Augmenting Official Citizen Science Data Collections with Social Media Data Related to Wildlife Observations

A thesis submitted in partial fulfilment

of the requirement for the degree of Doctor of Philosophy

Thomas J. Edwards

February 2022

Cardiff University
School of Computer Science & Informatics

Declaration

This work has not previously been accepted in substance for any degree and is not concurrently submitted in candidature for any degree.

Signed .... (candidate)

Date ....

Statement 1

This thesis is being submitted in partial fulfillment of the requirements for the degree of PhD.

Signed .... (candidate)

Date ....

Statement 2

This thesis is the result of my own independent work/investigation, except where otherwise stated. Other sources are acknowledged by explicit references.

Signed .... (candidate)

Date ....

Statement 3

I hereby give consent for my thesis, if accepted, to be available for photocopying and for inter-library loan, and for the title and summary to be made available to outside organisations.

Signed .... (candidate)

Date ....

.

To my wife Aleksandra and daughter Elizabeth
Without their love, support and kindness this study would not have been completed

Abstract

Studies of wildlife species distribution patterns are increasingly important in the face of rapid ecosystem changes that have implications for disease emergence and spread, food security, climate change, and invasive species biology. Citizen science campaigns can be very effective for observing wildlife behaviour, but they can also be a resource-consuming process and limited in coverage and sometimes their accuracy. Due to their wide usage, social media platforms represent an untapped source of potentially valuable wildlife observational data which is less costly to obtain but could complement citizen science data collections and support real-time species monitoring and analysis. There are however concerns about the correctness and completeness of social media data sources. Further, the exploitation of social media data related to wildlife involves challenges such as its heterogeneity, noisiness and lack of adequate labelled data. Previous research on using social media sites in ecology studies is limited and often involves manual or semi-automated approaches with few attempts to exploit advanced machine learning methods.

In this thesis, we aim to identify social media mining techniques that facilitate the usage of social media datasets as a source of wildlife observational data. First, we study the potential of social media data to supplement citizen science data collections and perform a range of statistical, spatial, and temporal analyses. We also present image and text-classification based verification approaches for identifying wildlife observations on social media which are suitable for large and diverse data collections. To address the fact the only a small proportion of social media posts have coordinates, we develop geo-referencing techniques that use state-of-the-art transformer-based neural network models, transfer learning, and regression models. These methods are extended with hybrid approaches incorporating machine learning and rule-based methods to improve the precision of geo-referencing models given limited amounts of training data. A preliminary study of how social media can be exploited for spatio-temporal analysis is conducted. The thesis shows that the image sharing platform, Flickr and the micro-blogging service Twitter can be valuable sources of wildlife observational data but require verification and preparation techniques to support their use. We show that combining neural network models, transfer learning, and/or rule-based approaches can facilitate the verification and georeferencing of social media datasets even in the presence of more specialised language and limited amounts of labelled data. We also present the largest collections of geo-referenced wildlife-related Twitter and Flickr datasets as well as a deep learning transformer model trained on wildlife Tweets. These resources can be beneficial for further studies into passive citizen science and social media mining.

Acknowledgements

With the completion of this study I would like to give thanks to Prof. Chris Jones and Dr Padraig Corcoran for their supervision throughout this study has gone beyond anything I could have ever asked for. I have gained knowledge and experience from working with them that has had a huge impact on not only this study but my professional life as a whole.

I also need to thank the School of Computer Science and Informatics at Cardiff University for the opportunity to conduct this study. Through many they have acted as a home away for over a decade now and I hope our journey together doesn't end here.

Finally, I would like to give a special thanks to my family, in particular I would like to give thanks to my wife Aleksandra, my daughter Elizabeth, my mother-in-law Jeni, and my grandparents-in-law Penka and Nikolaj. Their support and guiance both in good times and bad have helped guide this study to completion and will never be forgotten.

Contents

List of Publications

The work introduced in this thesis is based on the following publications.

- Edwards et al. [2019] — Thomas Edwards, Christopher B. Jones, and Padraig Corcoran. *Extracting Geometric Representations Of Trajectories Using Topological Data Analysis.* In *GeoComputation*, 2019. *The University of Auckland.* — This paper is the basis for the work presented in Chapter 6. Contributions include: methodology, analysis, paper draft.

- Edwards et al. [2021] — Thomas Edwards, Christopher B. Jones, Sarah E. Perkins, and Padraig Corcoran. *Passive citizen science: The role of social media in wildlife observations.* In *Plos one*, volume 16, 2021. *Public Library of Science San Francisco, CA USA.* — This paper is the basis for the work presented in Chapter 3. Contributions include: methodology, analysis, paper draft.

- Edwards et al. [2022a] — Thomas Edwards, Christopher B. Jones, and Padraig Corcoran. *Identifying wildlife observations on twitter.* In *Ecological Informatics*, 2022. — This paper is the basis for the work presented in Chapter 4. Contributions include: methodology, analysis, paper draft.

- Edwards et al. [2022b] — Thomas Edwards, Christopher B. Jones, and Padraig Corcoran. *A Hybrid Approach for Geo-referencing Tweets: Language Model Regression and Gazetteer Disambiguation.* In *International Journal of Geographical Information Science (IJGIS)*, 2022 [under review]. — This paper is the basis for the work presented in Chapter 5. Contributions include: methodology, analysis, paper draft.

List of Figures

List of Tables

List of Acronyms

BERT Bidirectional Encoder Representations from Transformers

BOW Bag-of-Words

CBOW Continuous Bag of Words Model

CNN Convolutional Neural Network

GMM Gaussian Mixture Model

GBIF Global Biodiversity Information Facility

GNB Gaussian Naive Bayes

GPS Global Positioning System

LG Logistic Regression

LBSNs Location-Based Social Networks

LSTM Long-Short Term Memory Neural Network

MLM Masked Language Model

NBN National Biodiversity Network

NER Named Entity Recognition

NLTK Natural Language Toolkit

NLP Natural Language Processing

OOV Out-Of-Vocabulary

OSM Open Street Map

POI Points of Interest

POS tagging Part-of-Speech Tagging

RNN Recursive Neural Network

RoBERTa Robustly Optimized BERT Pretraining Approach

SVM Support Vector Machines

SVR Support Vector Regression

SNS Social Network Site

uSIF unsupervised Smooth Inverse Frequency

Glossary

Neural Network: Neural networks are machine learning algorithms inspired by the structure of the human brain. Neural networks are comprised of a node layers, containing an input layer, one or more hidden layers, and an output layer. Each node, or artificial neuron, connects to another and has an associated weight and threshold. If the output of any individual node is above the specified threshold value, that node is activated, sending data to the next layer of the network[1].

Early Neural Networks: These neural networks are based on feed-forward approaches where text is processed in a sequential manner, word by word. Examples of such neural networks are Recurrent Neural Network (RNN) and Long Short Term Memory (LSTM) Neural Network. These sequential neural network architectures can fail at providing more context-specific word representations and tend to be computationally expensive.

Recurrent Neural Network (RNN): RNNs are feed-forward NN, which process text in a sequential manner where sentences are processed word by word. RNN process sequential information by recurrence. Previous input is represented as the hidden state of the recurrent computation and each new input is processed and combined with the

[1] Resource on Neural Networks: `https://www.ibm.com/cloud/learn/neural-networks`

hidden state. A limitation of RNN is that they process text from left-to-right or right-to-left and have limited capacity to remember long term dependencies words.

