IDF	98.60 km	127.03 km	156.82 km	181.32 km	0.0
BERT baseline	generic	93.63 km	119.34 km	94.90 km	121.37 km	0.0
GCN baseline	–	–	–	97.00 km	126.00 km	0.0
LLaMA 3	zero shot	–	–	87.86 km	363.45 km	21.94
LLaMA 3	five shot	–	–	81.86 km	301.46 km	21.27
GPT-4o	zero shot	–	–	62.93 km	244.826 km	0.46
GPT-4o	five shot	–	–	51.03 km	164.90 km	0.83
RoBERTa	generic	38.30 km	102.09 km	40.96 km	101.35 km	0.0
	wildlife Tweets	37.99 km	101.50 km	39.84 km	100.89 km	0.0
	wildlife Tweets + combined training set	36.81 km	101.05 km	38.04 km	100.44 km	0.0
semantic similarity + RoBERTa-based regression (Hybrid II: Semantic Similarity Matching)	5 km radial distance	-	-	38.24 km	100.36 km	0.0
	10 km radial distance	-	-	38.16 km	100.17 km	0.0
	20 km radial distance	-	-	38.78 km	100.26 km	0.0
NER + RoBERTa-based regression (Hybrid I: Location Name Resolution)	best single NER model	-	-	36.68 km	98.91 km	0.0
	voting mechanism	-	-	**36.47 km**	**98.22 km**	0.0

In contrast, the differences in the MedianED and MeanED values for the BERT and RoBERTa models between the dev and test sets are less than 3 km, demonstrating more stable performance. Table 2 further shows that the text generation models GPT-4o and LLaMA 3 consistently outperform other baselines in terms of MedianED, indicating their strong ability to infer locations in typical cases. However, their MeanED values are considerably higher, suggesting occasional large errors. This is likely due to difficulties these models face when tweets lack explicit location references, leading to inaccurate predictions. An additional limitation of these models, as well as prompting-based approaches, is their tendency to return a subset of tweets with invalid or missing coordinates (see Table 2, column *'Unlocated (%)'*). Despite these challenges, performance improves when a small number of illustrative examples are included in the prompt (i.e., the five-shot setting), highlighting the models' capacity to adapt quickly and generalize from limited data. However, the relatively small performance gap between the one-shot and five-shot settings, particularly for LLaMA 3, suggests that the model already captures essential geospatial cues from a single example and that additional examples may offer diminishing returns unless they are closely aligned with the input Tweet's context.

The results also demonstrate that using RoBERTa to build regression models for geo-referencing Tweets is more beneficial than using BERT (see Table 2). Even when the generic RoBERTa model is used, it still outperforms the BERT regression model where

the MedianED and the MeanED are about 50 km and 20 km, respectively, lower for the RoBERTa model than for BERT regression model. The reason for the better performance of RoBERTa versus BERT is that the RoBERTa model has been trained using a much larger training set than BERT. This indicates that using larger transformer models for regression, even when trained on generic datasets, is highly beneficial for the performance of regression models, particularly for sparse labelled data.

Fine-tuning a RoBERTa model on domain-specific data has resulted in additional improvements compared to the pre-trained RoBERTa, with 1–2 kilometers decrease in MedianED and MeanED ('RoBERTa generic' versus 'RoBERTa wildlife Tweets' in Table 2). Notably, the best-performing regression model uses a RoBERTa model fine-tuned to the domain and also a training set consisting of Twitter and Flickr posts, resulting in MedianED = 38.04 km and MeanED = 100.44 km for the test set. This indicates that augmenting the training corpus with labelled instances from diverse social network sites can be beneficial for building more accurate geo-referencing models for Twitter.

3.3. Analysis on Hybrid Approaches

As mentioned in Section 2, we developed two approaches for improving the regression models for geo-referencing Tweets. The approach based on radial distances and semantic similarity (described in Section 2.5) does lead to slightly lower MeanED values compared to the best-performing regression model, when a 10 km radial distance is used (see Figure 4 and Table 2), although there is no improvement in MedianED. However, as illustrated in Figure 5, the hybrid approach based on semantic similarity is notable for performing particularly well at the highest locational accuracy band, where the error is less than 5 km, especially when compared to the RoBERTa-based regression model.

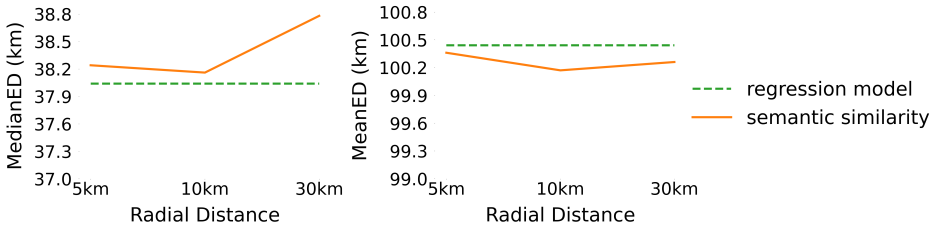

Figure 4. The effect of the radial distances on the performance of the Semantic Similarity + RoBERTa-based regression where the best performing RoBERTa-based regression model is used as a baseline.

In contrast, the hybrid location name extraction method that uses the RoBERTa regression method for disambiguation (described in Section 2.4) leads to marked overall improvement in performance relative to the best-performing RoBERTa-based regression model, achieving MedianED = 36.68 km and MeanED = 98.91 km when a single NER model is used for location name extraction. In this case, the spaCy NER model was used as it led to better geo-referencing results than the Flert NER method when combined with RoBERTa-based regression. Further, a voting procedure between the two location names extraction methods, the spaCy NER library and Flert NER model, led to additional improvements, with MedianED = 36.47 km and MeanED = 98.22 km. The error distribution results, presented in Figure 5, illustrate the fact that while the hybrid approach based on location name extraction and regression increases the accuracy of geo-referencing models significantly for all error distances up to 95 km, the improvement is most marked within distances of 5 km. These findings are also reflected in our accuracy-based analysis, presented in Figure 6. Note that lower scores for MeanED and MedianED indicate better performance, whereas higher scores for accuracy indicate improvement.

Additionally, a comparison between the NER voting-based approach (which uses a gazetteer to obtain coordinates) and the RoBERTa-based regression model (see Table 3)

showed that, for the subset of posts in which place names are detected, the NER/gazetteer method using RoBERTa-based regression for location disambiguation ('NER + regression-based disambiguation') outperforms the purely transformer-based regression models and the pure location name extraction method ('NER + Nominatim-based disambiguation') for geo-referencing social media posts, obtaining a median ED of 1.32 km. For this restricted dataset, in which all Tweets contain a place name, a smaller MedianED of 0.86km can be obtained by informing the Nominatum geocoder of the UK context ('NER + Nominatim-based disambiguation with UK context'), as opposed to letting the regression method perform the disambiguation, though in this case the MeanED is still slightly inferior to using regression-based disambiguation. Without such prior knowledge, however, the regression-based disambiguation approach is clearly beneficial overall for geo-referencing Tweets. It should also be stressed that using only the NER method with gazetteer coordinates has a major limitation for datasets such as the one employed here as only about 5% of the Tweets contain detectable place names (see Table 4). For the remaining 95% of the dataset, the NER/gazetteer method fails entirely. Our hybrid approach that uses coordinates both from gazetteers and predicted from regression is therefore clearly advantageous for such datasets.

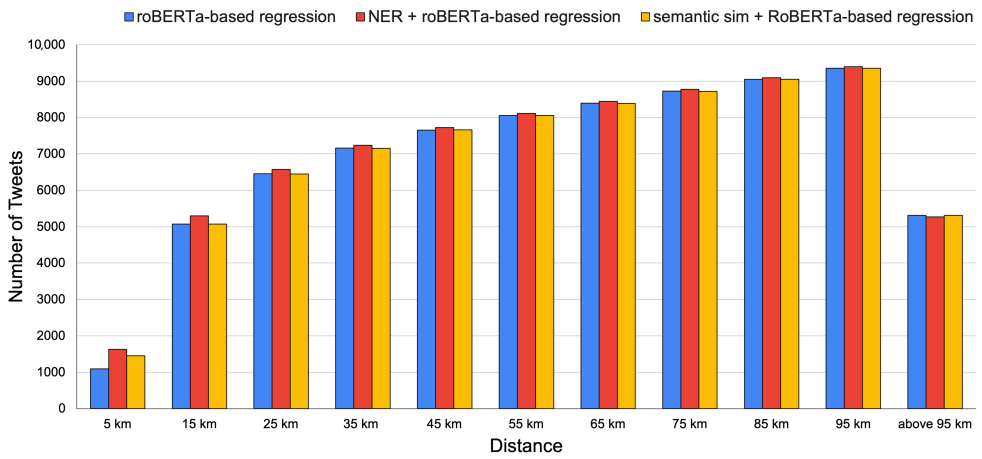

Figure 5. Distribution of error results showing proportion of results within 5 km, 15 km, etc. for the approaches RoBERTa-based regression, NER + RoBERTa-based regression, and Semantic Similiarity + RoBERTa-based regression.

Table 3. A comparison between different location name disambiguation techniques for the location name extraction approach just for those 872 Tweets in which place names could be detected, where 'NER+regression-based disambiguation' refers to using the RoBERTa-based regression model for performing location disambiguation, 'NER + Nominatim-based disambiguation' refers to using the top-ranked location returned by Nominatim for a given place name, 'NER + Nominatim-based disambiguation with UK context' refers to using the the top-ranked location returned by Nominatim but limiting the search to UK-based locations, and 'RoBERTa-based Regression' refers to using only regression for inferring the coordinates for the Tweets.