Long Short Term Memory (LSTM) Neural Network: Long-short term memory neural models (LSTMs) are an extension to RNNs and they address the problem of RNN (learning only short-term dependencies) by using a gating unit which allows it to selectively determine what to remember over long spans reducing the number of successive gradient calculations. Despite, this improvement, these neural models can still fail at providing more context-specific representations and tend to be computationally expensive.

Transformer-based Neural Network: Transformer type neural network architecture addresses the problems associated with earlier neural network models by using an attention mechanism where each word representation is directly connected with the representation of every other word. The non-sequential manner in which data is processed enables capturing more relationships between words and thus provides better contextual representation.

Skip-gram approach: An approach for building word embedding models where during training it tries to predict the source context words (surrounding words) given a target word (the center word).

CBOW approach: An approach for building word embedding models where during training it predicts the target word according to its context words.

Pre-trained Model: Neural network architectures allow model pre-training where word or language representation models can be trained on large generic corpora with the possibility of subsequently being adapted to specific tasks using an application-specific training dataset to fine-tune the model.

Corpus-trained Model: Neural network language representations learned from scratch using the application training set (task-specific dataset). Note though that all pre-trained models have been trained on generic corpora.

Fine-tuning technique: This a technique, mainly used in transformer-based architectures where a pre-trained word model is adapted (fine-tuned) to the classification task by adding a single additional neuron layer which is task-specific and requires labelled training data.

Word Embedding Model: Multi-dimensional vector space representations of words generated using dimensionality reduction methods that represent the semantics of words and capture semantic relationships between words. Word embeddings can be created using principles from the neural network architectures. Some of the most efficient techniques used to generate word embedding models are skip-gram and CBOW. A problem with standard word embedding models is that they produce a single vector representation per word independent of the context in which they appear.

Language Model: These are word representations also referred to as contextualised word embeddings built using transformer-based principles. They address the limitations associated with conventional word embeddings by computing dynamic representations for words based on the context in which they are used.

Bidirectional Encoder Representations from Transformers (BERT): A state-of-the-art language model. It is available as a pre-trained model for various domains. However, one of the biggest and most widely used pre-trained BERT models is trained on Books corpus and Wikipedia data. This pre-trained model can be fine-tuned for various tasks by adding a single output layer.

GloVe: A count-based word embedding model where dimensionality reduction uses a co-occurrence counts matrix. For this paper, we used GloVe model pre-trained on a large corpus of generic Tweets.

Word2Vec: A word embedding model which uses the skip-gram approach to build term representations. it is a two-layer neural network which gives as an output an embedding matrix, where each term (single or multi-token) from the corpus vocabulary is represented as an n-dimensional vector. A problem with the Word2Vec model is that it ignores the morphology of words by assigning a distinct vector to each word. For the paper, we used a pre-trained Word2Vec model trained on Google news datasets.

fastText: A word embedding model which generates vector representations of each character n-gram and words are represented as the sum of these representations. This allows the creation of representations of rare and misspelled words.

fastText classification pipeline: A one layer neural network which has been developed to deal with unbalanced large datasets with fast training time. The classification pipeline learns embeddings for each word in a sentence. These word representations are then averaged to create a sentence representation, which is fed into the classifier layer.

Citizen science: The scientific work undertaken by members of the public, often in collaboration with or under the direction of professional scientists and scientific institutions to collect biodiversity data.

Passive citizen science: The use of social media that are unconnected to any particular citizen science program, but represent an unexploited source of valuable ecological data.

Twitter: Twitter is a micro-blogging social network which was established in 2006. The platform has 302 million active users that send over 500 million tweets every day. Twitter architecture is based on posting short messages, i.e. 'tweets' and user connections are established based on 'following' principles.

Tweet: A tweet is a piece of user-generated text with its length up to 280 characters. It may describe anything a user wants to post, e.g. mood, events, observations. In addition to posting original content, users can also retweet other's tweets. Tweets and retweets from a user will be pushed to their followers Twitter interface for them to read. A Tweet may include hashtags (words or phrases starting with '#' and mentions of another user's name identified with a preceding '@')s.

Followers: Besides posting tweets, a user may subscribe to others' tweets by following them. These relationships are unidirectional where a user can follow another user without the opposite being true.

Twitter Date Information: A Tweet is associated with its posting timestamp so the date of which a Tweet is posted is given automatically by Twitter architecture.

Twitter Location Information: Users may optionally publish their location information. Further, users may complete their profiles to include information like home cities, timezones, and personal websites. Timestamps, geo-tags, and user profiles serve as contextual information for tweets, and we refer to them as tweet context. The optional sharing of location data may cause incompleteness in the data collections (i.e., lack of coordinates) which is a critical issue for research on wildlife observations.

Flickr: Flickr, established in 2004, is one of the largest photo-sharing social network platforms with more than 100 million registered users, and 10 billion photographs uploaded. It is used by professional photographers and amateurs to upload and organise photos within collections on different topics.

Flickr Date Information: A Flickr post is associated with two dates, ‘taken date’ and ‘posted date’. ‘Taken date’ refers to the time at which the photo was taken while ‘posted date’ represents the time at which the photo was uploaded to Flickr. For the purposes of the thesis, we focus mainly on the ’taken date’.

Flickr Location Information: Providing location information per Flickr post is optional where the user has to choose whether they want to share the location of the photo or not. Once the user has agreed to share the location, photos are automatically geotagged using GPS coordinates of the device used to take the photo. A similar problem associated with the lack of location information as for Twitter exists for Flickr, however in a smaller degree.

Chapter 1

Introduction

Observations on the distribution of wildlife species have always formed a crucial part of conservation and species management [Amano et al., 2016, Barve, 2014]. Wildlife-related data is becoming increasingly important in the face of rapid ecosystem changes that can be brought about, for example, by climate change and invasive species. The consequences of such changes have implications for disease emergence and spread, as well as food security [Barve, 2014].

High-quality species distribution data are typically collected by professionals, but such data can be time-consuming, and expensive to gather, and hence often lack broad coverage [Amano et al., 2016]. To overcome this knowledge gap, especially over a large spatial and/or temporal scale citizen scientists are often engaged; members of the public who volunteer to record the presence of a given species and associated metadata, such as time, date, and location [Silvertown, 2009, Barve, 2014, Cohn, 2008]. Thus, citizen science can be defined as the scientific work undertaken by members of the public, often in collaboration with or under the direction of professional scientists and scientific institutions to collect biodiversity data [Cohn, 2008, Brown and Williams, 2019]. Citizen science projects can effectively crowd source data [Cohn, 2008]. Due to the fact of using non-professionals, however, projects frequently come under criticism in terms of the accuracy of species identification, and associated data [Guerrini et al., 2018]. Further, organising citizen science campaigns and recruitment of volunteers can be cost-consuming and challenging process [Adler et al., 2020]. More recent citizen science projects have tried to address the problem of organising campaigns by using internet-based platforms. An example of such platform is iNaturalist (https://www.inaturalist.org/), a web-based and mobile-supported social network which allows individuals to upload photo observations and identify organisms [Aristeidou et al., 2021]. However, the problems of attracting volunteers to participate, the resource-consuming process of organising campaigns, and the lack of broad coverage of the collected datasets still remain. Social media websites such as Flickr, Twitter, and Facebook have built a network of more than 2 billion users worldwide, generating millions of messages daily that are easily accessible, and reflect the observed reality of a quarter of the human population [Daume, 2016]. Therefore, they emerged as an informal real-time information source that can contribute to the detection of trends and early warnings in critical fields such as ecological change, environmental problems, and shifts in ecosystems [Daume, 2016, Di Minin et al., 2015, August et al., 2020]. We define the use of social media that are unconnected to any particular citizen science program, but represent unexploited source of valuable ecological data as *passive citizen science*. In contrast to citizen science campaigns, the passive citizen science approach provides a cost and time-efficient method for collecting wildlife-related data on a larger scale and for wider time-span. Similar to the citizen science approach, it involves the participation of non-experts. However, the passive citizen science approach consists of crowdsourcing wildlife-related datasets uploaded by the public, independent of campaigns.