Method	MedianED	MeanED
NER + regression-based disambiguation	1.32 km	39.22 km
NER + Nominatim-based disambiguation	1.85 km	50.27 km
NER + Nominatim-based disambiguation with UK context	0.86 km	39.28 km
RoBERTa-based regression	14.95 km	59.83 km

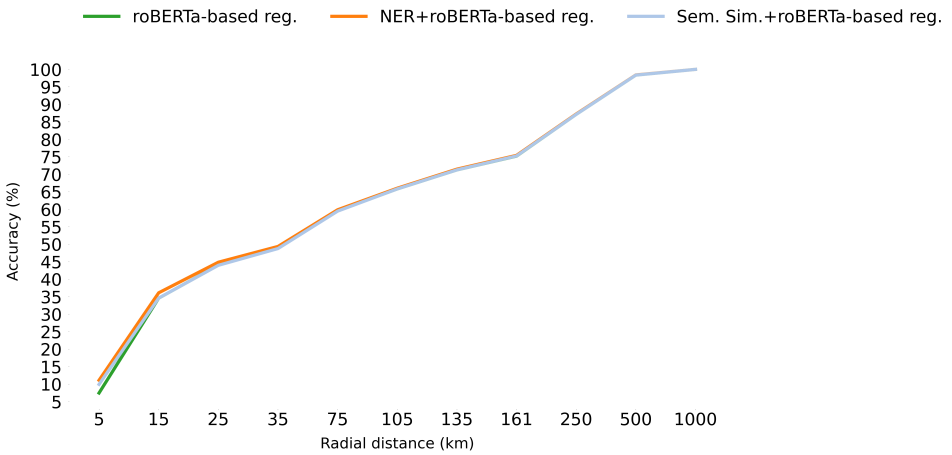

Figure 6. Accuracy within 5 km, 15 km, etc. of the actual location for the approaches RoBERTa-based regression, NER + RoBERTa-based regression, and Semantic Similarity + RoBERTa-based regression ('reg.' stands for regression).

Table 4. Data distribution for the two hybrid approaches, i.e., semantic similarity approach ('*App.*' refers to approach), '*#Tweets*' refers to number of Tweets.

	-	**#Tweets**
Locations Names Extraction	Tweets with detected location names	872
	Tweets with no detected location names	13,794
Semantic Similarities	Training instances within region (10 km)	14,308
	No training instances within region (10 km)	358
Total number of Tweets		14,666

4. Discussion

Geo-referencing social media data, especially Tweets, presents a unique set of challenges. These include modelling informal language, limited context, and the limited presence of explicit place names. Our results highlight several key issues in these respects and demonstrate how different methodological choices impact the effectiveness of location prediction.

One major unexplored research area is the use of regression models for coordinates prediction versus cluster, region-based approaches. As shown in Sections 3.2 and 3.3, our use of a transformer language model-based regression approach, especially when combined with a location name disambiguation strategy, outperforms four baseline methods, including a conventional SVR model, a graph convolutional network (GCN), and larger generative models such as GPT-4o and LLaMA 3. These results suggest that the contextual language representation of MLM transformer-based regression, particularly using the RoBERTa architecture, offers significant advantages for predicting Tweet coordinates. Notably, RoBERTa outperformed BERT, indicating that when domain-specific training data is limited, more robust pre-trained models are better suited to capture the subtle linguistic cues needed for geolocation tasks. Moreover, we find that domain adaptation and task-specific fine-tuning, even on relatively small corpora, can significantly improve performance, confirming the utility of transfer learning in this context. The domain-adapted RoBERTa model developed in this study can potentially benefit other Twitter-based geospatial or wildlife monitoring tasks.

Another critical issue is data sparsity, especially in the availability of geo-annotated training examples. Augmenting training data with additional sources from other social media platforms helped mitigate this problem. Our approach relates to prior work by [20,65], which showed the value of using Flickr and Twitter data for geolocating Wikipedia content. A key

advantage of our method is its simplicity: unlike traditional statistical models, our neural network-based approach does not require complex pre-processing such as feature vector normalization. Future work could extend this by using even more diverse and multilingual data sources to pre-train or fine-tune language models for greater robustness and coverage.

Disambiguation of place names remains a persistent obstacle. Our hybrid strategies, discussed in Sections 2.4, 2.5, and 3.3, combine semantic similarity and location name extraction with regression. These techniques address the ambiguity of place names and improve precision, especially for Tweets where names appear in isolation or contextually ambiguous forms. The best-performing configuration integrates location name extraction with RoBERTa-based regression, which yields improved precision in predicting coordinates with distance errors below 5 km. However, a major limitation of NER/gazetteer-based approaches is their dependency on the presence of identifiable place names, which are found in practice in only a small subset of Tweets. In contrast, transformer-based regression methods are capable of assigning coordinates to all inputs, independent of explicit location mentions.

To further explore this capability, we analysed Tweets where NER failed to detect location names. Interestingly, the regression model predicted coordinates within 15 km for 32% of these Tweets and within 55 km for 51%, highlighting its broader applicability. The spatial distribution analysis in Figure 7 also indicates that the model introduces minimal geographic bias in its predictions.

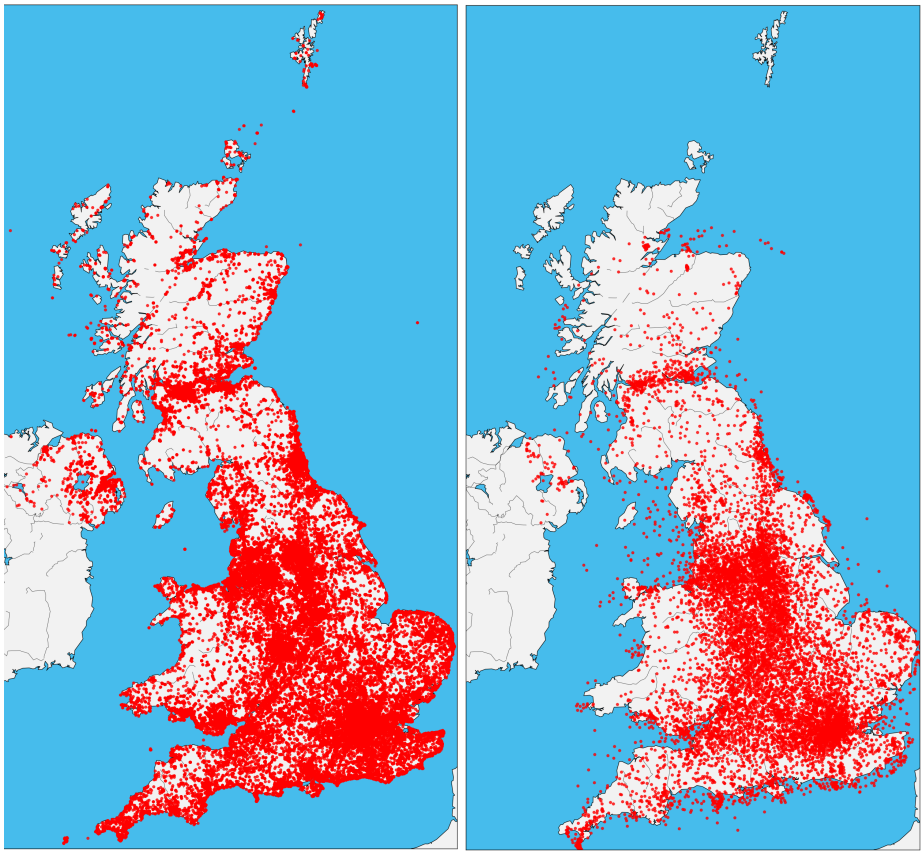

Figure 7. A comparison of the geographical distribution of original Tweets coordinates (**left**) versus predicted coordinates (**right**).

Examples of Tweets where NER-based methods fail and where regression succeeds are given in Table 5. An obvious characteristic of Tweets that cannot be geo-referenced with NER and gazetteers is illustrated with examples such as: *'This Herring gull was harassing returning guillemots to give up their catch. #wildlife...'*, *'@BBCSpringwatch saw a jackdaw this evening eating seeds at our bird feeders. This is a first for us. Is this normal behaviour?'*, and *'Down*

side to lots of #Clover in your lawn? Bee sting in your foot, that's what #ouch' is that there are no detectable place names. They often include fine-grained locations using generic place words such as 'our bird feeders' and 'lawn' that cannot be associated with coordinates using gazetteer methods. Such descriptions are also a challenge for the language model methods, but in some cases surprisingly good results can be obtained (as in the second example in Table 5), which can be attributed to the locations being learnt from similar language in the training examples. There are other situations in which Tweets include actual locations that have been misclassified by the NER approaches, for example, because of adjectives attached to the proper location names, as in *'Lovely #daffodils @ Sunny Adlington'*, where the NER methods have labelled the phrase 'Sunny Adlington' as a person. Another failure of the NER/gazetteer methods occurs when Tweets are associated with locations that have not been identified by the pre-trained NER methods and are not present in the gazetteer, such as "Uttoxeter Quarry" in the Table 5 example *'@Staffsbirdnews Uttoxeter Quarry: Common Tern,Common Sand, 4 Green Sand, 4 Snipe, 3 Pintail, 19Wigeon, 4 Pochard and 2 Blue Snow Geese'*.

Table 5. Examples of Tweets for which the regression model performed well but the NER/gazetteer location extraction-based approach failed.

Tweet	Dist. Error (km)
13 spoonbills and one with a avocet sitting on ones head @RSPBtitch-wellmarsh	4.00 km
Morning all. Yes indeed, it's a marshmallow world again round here. Deep joy. And pity me poor Robin; Blackbird on their nests!	2.69 km
Great Black Backed Gull spotted on 09-Jul-2013. Sent from Birds of Britain HD app by @CleverMatrix	2.71 km
@Staffsbirdnews Uttoxeter Quarry: Common Tern, Common Sand, 4 Green Sand, 4 Snipe, 3 Pintail, 19 Wigeon, 4 Pochard and 2 Blue Snow Geese	3.89 km
What beauty, Buddleja and a Peacock butterfly! #buddleja #buddleia #butterflybush #peacockbutterfly #beauty #nature #garden #betwsycoed	4.72 km
@Staffsbirdnews Uttoxeter Quarry: Redstart, Black-tailed Godwit, 3 Green Sand, 6 Common Sand, 5 LRP, Willow Tit	1.18 km
Tiny bee type thingy on my pink daisy #beetypething #tinybee #pinkdaisy #daisy #pink #gardening #gardensofinstagram #lbloggers #lbloggersuk #instagarden #growyourown #plants #plantsofinstagram #gbloggersuk…	3.26 km
discovered today that there's a #wren pair #nesting in our #compost bin! #eye_spy_birds @Natures_Voice @GWmag @bbcspringwatch birdsofinstaqram best_birds_of_world @chesterelements #wren	3.56 km
#wmbirdclub #Belvide 12/10: 68 Golden; 3 Ringed Plover, Ruff, 8 Dunlin, 40 Gadwall, 27 Shoveler, 14 Wigeon, 163 Teal; 55 Pochard.	0.43 km