A quantitative review of the application of social media in environmental research, conducted by Ghermandi and Sinclair [2019] suggests a very rapid growth in the field of environmental monitoring, with Twitter and Flickr being most frequently used as data sources. Among the identified strengths of social media is the large volume of available data samples which makes data collection a less labour-intensive, time-consuming and costly procedure [Ghermandi and Sinclair, 2019, Antoniou et al., 2016, Soliman et al., 2017]. Social media data also allows for a timely and (near) real-time monitoring and analysis of species distribution [Ghermandi and Sinclair, 2019, Daume et al., 2014, ElQadi et al., 2017, Jarić et al., 2020].

1.1 The Problem

Despite the potential of social media to be used for species distribution models there are still some concerns about the quality, reliability, and completeness of information mined from social media [Ghermandi and Sinclair, 2019, Daume, 2016, Kent and Capello Jr, 2013]. There are also concerns about the data ownership and future availability of social network data [Daume, 2016, Palomino et al., 2016, Ghermandi and Sinclair, 2019]. Further, datasets related to wildlife observations need to be associated with location information to allow for species tracking and observation of movement patterns. However, often users refuse to share their location on social media sites which leads to large quantities of potentially valuable wildlife-related social media data that lack coordinates information.

Recent research on using social media data as a source for ecology related studies has focused on addressing some of these problems by proposing verification approaches and estimating the value of social media platforms to supplement official citizen science portals [Daume, 2016, ElQadi et al., 2017, Barve, 2014, Ghermandi and Sinclair, 2019]. Most of the proposed approaches are limited in scale (suitable for verifying data associated with a few species) and involve manual or semi-automatic verification. More recent research [Jarić et al., 2020, August et al., 2020, Skreta et al., 2020] investigates automated image verification methods suitable for verifying larger collections, specifically plants or butterflies. However, the aforementioned research is still limited in scale and there is lack of verification techniques suitable for textual data.

Text classification approaches are suitable for identifying wildlife-related text-based social media posts. Machine learning methods [Al-Garadi et al., 2021, Guo et al., 2020, Liu et al., 2021, Lopez-Lopez et al., 2021] are widely adopted in social media mining where most of the work is based on using neural network models. However, most of the existing solutions are suitable for big data analysis and lack extensive comparison between different classification strategies and their suitability for verifying wildlife-related data. Additionally, work on text classification for wildlife data is very limited and it is based on using statistical machine learning algorithms [Jeawak et al., 2017, 2020, Leung and Newsam, 2012]. Similarly, research on geo-referencing approaches [Scherrer et al., 2021, Eisenstein et al., 2010b, Priedhorsky et al., 2014, Rahimi et al., 2017a, De Rouck et al., 2011, Laere et al., 2014a] lack comprehensive investigation into methods suitable for smaller training datasets. Finally, existing geo-referencing approaches do not fully take an advantage of recently created transformer-based neural network models and transfer learning techniques which give state-of-the-art performance for various Natural Language Processing (NLP) tasks.

The significant need for establishing methodologies for verifying and preparing social media data to serve as a useful supplement to official citizen science campaigns is the main motivation for this thesis. This involves the execution of number of steps which have not been fully researched and can be challenging when dealing with social media datasets which tend to be noisy and heterogeneous but also may include more specialised language when search is limited to wildlife observations and also lack large labelled datasets. In particular, these steps involve validation and geo-referencing. Additionally, there is need for establishing methods for analysing the wildlife-related social media data such as with respect to spatio-temporal patterns. We build towards establishing such methods by extending on spatio-temporal techniques for object location identification to support extraction of object's trajectory data.

1.2 Research Questions

This thesis is motivated by the hypothesis that social media provide the potential to supplement active citizen science efforts to acquire observations of wildlife, and that computing methods can be developed to assist in recognising and geo-referencing such observations.

The validity of this hypothesis is tested through the conduct of large scale analysis evaluating the value of social media datasets to supplement citizen science data collections and identification of fully automated verification methods suitable for performing large scale validation of image and text social media data related to wildlife. The thesis also presents approaches for geo-referencing social media posts related to wildlife focusing on scenarios with small quantities of training data. Previous work on analysing the potential of social media data to support wildlife-related studies and verification techniques (discussed in Sections 2.2.3, 2.2.5 of the Background chapter) are scarce, limited in scale, or present only semi-automatic approaches. Further, recent research on geo-referencing social media posts and text classification methods (discussed in Sections 2.3 and 2.4) usually rely on large amounts of training data and lack extensive analysis on how state-of-the-art neural network models and transfer learning techniques can be incorporated in creating more accurate text verification and geo-referencing methodologies suitable for smaller training datasets and wildlife-related social media data. Finally, we look at extending on existing spatio-temporal based methods, presented in Section 2.5, to support trajectory extraction for objects and thus facilitate studies on movement patterns and tracking of wildlife and weather data objects. The central point addressed by this research is that incorporating transfer learning techniques, state-of-the-art neural network models, and less data consuming rule-based approaches for geo-referencing can facilitate the creation of validation and preparation techniques for social media data related to wildlife. In this work, the following research questions help illustrate the steps towards realising this thesis:

- **RQ 1:** Can social media data serve as a useful supplement to citizen science data portals in representing the spatial and temporal distribution of bio-diversity data?
- **RQ 2:** What are the most efficient text classification approaches for verifying that social media postings are genuine wildlife observations?
- **RQ 3:** Can deep learning transformer regression models provide an effective means of geo-referencing social media posts?
- **RQ 4:** Do zig zag persistent homology methods have good potential for extracting trajectories of spatio-temporal objects?

1.3 Solution Framework

In Figure 1.1 we have outlined the framework which we will follow in Chapters 3, 4, 5, and 6 in order to answer the research questions presented in Section 1.2. In the 'Verification' stage, we looked to verify social media postings as true wildlife observations. The verified datasets which lack coordinates are passed to the geo-referencing module for assigning coordinates and finally the trajectory extraction stage can use the georeferenced data to extract object trajectories. This will facilitate analysis into species movement patterns.

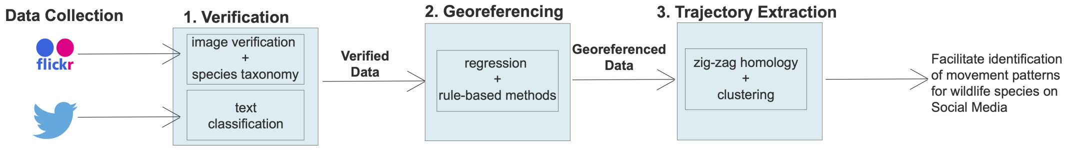

Data Collection
flickr
1. Verification
image verification + species taxonomy
text classification
Verified Data
2. Georeferencing
regression + rule-based methods
Georeferenced Data
3. Trajectory Extraction
zig-zag homology + clustering
Facilitate identification of movement patterns for wildlife species on Social Media

Figure 1.1: Solution Framework

1.4 Contributions

The main contributions made in this research work are outlined below.

- **Contribution 1:** We conducted a large scale study (including the largest number of species considered to date) investigating the potential of social media data to supplement official citizen science data portals. Specifically, a comparison between image-sharing social media platform and citizen science data collection has been performed using statistical, spatial, and temporal analysis considering different spatial and temporal settings. The analysis revealed that image-based social media platforms could offer a rich source of observation data for certain taxonomic groups, and/or as a resource for dedicated projects. In particular, spatial and temporal analysis suggest that the social media dataset best reflects the citizen data collection when considering a purely spatial distribution with no time constraints. Further, we develop a fully automated verification method for image-based social media platforms suitable for verifying large and diverse collections of species. The approach is based on the the use of an image recognition tool in combination with species taxonomic data to determine the likelihood that the mention of a species on social media platform represents a given species. The work relevant to this contribution is presented in **Chapter 3**.

- **Contribution 2:** A comparison between three different classification algorithms, and various feature extraction and feature integration methods allowed us to identify techniques suitable for verifying that postings in text-based social media data collections are relevant to wildlife observations, using limited amounts of training data. This analysis revealed the potential of state-of-the-art large pre-trained neural network models that are fine-tuned to the classification task to correctly classify instances relevant to wildlife even when more specialised language is used. Further, an investigation into the specific components of the social media posts that are indicative for genuine wildlife observations on social media revealed trends concerning the use of hashtags that are unrelated to official citizen science campaigns. Such hashtags could therefore be exploited in automated feature selection techniques for improving classification performance, as well as used as part of more informal campaigns encouraging people to use these hashtags when wildlife observations are posted. The work relevant to this contribution is presented in **Chapter 4**.

- **Contribution 3:** We conducted analysis into less data-consuming geo-referencing approaches based on regression, transformer-based models and transfer learning techniques. Findings showed that using a domain-trained state-of-the-art language model that is adapted for multivariate regression and augmenting the training set with datasets from multiple social media platforms can be beneficial for geo-referencing social media posts. Evaluation has been performed using wildlife-related social media posts and two regression models — one based on a widely used statistical regression approach and the other using neural network-based regression. Further, we provided the largest collection of geo-referenced wildlife-related Tweets and a domain-trained transformer model which can be used in future research on geo-referencing and analysing social media data relevant to wildlife. Finally, we proposed two hybrid approaches incorporating multivariate regression models based on transformer architecture and rule-based strategies, i.e., location name extraction and semantic similarities between the training and test instances. Both strategies and especially location name extraction combined with regression, led to improvements in the precision of geo-referencing models without requiring large amounts of training data. The work relevant to this contribution is presented in **Chapter 5**.

- **Contribution 4:** A methodology for extracting and normalising geometric representations of trajectories for tracking spatio-temporal phenomena has been developed. The methodology used the spatio-temporal objects resulting from topological data analysis based on zig-zag persistent homology. Further, clustering and normalising the trajectories helped identify similar trajectories and similar patterns of movement for weather-related dataset. The study also indicates the potential for these methods to be applied to social media wildlife data. The work relevant to this contribution is presented in **Chapter 6**.

1.5 Thesis Outline

The chapters containing the remainder of this thesis are laid out as follows.

- **Chapter 2: Background and Research Domain** — This chapter introduces main concepts in citizen science and passive citizen science. It discusses related work on social media mining techniques, including text classification, geo-referencing, and trajectory extraction methods. It further identifies gaps in literature related to verifying and geo-referencing sparse collections of social media content related to wildlife, as well as gaps related to trajectories extraction approaches.

- **Chapter 3: Suitability of Social Media as a Supplement to Citizen Science Portals** — Chapter 3 presents a large scale study, including a range of statistical, spatial and temporal analysis for identifying whether photo-sharing media platforms can serve as a useful source of wildlife data that can complement citizen science data collections. Further, a fully automated verification method has been presented, suitable for verifying large and diverse collections of image-based social media datasets. The chapter shows that verified and geo-referenced image-based social media data can be used to observe certain types of species and taxonomic groups. The work in this chapter relates to **Contribution 1**.

- **Chapter 4: Text Classification for Verifying Social Media Relevant to Wildlife** — This chapter analyses the suitability of state-of-the-art classification approaches for verifying text-based social media content related to wildlife with the presence of limited training data. Chapter 4 shows the potential of state-of-the-art neural network techniques to facilitate the discovery of valuable wildlife related data on social networks without the need of human verification steps or officially organised citizen science campaigns. The work in this chapter relates to **Contribution 2**.

- **Chapter 5: Geo-referencing Social Media Data Related to Wildlife Observations** — This chapter investigates the benefits of using state-of-the-art neural network models and transfer learning techniques for building regression models for geo-referencing social media posts. The research focuses on scenarios with limited training data and the geo-referencing of wildlife-related posts. Chapter 5 shows that enriching small training sets with additional labelled data from diverse social media networks can be beneficial for the performance of regression models, especially when combined with domain-trained contextual word representations. Further, we present analysis into two hybrid approaches for improving the precision of geo-referencing social media posts that do not require additional training data. The work in this chapter relates to **Contribution 3**.

- **Chapter 6: Extracting Geometric Representations Of Trajectories** — This chapter presents a method for trajectory extraction based on using objects locations extracted from imagery data using temporal analysis approaches. The chapter also shows that normalisation techniques such as clustering help identify object movement directions. The work in this chapter relates to **Contribution 4**.

- **Chapter 7: Conclusions and Future Work** — This chapter concludes this thesis and summarises our contributions and findings. It also highlights work that could be undertaken to take this project further and covers future plans for building on approaches for facilitating the use of social media data in wildlife studies.

Chapter 2

Background and Research Domain

As discussed in Chapter 1, wildlife data is often used for studying changes in species movement patterns and invasive species occurrences. Such information is important for detecting early environmental and climate changes as well as supporting species preservation campaigns Amano et al. [2016], Barve [2014]. Traditional methods for collecting environmental data involve professionals or more often volunteers (i.e., citizen scientists) who take part in official campaigns to observe given species. These campaigns or the involvement of professionals to collect wildlife-related data are often costly and time-consuming to organise and execute [Amano et al., 2016, Silvertown, 2009, Doyle et al., 2019]. Due to its wide usage, social media platforms such as Flickr and Twitter, represent an unexploited source of potentially valuable wildlife data which is less costly to obtain but yet can complement citizen science data collections and provide a real-time species monitoring [Daume, 2016]. However, concerns about the correctness and completeness of social media data still remain unresolved. Previous research is limited in scale and often involves manual processing. Further, preparing social media data to serve as a supplement to citizen science data collections involve the execution of a number of steps which have not been fully researched and can be challenging when dealing with wildlife-related collections which lack large amounts of labelled data. In particular, preparation of data involves validation, geo-referencing, and extraction of spatio-temporal movement data. Thus, the problem of applying social media mining techniques to validate social media data relevant to wildlife observations is the basis of the research in this thesis. This chapter provides a review of main techniques and concepts in social media mining including a survey of some of the most relevant works of the area. Further, we point to gaps in the literature in relation to our problem focus set out in Chapter 1, i.e., verifying, geo-referencing, and analysing social media data relevant to wildlife.