In future, it is possible to envisage that such false negatives for the NER/gazetteer method could be reduced by improved training of the NER methods with location-rich Twitter data, as well as access to richer gazetteer resources. A significant advantage of the regression-based transformer model is its ability to assign coordinates to such Tweets (especially those that do not mention gazetteered place names), based on learned trends from the training set.

5. Conclusions

This paper has addressed the problem of geo-referencing Twitter data related to wildlife observation, using only the text message. A main challenge in developing machine learning methods for geo-referencing social media data is the need for large amounts of data that have coordinates that can be used in a training dataset. However, many social media posts such as Twitter/X lack coordinate information, especially when search is limited to a specific region or topic. We addressed this challenge by proposing a simple but effective method for augmenting the Twitter training data with a Flickr dataset. We also used neural network transformer methods for building multivariate regression models and word representations, which to date have received only limited attention in geo-referencing social media studies, especially when related to wildlife. In particular, we adapted the RoBERTa transformer model for the regression task. Results showed that the RoBERTa model, when fine-tuned to the domain and when augmenting the training set with diverse social media sources, can be highly beneficial for geo-referencing Tweets. Finally, we demonstrated that a hybrid approach of using NER methods and a gazetteer for geocoding when place names are present, supported by the RoBERTa-based regression for disambiguation, and using only the regression method when place names are absent, can significantly improve the performance of geo-referencing models. This benefit is most noticeable when generating the highest accuracy results.

Author Contributions: Conceptualization, Thomas Edwards, Christopher B. Jones, and Padraig Corcoran; methodology, Thomas Edwards; software, Thomas Edwards; validation, Thomas Edwards, Christopher B. Jones, and Padraig Corcoran; formal analysis, Thomas Edwards; investigation, Thomas Edwards; resources, Thomas Edwards; data curation, Thomas Edwards; writing—original draft preparation, Thomas Edwards; writing—review and editing, Thomas Edwards, Christopher B. Jones, and Padraig Corcoran; visualization, Thomas Edwards; supervision, Christopher B. Jones and Padraig Corcoran; project administration, Christopher B. Jones and Padraig Corcoran. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding

Data Availability Statement: The data and code are available at Figshare: https://figshare.com/s/556f6fe76685a38932b8 (accessed on 13 May 2025).

Acknowledgments: During the preparation of this study, the author(s) used GPT-4o and LLaMA 3 for the purposes of predicting coordinates for Tweets as part of the analysis. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest

Appendix A. Keywords Used for Data Collection

We collected Tweets using search phrases of common and scientific species names to create a dataset for the invasive species in the UK with occurrences on the NBN data portal, as well as the 10 most numerous species on NBN, and 10 ten most numerous species on Flickr, some of which overlap. This collection has been informed by previous research within the area [66,67]. The keywords are given in Table A1.

Table A1. Examples of keywords used to collect Twitter/X dataset.

Scientific Name	Common Name
Fagus sylvatica	Beech
Gallinago gallinago	Snipe
Parus major	Great Tit
Pteridium aquilinum	Bracken
Cyanistes caeruleus	Blue Tit
Hedera helix	Ivy
Bellis perennis	Daisy
Turdus merula	Blackbird
Scirurus carolinensis	Grey squirrel
Fringilla coelebs	Chaffinch
Passer domesticus	House Sparrow
Anas platyrhynchos	Mallard
Columba palumbus	Woodpigeon
Chloris chloris	Greenfinch
Prunella modularis	Dunnock
Taraxacum officinale agg.	Dandelion
Heracleum mantegazzianum	Giant Hogweed
Hyacinthoides non-scripta	Bluebell
Branta canadensis	Canada Goose
Aix sponsa	Wood Duck

Appendix B. Prompts Used for GPT-4o and LLaMA 3 Models

We used the same prompt to geo-reference Tweets for both models. The prompt was designed based on design principles from OpenAI and Meta, as well as examples provided by Reynolds and McDonell [68]. We used the same prompt for zero- and few-shot settings. However, for few-shot settings, we added five randomly selected examples from the training data.

Prompt used for geo-referencing Tweets

Given the following tweet, provide the location name like 'Location name:' (if mentioned or inferred), followed by the latitude ('Latitude:') and longitude ('Longitude') values each on a separate line. If multiple locations are mentioned, return the coordinates for the most relevant or prominent one. If no explicit location is mentioned, infer the coordinates based on other clues, such as references to landmarks, events, or notable geographical details. Prioritize locations within the United Kingdom and Southern Ireland, if applicable. If the location is not fully clear, provide the best guess and explain the reasoning behind the uncertainty (using 'Reason:' as a last row of the response). Ensure that a result is returned, even if the confidence in the coordinates is low.