2.1 Citizen Science

2.1.1 Definition and Applications

Wildlife observation is the practice of noting the occurrence or abundance of a dead or living animal species at a specific location and time [Davis and Winstead, 1980]. High-quality species distribution data are typically collected by professionals, but such data can be time-consuming, and expensive to gather, and hence often lack broad coverage [Amano et al., 2016, Silvertown, 2009, Doyle et al., 2019]. To overcome this knowledge gap, especially over a large spatial and temporal scale many wildlife-related studies involve the participation of members of the public who volunteer to record the presence of a given species and associated metadata, such as time, date, and location [Silvertown, 2009, Barve, 2014, Cohn, 2008]. These volunteers are referred to as citizen scientists. In this context (recalling from Chapter 1), citizen science, also called participatory science and crowd-sourced science, can be defined as ‘the scientific work undertaken by members of the public, often in collaboration with or under the direction of professional scientists and scientific institutions to collect biodiversity data’ [Cohn, 2008, Brown and Williams, 2019, Doyle et al., 2019]. Engaging non-professionals in wildlife observation campaigns allows for the collection of biodiversity data on a larger scale at a lower cost, compared to traditional wildlife-observation approaches involving only professionals. Therefore, citizen science projects have been remarkably successful in advancing research in bio-geography, ecology, invasive species biology, climate change, and land cover use [Bonney et al., 2009, Barve, 2014, Amano et al., 2016, Laso Bayas et al., 2021]. The data collected through citizen science campaigns is usually available through digital biodiversity data portals Barve [2014]. Some of the most successful UK-based citizen science campaigns have been organised by the British Trust for Ornithology's (BTO) [1] and Royal Society for Protection of Birds (RSPB) [2] [Hart et al., 2012]. In particular, garden-based citizen science programs have been very successful in collecting biodiversity data, particularly on avian species. Examples of such citizen science campaigns are the 'Garden BirdWatch' and 'Big Garden Weigh-In' organised by the BTO and 'Make Your Nature Count survey' organised by the RSPB.

2.1.2 Citizen Science Data Portals

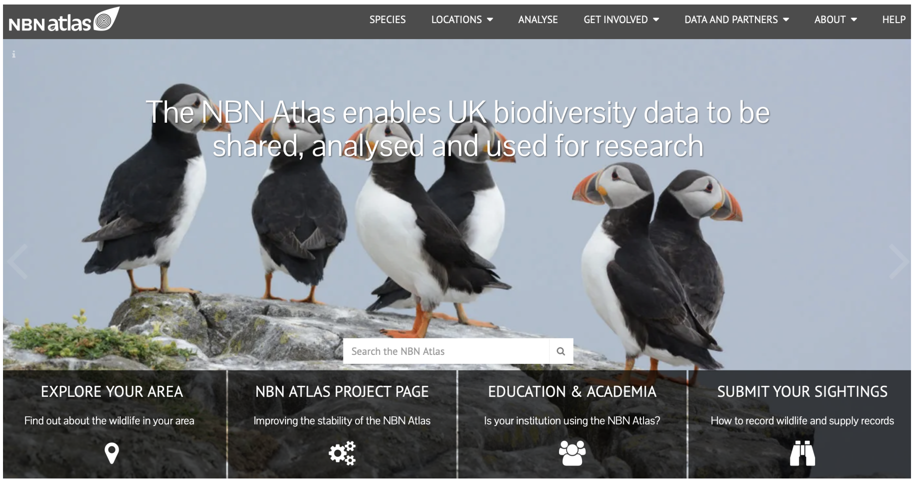
The emergence of Internet technologies stimulated a process of integrating the data, collected by citizen science campaigns into digital data portals which provide easy access to diverse data collections, facilitate the conduct of ecology and wildlife-related analysis, and further help for the organisation of citizen science campaigns [Yesson et al., 2007, Heberling et al., 2021]. Some of the most diverse and large citizen science data portals are explained in the rest of this section.

Global Biodiversity Information Facility (GBIF) Global Biodiversity Information Facility (GBIF) is the world's largest biodiversity data network [Heberling et al., 2021]. It provides the largest single gateway to wildlife observational data collected by citizen scientists [Yesson et al., 2007]. It has been created with the aim to make the world's biodiversity data freely and universally available via the Internet and enable scientific research Bridgewater et al. [2010]. It has been funded by world's governments and aimed at providing anyone, anywhere, open access to data about all types of life on Earth.

[1] BTO: `http://www.bto.org/`

[2] RSPB: `http://www.rspb.org.uk/`

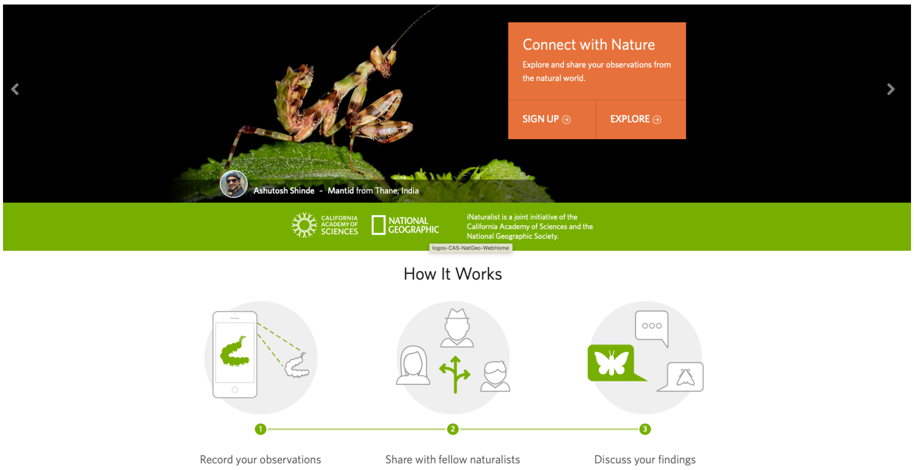
National Biodiversity Network (NBN) National Biodiversity Network (NBN) is the UK node of GBIF (see Figure 2.1). The NBN is a collaborative project committed to making biodiversity information available via the NBN Atlas. The National Biodiversity Network (NBN) is registered as a charity and supports the sharing of ecological data in the UK since 2000. The goal of the project is to improve the availability of high-quality species occurrence data in the UK. It is the largest collection of biodiversity information within the UK and Ireland and has revolutionised the use of biodiversity data by allowing it to be shared, downloaded, analysed, and researched by the public. NBN Atlas holds more than 220 million species occurrence records combined from individual observations and official organisations such as the 'Royal Society for the Protection of Birds (RSPB)'. NBN datasets have proved beneficial in studying wildlife distribution in a number of studies [Leivesley et al., 2021, Blight et al., 2009]. It also provides a programmatic API which facilitates fast and efficient data acquisition and processing.