References

1. Di Rocco, L.; Dassereto, F.; Bertolotto, M.; Buscaldi, D.; Catania, B.; Guerrini, G. Sherloc: A knowledge-driven algorithm for geolocating microblog messages at sub-city level. *Int. J. Geogr. Inf. Sci.* **2021**, *35*, 84–115. [CrossRef]
2. Zheng, X.; Han, J.; Sun, A. A survey of location prediction on twitter. *IEEE Trans. Knowl. Data Eng.* **2018**, *30*, 1652–1671. [CrossRef]
3. Stock, K. Mining location from social media: A systematic review. *Comput. Environ. Urban Syst.* **2018**, *71*, 209–240. [CrossRef]
4. Amano, T.; Lamming, J.D.; Sutherland, W.J. Spatial gaps in global biodiversity information and the role of citizen science. *Bioscience* **2016**, *66*, 393–400. [CrossRef]
5. Barve, V. Discovering and developing primary biodiversity data from social networking sites: A novel approach. *Ecol. Inform.* **2014**, *24*, 194–199. [CrossRef]
6. Middleton, S.E.; Kordopatis-Zilos, G.; Papadopoulos, S.; Kompatsiaris, Y. Location extraction from social media: Geoparsing, location disambiguation, and geotagging. *ACM Trans. Inf. Syst. (TOIS)* **2018**, *36*, 1–27. [CrossRef]
7. Paraskevopoulos, P.; Palpanas, T. Fine-grained geolocalisation of non-geotagged tweets. In Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015, Paris, France, 25–28 August 2015; pp. 105–112.
8. Kelm, P.; Murdock, V.; Schmiedeke, S.; Schockaert, S.; Serdyukov, P.; Van Laere, O. Georeferencing in social networks. In *Social Media Retrieval*; Springer: Berlin/Heidelberg, Germany, 2013; pp. 115–141.
9. Melo, F.; Martins, B. Automated geocoding of textual documents: A survey of current approaches. *Trans. GIS* **2017**, *21*, 3–38. [CrossRef]
10. Li, J.; Qian, X.; Lan, K.; Qi, P.; Sharma, A. Improved image GPS location estimation by mining salient features. *Signal Process. Image Commun.* **2015**, *38*, 141–150. [CrossRef]
11. Inkpen, D. Text mining in social media for security threats. In *Recent Advances in Computational Intelligence in Defense and Security*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 491–517.
12. Bassi, J.; Manna, S.; Sun, Y. Construction of a geo-location service utilizing microblogging platforms. In Proceedings of the 2016 IEEE Tenth International Conference on Semantic Computing (ICSC), Laguna Hills, CA, USA, 3–5 February 2016; pp. 162–165.
13. Van Laere, O.; Schockaert, S.; Dhoedt, B. Georeferencing Flickr resources based on textual meta-data. *Inf. Sci.* **2013**, *238*, 52–74. [CrossRef]
14. Kulkarni, S.; Jain, S.; Hosseini, M.J.; Baldridge, J.; Ie, E.; Zhang, L. Spatial Language Representation with Multi-Level Geocoding. *arXiv* **2020**, arXiv:2008.09236. [CrossRef]
15. Rahimi, A.; Cohn, T.; Baldwin, T. Semi-supervised User Geolocation via Graph Convolutional Networks. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Melbourne, Australia, 15–20 July 2018; pp. 2009–2019.
16. Han, B.; Cook, P.; Baldwin, T. Text-based twitter user geolocation prediction. *J. Artif. Intell. Res.* **2014**, *49*, 451–500. [CrossRef]
17. Zhou, F.; Qi, X.; Zhang, K.; Trajcevski, G.; Zhong, T. MetaGeo: A General Framework for Social User Geolocation Identification With Few-Shot Learning. *IEEE Trans. Neural Netw. Learn. Syst.* **2023**, *34*, 8950–8964. [CrossRef] [PubMed]
18. Scherrer, Y.; Ljubešić, N. HeLju@ VarDial 2020: Social media variety geolocation with BERT models. In Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects. International Committee on Computational Linguistics (ICCL), Online, 13 December 2020; pp. 202–211.
19. Scherrer, Y.; Ljubešić, N.; Tiedemann, J.; Scherrer, Y.; Jauhiainen, T. Social media variety geolocation with geobert. In Proceedings of the Eighth Workshop on NLP for Similar Languages, Varieties and Dialects. The Association for Computational Linguistics, Kiyv, Ukraine, 20 April 2021; pp. 135–140.
20. Van Laere, O.; Schockaert, S.; Tanasescu, V.; Dhoedt, B.; Jones, C.B. Georeferencing Wikipedia Documents Using Data from Social Media Sources. *ACM Trans. Inf. Syst.* **2014**, *32*, 1–32. [CrossRef]
21. DeLozier, G.; Baldridge, J.; London, L. Gazetteer-independent toponym resolution using geographic word profiles. In Proceedings of the AAAI Conference on Artificial Intelligence, Austin, TX, USA, 25–30 January 2015; Volume 29.
22. Jauhiainen, T.; Lui, M.; Zampieri, M.; Baldwin, T.; Lindén, K. Automatic language identification in texts: A survey. *J. Artif. Intell. Res.* **2019**, *65*, 675–782. [CrossRef]
23. Jeawak, S.S.; Jones, C.B.; Schockaert, S. Predicting the environment from social media: A collective classification approach. *Comput. Environ. Urban Syst.* **2020**, *82*, 101487. [CrossRef]
24. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. 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, Volume 1 (Long and Short Papers), Minneapolis, MN, USA, 2–7 June 2019; pp. 4171–4186.
25. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. Roberta: A robustly optimized bert pretraining approach. *arXiv* **2019**, arXiv:1907.11692.
26. Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I. Language models are unsupervised multitask learners. *OpenAI Blog* **2019**, *1*, 9.

27. Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv* **2023**, arXiv:2307.09288. [CrossRef]
28. Schick, T.; Schütze, H. Exploiting Cloze-Questions for Few-Shot Text Classification and Natural Language Inference. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, Online, 19–23 April 2021; pp. 255–269.
29. Le Scao, T.; Rush, A.M. How many data points is a prompt worth? In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Online, 6–11 June 2021; pp. 2627–2636.
30. Zheng, C.; Jiang, J.Y.; Zhou, Y.; Young, S.D.; Wang, W. Social media user geolocation via hybrid attention. In Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, Online, 25–30 July 2020; pp. 1641–1644.
31. Cheng, Z.; Caverlee, J.; Lee, K. You are where you tweet: A content-based approach to geo-locating twitter users. In Proceedings of the 19th ACM International Conference on Information and Knowledge Management, Toronto, Canada, 26–30 October 2010; pp. 759–768.
32. Melo, F.; Martins, B. Geocoding textual documents through the usage of hierarchical classifiers. In Proceedings of the 9th Workshop on Geographic Information Retrieval, Paris, France, 26–27 November 2015; pp. 1–9.
33. Eisenstein, J.; O'Connor, B.; Smith, N.A.; Xing, E.P. A Latent Variable Model for Geographic Lexical Variation. In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, Cambridge, MA, USA, 9–11 October 2010; pp. 1277–1287.
34. Fornaciari, T.; Hovy, D. Identifying Linguistic Areas for Geolocation. In Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019), Hong Kong, China, 4 November 2019; pp. 231–236. [CrossRef]
35. Kordopatis-Zilos, G.; Papadopoulos, S.; Kompatsiaris, I. Geotagging Text Content With Language Models and Feature Mining. *Proc. IEEE* **2017**, *105*, 1971–1986. [CrossRef]
36. Wing, B.; Baldridge, J. Simple supervised document geolocation with geodesic grids. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Portland, OR, USA, 19–24 June 2011; pp. 955–964.
37. Hulden, M.; Silfverberg, M.; Francom, J. Kernel density estimation for text-based geolocation. In Proceedings of the AAAI Conference on Artificial Intelligence, Austin, TX, USA, 25–30 January 2015; Volume 29, pp. 145–150.
38. Serdyukov, P.; Murdock, V.; Van Zwol, R. Placing flickr photos on a map. In Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, Boston, MA, USA, 19–23 July 2009; pp. 484–491.
39. Masis, T.; O'Connor, B. Where on Earth Do Users Say They Are?: Geo-Entity Linking for Noisy Multilingual User Input. In Proceedings of the Sixth Workshop on Natural Language Processing and Computational Social Science (NLP+ CSS 2024), Mexico City, Mexico, 21 June 2024; pp. 86–98.
40. Yan, Z.; Yang, C.; Hu, L.; Zhao, J.; Jiang, L.; Gong, J. The Integration of Linguistic and Geospatial Features Using Global Context Embedding for Automated Text Geocoding. *ISPRS Int. J. Geo-Inf.* **2021**, *10*, 572. [CrossRef]
41. Gritta, M.; Pilehvar, M.T.; Collier, N. Which Melbourne? Augmenting Geocoding with Maps. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Melbourne, Australia, 15–20 July 2018; Gurevych, I., Miyao, Y., Eds.; Association for Computational Linguistics: Melbourne, Australia, 2018; pp. 1285–1296. [CrossRef]
42. Cardoso, A.B.; Martins, B.; Estima, J. Using recurrent neural networks for toponym resolution in text. In *EPIA Conference on Artificial Intelligence*; Springer: Cham, Switzerland, 2019; pp. 769–780.
43. Molina-Villegas, A.; Muñiz-Sanchez, V.; Arreola-Trapala, J.; Alcántara, F. Geographic Named Entity Recognition and Disambiguation in Mexican News using word embeddings. *Expert Syst. Appl.* **2021**, *176*, 114855. [CrossRef]
44. Fize, J.; Moncla, L.; Martins, B. Deep Learning for Toponym Resolution: Geocoding Based on Pairs of Toponyms. *ISPRS Int. J. Geo-Inf.* **2021**, *10*, 818. [CrossRef]
45. Nakov, P.; Ritter, A.; Rosenthal, S.; Sebastiani, F.; Stoyanov, V. SemEval-2016 task 4: Sentiment analysis in Twitter. *arXiv* **2019**, arXiv:1912.01973. [CrossRef]
46. Nguyen, D.Q.; Vu, T.; Nguyen, A.T. BERTweet: A pre-trained language model for English Tweets. *arXiv* **2020**, arXiv:2005.10200. [CrossRef]
47. Born, J.; Manica, M. Regression Transformer: Concurrent Conditional Generation and Regression by Blending Numerical and Textual Tokens. In Proceedings of the ICLR2022 Machine Learning for Drug Discovery, Virtual, 25 April 2022.
48. Simanjuntak, L.F.; Mahendra, R.; Yulianti, E. We know you are living in bali: Location prediction of twitter users using bert language model. *Big Data Cogn. Comput.* **2022**, *6*, 77. [CrossRef]
49. Li, M.; Lim, K.H.; Guo, T.; Liu, J. A Transformer-Based Framework for POI-Level Social Post Geolocation. In Proceedings of the Advances in Information Retrieval–45th European Conference on Information Retrieval, ECIR 2023, Lecture Notes in Computer Science Dublin, Ireland, 2–6 April 2023; Kamps, J., Goeuriot, L., Crestani, F., Maistro, M., Joho, H., Davis, B., Gurrin, C., Kruschwitz, U., Caputo, A., Eds.; Proceedings Part I; Springer: Cham, Switzerland, 2023; Volume 13980, pp. 588–604. [CrossRef]