The NBN Atlas enables UK biodiversity data to be shared, analysed and used for research

Figure 2.1: NBN interface

iNaturalist iNaturalist is a web-based and mobile-supported social network which allows individuals to upload photo observations and identify organisms [Aristeidou et al., 2021] (see Figure 2.2). It can be used to record species observations, share findings with others, get help with identifications, or access the observational data collected by iNaturalist users. Therefore, it is a crowd-sourced species identification system and an organism occurrence recording tool. Further, it aims to generate scientifically valuable biodiversity data from volunteer's observations.



Connect with Nature

Explore and share your observations from the natural world.

Ashutosh Shinde - Mantid from Thane, India

iNaturalist is a joint initiative of the California Academy of Sciences and the National Geographic Society.

How It Works

Figure 2.2: iNaturalist interface

2.1.3 Issues with Citizen Science Campaigns

Despite the large efforts in collecting, sharing, and providing easy accessibility to citizen science data, projects frequently come under criticism in terms of the accuracy of species identification, and associated data due to the fact of using non-professionals [Guerrini et al., 2018]. Further, organising citizen science campaigns and recruitment of volunteers can be cost-consuming and challenging process especially when observational data need to be collected over long periods of time [Adler et al., 2020]. Therefore, studies are often conducted on a limited spatial and temporal scale and focus on a limited number of species. These lead to gaps in the biodiversity data collections, both geographically and taxonomically [ElQadi et al., 2017, Barve, 2014]. More recent citizen science projects have tried to address the problem of organising campaigns by creating Internet-based platforms (described in Section 2.1.2) or using existing social networks for attracting and gathering volunteers. For instance, the citizen science platform iNaturalist has been successfully organising campaigns for observing wildlife [Aristeidou et al., 2021]. The authors of [Aplin et al., 2021] uses a mobile application for collecting observational data about parrots and studying their social organisation. Another example includes urban residents reporting occurrences of tagged birds through a Facebook group, a smartphone application and email [Davis et al., 2017]. A crowdsourcing tool was employed in [Fritz et al., 2012] to collect data for the creation of a land cover map, while in [Paul et al., 2018, Lowry and Fienen, 2013] crowdsourcing was used as a supplemental method for collecting hydrologic data. However, the problems of attracting volunteers to participate and the cost of organising campaigns still remain.

2.1.4 Summary

Citizen science, consisting of organising campaigns for non-professional volunteers has become a standard approach for collecting wildlife observational data which can be used to facilitate ecology-related studies. Despite the efforts of integrating and making citizen science-related data easily accessible through data portals, the problems associated with the correctness, completeness, and diversity of the wildlife data collected still remain. Additionally, organising specific campaigns for gathering volunteers can be a resource-consuming process. The emergence of social media sites which provide a large sharing platform for information has led to an increasing interest in the suitability of this data for studying wildlife. The need to further explore how social media mining techniques can be used to validate and prepare such sources for supporting wildlife studies is the main motivation for the research in the thesis and thus we focus on this problem in the next sections.

2.2 Social Media

2.2.1 Definition, Usage, Types

The rapid emergence of Web 2.0 and the easy public access to Web 2.0-based technologies marked a new era for Internet use allowing users not only to view content but also to communicate, share and exchange information through the use of social media platforms. In this context social networking sites (SNSs) can be defined as 'web-based services that allow individuals to (1) construct a public or semi-public profile within a bounded system, (2) articulate a list of other users with whom they share a connection, and (3) view and traverse their list of connections and those made by others within the system' [Boyd and Ellison, 2010]. The wide accessibility and real-time manner in which information is exchanged on SNSs led to an increasing use of these platforms and the generation of large volumes of content, including text, photo, audio, and video data [Camacho et al., 2020, Magge et al., 2021, Daume, 2016, Di Minin et al., 2015, August et al., 2020, Gundecha and Liu, 2012]. Therefore, SNS have become not only a valuable communication platform for people but also a valuable information source that can contribute to various applications such as trend and event detection in the economy, public opinion and population health [Magge et al., 2021, Burnap et al., 2016, Preis et al., 2013, Ortiz et al., 2011, Mellon, 2014, Ahani and Nilashi, 2020], fake news and propaganda identification [Ozbay and Alatas, 2020, Shu et al., 2017, Wu and Liu, 2018], emergency and disaster management [Imran et al., 2020, Luna and Pennock, 2018, Daume, 2016, Martin et al., 2020], marketing research [von Scheel et al., 2015], and many more.

Based on the definition for SNSs by Boyd and Ellison [2010] and the type of data exchanged on these platforms, we distinguish the following types of SNSs — social communication networks, photo-sharing network sites, micro-blogging network sites, location-based social networks (LBSNs).

Social Communication Networks: The purpose of these Internet platforms is to simulate the establishment of friendships and social connections from real life [Zheng, 2011]. These SNSs allow users to share with one another hybrid type of content, including text, photos, and videos [Pittman and Reich, 2016]. Further, they require a mutual connection between users based on 'friendships' where users in order to interact have to be part of each other's 'friendship groups'. Facebook, established in 2004, is the biggest social communication network worldwide with 2.85 billion monthly active users as of the first quarter of 2021 [3]. Another example of social communication network is LinkedIn which is mainly used for professional networking, and facilitates connections between job seekers and employers. Some of these social networks allow users to share their location or attach GPS location information to their posts.

Micro-blogging Network Sites This type of SNSs allow users to share short statements (i.e., micro-blogs) with other users of the network where micro-blogs are primarily text-based but can also include links to other types of media (photos, videos, references to web pages). In contrast to Social Communication Networks, micro-blogging platforms do not require mutual 'friendship' connections for sharing and viewing information. Twitter is one of the most popular micro-blogging network sites that lets users share 280-character 'tweets' of text which might link to other sites or photo/video files.

Photo-sharing network sites Photo-sharing platforms allow people to store, organise, share and search photos collections related to a given topic or interest. Photos are usually associated with a title, description, and list of tags describing the content of the photo. Unlike the other types of social networks, the purpose of these SNSs is image sharing rather than text-based message exchange. The most widely used photo-sharing network sites are Flickr, Snapchat, Instagram. Additionally, users can share

[3] `https://s21.q4cdn.com/399680738/files/doc_financials/2021/FB-Earnings-Presentation-Q1-2021.pdf`

GPS location associated with each photo as well as personal location.

Video-sharing Network Sites These platforms allow sharing video content where similarly to the photo-sharing network sites videos are usually associated with a title, description, and list of tags describing the video. TikTok is one of the most recent and popular video-sharing network sites used for creating and sharing short videos. TikTok was created in 2016 and it has 689 million users worldwide [Mohsin, 10]. However, analysing video content is outside the scope of our research.

Location-Based Social Networks (LBSNs) The increasing availability of location-acquisition technologies such as GPS led to the emergence of LBSNs where users can share location-related data such as the venues, places, or Points-of-Interest (POI) they visit. These platforms help users establish connections with other users based on common location interests (nature walks, restaurant preferences), and also check venue properties, such as their opening times, opinions, and pictures [Torrijos et al., 2020] as well as provide recommendations based on previously visited locations. A popular LBSN which allow users to share their current location and explore venue's information is Foursquare.

2.2.2 Passive Citizen Science

As described in Section 2.2.1, the exponential growth of connections and information on social media platforms make them a valuable source of knowledge for many applications. Throughout the thesis, we focus on analysing the potential of social media datasets to be used for wildlife observation studies and supplement citizen science data portals. As mentioned in Chapter 1, we define the use of social media that are unconnected to any particular citizen science program, but represent an unexploited source of valuable ecological data, as *passive citizen science*. In contrast to citizen science, the passive citizen science approach provides a cost and time-efficient method for collecting wildlife-related data on a larger scale and for a wider time-span. Similar to the citizen science approach, it involves the participation of non-experts. However, wildlife-related data can be obtained without organising specific campaigns. Further, there are no restrictions on species observations and time-frames. Instead it consists of a process of crowdsourcing in which data are retrieved from Internet resources, particularly social media, to which members of the public have uploaded observations such as annotated photos of wildlife.

In particular social media data have the following advantages in being used to supplement official citizen science campaigns [Ghermandi and Sinclair, 2019]:

- Large data volumes at no costs — Social networks generate large amounts of user content daily, some of which might be useful for wildlife-related research. The large volume of available data makes data collection a less labour-intensive, time-consuming and costly procedure [Ghermandi and Sinclair, 2019, Antoniou et al., 2016, Soliman et al., 2017].
- Support cross-validation of data collected by citizen science projects [Di Minin et al., 2015]
- Support large scale of analysis — Traditional approaches are limited to exploring only a few species depending on the purpose of the citizen science campaign that has been conducted. Social media data can be easily applicable to scales, such as entire populations, ecosystems or biomes.
- In locations where resources for field work and data collection are scarce, social media can help save resources and allow directing professional data collection to less known or more poorly accessible areas [Di Minin et al., 2015].
- Support real-time monitoring — Social media data allows for a timely and (near) real-time monitoring and analysis of land use changes [Sitthi et al., 2016], species distribution [Jeawak et al., 2020, Ghermandi and Sinclair, 2019, Daume

et al., 2014, ElQadi et al., 2017], early warning to natural hazards [López-Cuevas et al., 2017] and provide early alerts for pending and potentially irreversible shifts in ecosystems [Daume, 2016, Di Minin et al., 2015, August et al., 2020]

Despite these benefits of social media to be used for species distribution models, there are still some concerns about the quality and reliability of information mined from social media [Ghermandi and Sinclair, 2019, Daume, 2016, Kent and Capello Jr, 2013]. There are also concerns about the data ownership and future availability of social network data [Daume, 2016, Palomino et al., 2016, Ghermandi and Sinclair, 2019]. Further, datasets related to wildlife observations need to be associated with location information to allow for species tracking and observation of movement patterns. However, often users refuse to share their location information on social media sites which leads to large, potentially valuable, quantities of wildlife-related social media data that lack coordinates information.

2.2.3 Social Media Mining

Social media mining is the process of representing, analysing, and extracting knowledge patterns from unstructured social media content [Gundecha and Liu, 2012, Zafarani et al., 2014]. It is an interdisciplinary field encompassing techniques from computer science, statistics, social sciences and more.

Some of the main challenges associated with analysing social media data include noisy user-generated content written in informal manner where it is possible to have misspellings, jargon language, and unfinished sentences [Hua et al., 2012]. Further, posts are often uploaded as micro-blogs with a limited length which can include text, images, links. Therefore, social mining techniques need to deal with short sequences with heterogeneous nature [Liu et al., 2016, Imran et al., 2020]. Further, performing Natural Language Processing (NLP) tasks such as Named Entity Recognition (NER) and Part-of-Speech (POS) tagging are more challenging for short sequences and thus require different methods versus when analysing documents [Van Laere et al., 2013]

Social media mining techniques have been extensively used in numerous fields of science such as identifying crime hotspots [Yang et al., 2018a], real time disaster management detection [Ekta et al., 2017], and monitoring public health [Magge et al., 2021]. Much of the research in social media mining focuses on performing big data-related analysis and using data-consuming machine learning approaches [Purohit and Peterson, 2020, Usero et al., 2022, Manoharan and Senthilkumar, 2020]. Examples include classification approaches for disaster [Purohit and Peterson, 2020] and business management [Usero et al., 2022], and for support for drug development [Manoharan and Senthilkumar, 2020]. Further, machine learning techniques have been extensively used for categorising and georeferencing social media data [Çöltekin and Rama, 2018, Mohammad et al., 2018, Jeawak et al., 2017, 2018, Jauhiainen et al., 2019, Martinc and Pollak, 2019, Jeawak et al., 2020].

There has also been a rapid growth of the use of social media mining techniques in the field of environmental monitoring, with Twitter and Flickr being most frequently used as data sources [Ghermandi and Sinclair, 2019]. An overview of the impact of internet social networks on traditional biodiversity data collection methods by Di Minin et al. [2015] is optimistic and concludes that social media can potentially play an important role in conservation science. A couple of studies used crowdsourcing tools for the creation of a land cover map [Fritz et al., 2012] and as a supplement to citizen science campaigns for collecting hydrologic data [Lowry and Fienen, 2013]. Other work by Daume [2016], ElQadi et al. [2017], Barve [2014] focused on evaluating social network sites (Flickr and Twitter) relative to biodiversity data portals in order to identify the potential use of ad-hoc methods for augmenting traditional citizen science data collections. However, this previous research was conducted on a narrow range of species (between two and four) and verification of social network datasets was performed semi-automatically or manually. For instance, the authors of Barve [2014] collected geo-referenced Flickr image data related to Snowy Owl (Bubo scandiacus) and the Monarch Butterfly (Danaus plexippus) in order to compare the Flickr collection with the collection available on the citizen data portal GBIF. They did not perform validation, instead they simply compared the map species distributions between the two collections. The research by Daume [2016] used Twitter for collecting wildlife data and performed verification manually. ElQadi et al. [2017] verify Flickr data using an image content recognition tool. Google Reverse Image Search was used in order to return labels per photo which best describe the content of the given photo. All labels per species are ordered in descending order of frequency. Then, species-relevant tags and species-irrelevant tags were identified among the most frequent ones which indicate that the given photo is a true representation of the given species.

Recent research by Jarić et al. [2020] presents a review of the iEcology approach which encompasses the use of automated tools for discovering patterns in the natural world using data accumulated in digital sources collected for other purposes. The authors highlight the value of social media such as Flickr. The use of iEcology approaches is increasing as it provides low-cost and fast data collection, pattern identification, and visualisation of nature-related data. In particular, the study of August et al. [2020] investigates whether a plant species image classifier can be used to extract relevant plant observations from Flickr, using a general search term of ’flower’. Analyses showed that automated methods have the potential to help identify wildlife-related imagery data on social media, especially when photos were focused on single native species in rural situations or when classification was performed at a genus or class level. It was suggested that future work could usefully focus on searching for individual species including invasive species. Work by [Skreta et al., 2020] presents an image-based classification approach which also considers image metadata information such as latitude, longitude, and date to support a fine-grained distinction between different butterfly species. The aim is to support the automated verification of butterfly images for the citizen science portal eButterfly. Currently, the verification of the images has been conducted manually by entomologists. The research shows that an automated model that, along with the image, incorporates geographic and temporal information, can support the labelling of butterfly species without the need of human experts. A drawback of this approach is the need of large amounts of labelled images to build an accurate classifier. Image sharing social media platforms are also used in [Roos and Longo, 2021] where authors collect fish-related images from three different social networks to obtain fisheries information in order to detect illegal actions and help the development of conservation strategies.