50. Bhandari, P.; Anastasopoulos, A.; Pfoser, D. Are large language models geospatially knowledgeable? In Proceedings of the 31st ACM International Conference on Advances in Geographic Information Systems, Hamburg, Germany, 13–16 November 2023; pp. 1–4.
51. Sultanov, A. Leveraging Large Language Models for Textual Geotagging: A Novel Approach to Location Inference. *Comput. Tools Educ.* **2024**, *3*, 48–65. [CrossRef]
52. Tucker, S. A systematic review of geospatial location embedding approaches in large language models: A path to spatial AI systems. *arXiv* **2024**, arXiv:2401.10279. [CrossRef]
53. Han, B.; Cook, P.; Baldwin, T. Geolocation prediction in social media data by finding location indicative words. In Proceedings of the COLING 2012, Mumbai, India, 8–15 December 2012; pp. 1045–1062.
54. Wing, B.; Baldridge, J. Hierarchical discriminative classification for text-based geolocation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; pp. 336–348.
55. Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; et al. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *arXiv* **2019**, arXiv:1910.03771.
56. Honnibal, M.; Montani, I. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. *Appear* **2017**, *7*, 411–420.
57. Schweter, S.; Akbik, A. FLERT: Document-Level Features for Named Entity Recognition. *arXiv* **2020**, arXiv:2011.06993.
58. Tjong Kim Sang, E.F.; De Meulder, F. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003, Edmonton, Canada, 31 May–1 June 2003; pp. 142–147.
59. Reimers, N.; Gurevych, I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, 3–7 November 2019; pp. 3982–3992.
60. Bojanowski, P.; Grave, E.; Joulin, A.; Mikolov, T. Enriching word vectors with subword information. *Trans. Assoc. Comput. Linguist.* **2017**, *5*, 135–146. [CrossRef]
61. Gritta, M.; Pilehvar, M.T.; Collier, N. A pragmatic guide to geoparsing evaluation. *Lang. Resour. Eval.* **2019**, *54*, 683–712. [CrossRef] [PubMed]
62. Savelka, J.; Ashley, K.D.; Gray, M.A.; Westermann, H.; Xu, H. Can gpt-4 support analysis of textual data in tasks requiring highly specialized domain expertise? *arXiv* **2023**, arXiv:2306.13906. [CrossRef]
63. Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 1877–1901.
64. Grattafiori, A.; Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Vaughan, A.; et al. The llama 3 herd of models. *arXiv* **2024**, arXiv:2407.21783. [CrossRef]
65. De Rouck, C.; Van Laere, O.; Schockaert, S.; Dhoedt, B. Georeferencing Wikipedia pages using language models from Flickr. In Proceedings of the 10th International Semantic Web Conference (ISWC 2011), Bonn, Germany, 23–27 October 2011; pp. 1–8.
66. Edwards, T.; Jones, C.B.; Perkins, S.E.; Corcoran, P. Passive citizen science: The role of social media in wildlife observations. *PLoS ONE* **2021**, *16*, e0255416. [CrossRef]
67. Edwards, T.; Jones, C.B.; Corcoran, P. Identifying wildlife observations on twitter. *Ecol. Informatics* **2022**, *67*, 101500. [CrossRef]
68. Reynolds, L.; McDonell, K. Prompt programming for large language models: Beyond the few-shot paradigm. In Proceedings of the Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems, Virtual, 8–13 May 2021; pp. 1–7.