Most of the aforementioned approaches do not offer automated verification of social media data that is suitable for large scale studies, especially when the data that is analysed is text-based. Additionally, there is need for the development of methods which help object's trajectory extraction and facilitate better visualisation of movement patterns rather than simply plotting the coordinates of wildlife-observations obtained from social media posts.

2.2.4 Twitter and Flickr

As identified in Section 2.2.3, Twitter and Flickr have been successfully used as part of official citizen science campaigns and also as data sources for ecology and wildlife observation studies. This shows their potential as passive citizen science sources. Both networks provide a relatively easy access to the data uploaded and represent different types of SNSs (i.e. images versus text). Twitter is a micro-blogging social network while Flickr is is one of the largest photo-sharing social network platforms. Thus, they provide a platform for experimenting with verification methods suitable for images and text-based posts, which is the reason for using Twitter and Flick as exemplary social media platforms for the research in the thesis.

2.2.5 Summary

Social media data, especially Flickr and Twitter, has been identified as a useful source of ecological data. However, such informal sources of information require verification before data being used in ecology-related research. Additionally, most of the research on using social mining techniques related to wildlife is limited in scale and it is using manual or semi-automatic verification analysis. Additionally, most of the automated image verification techniques require large amounts of annotated photos which are hard to obtain as the annotation needs to be performed by experts. Research is also lacking a large scale comparison between traditional citizen science portals and social media platforms. We address this research gap with research question **RQ 1**.

Finally, this section outlined three main aspects of preparing social network datasets to be used as part of wildlife-related research. These are: verification of genuine wildlife-related posts which requires classification approaches, need for geo-referencing tools, and ability to plot movement data to support identifying and tracking moving objects. These are further discussed in the subsequent sections, Section 2.3, Section 2.4, and Section 2.5.

2.3 Supervised Text Classification

2.3.1 Definition and Applications

Text or sentence classification methods typically use supervised machine learning to assign one or more labels to a given sentence or document [Deng et al., 2019, Zhong and Enke, 2019]. Text classification for social media data can be particularly challenging because of the short text sequences [Chen et al., 2019], noisy data, and large number of misspellings and jargon language used, as well as the presence of polysemous words [Bouazizi and Ohtsuki, 2019].

The main steps of the text classification process involve feature extraction, feature integration, and using a machine learning algorithm to build a predictive model for labeling unseen text instances (see Figure 2.3).

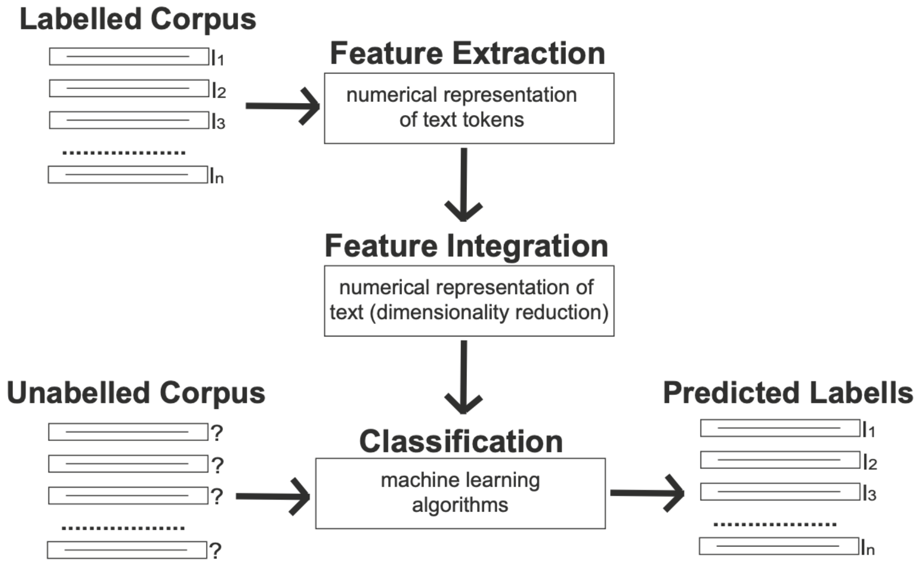

Labelled Corpus
l1
l2
l3
ln
Feature Extraction
numerical representation of text tokens
Feature Integration
numerical representation of text (dimensionality reduction)
Unabelled Corpus
?
?
?
?
Classification
machine learning algorithms
Predicted Labells
l1
l2
l3
ln

Figure 2.3: Overview of Text Classification Pipeline

Text classification methods have been extensively used in social media mining for categorising and filtering data. Examples include detection of depression [Burdisso et al., 2019], real-time emergency response [Imran et al., 2020, Luna and Pennock, 2018, Daume, 2016, Martin et al., 2020], marketing research [Hartmann et al., 2019], and many more. However, text classification has not been widely deployed in categorising and identifying wildlife-related observations. The thesis is particularly concerned with using sentence classification for verifying text-based social media posts related to wildlife observations which can be used for supplementing citizen science collections. We discuss relevant work and challenges associated with this task in Sections 2.3.5 and 2.3.6.

Our text verification approach consists of performing experiments with various machine learning and feature extraction and integration methods in order to identify classification methodologies suitable for identifying wildlife-related posts on social media. Thus, in the following, we describe classical machine learning approaches as well as the state-of-the-art neural network models used for text classification (see Section 2.3.2) which have become commonly adopted for categorising social media data. Further, we explain methods for extracting and integrating the features often used as input for classification models (see Section 2.3.3 and Section 2.3.4).

2.3.2 Machine Learning Approaches

Classical Machine Learning

We refer to classical machine learning as a group of algorithms which are often coupled with feature vectors that represent the frequency of occurrence of individual words. A drawback of such approaches is that they are limited in their capacity to deal with out-of-vocabulary words (i.e. words in the test data that were not observed in training) and with fine-grained distinction between classes [Joulin et al., 2017]. However, classical machine learning algorithms such as Support Vector Machines (SVM) and Logistic Regression are still widely used for many social network classification tasks [Çöltekin and Rama, 2018, Mohammad et al., 2018] and ecology studies [Jeawak et al., 2017, 2018, Jauhiainen et al., 2019, Martinc and Pollak, 2019, Jeawak et al., 2020]. For the rest of this section, we review some of these machine learning algorithms which we have also experimented with in the thesis.

Support Vector Machines (SVM) Boser et al. [1992]: The objective of the support vector machine algorithm is to find a hyperplane (decision boundary) in an N-dimensional space(N - the number of features) that distinctly classifies the data points. The optimum hyperplane is the one with the maximum distance between data points of both classes. Hyperplanes are decision boundaries that help classify the data points. Data points falling on either side of the hyperplane can be attributed to different classes. Also, the dimension of the hyperplane depends upon the number of features. SVM classifiers tend to perform well with a limited amount of labelled data and give a strong baseline performance for many classification tasks.

Naive Bayes classifier: These algorithms are based on applying Bayes' theorem with the assumption of conditional independence between every pair of features given the value of the class variable. Despite their simplicity, they tend to perform classification fast and have led to strong performance in various classification tasks [Prabhat and Khullar, 2017]. Further, these algorithms provide explanations of the features that were most significant for classifying given test instance.

Logistic Regression (LR): Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable and it is used to predict discrete data. It is widely used in classification to solve problems such as identifying spam emails, fraud in online transactions, and medical notes applications for identifying mentions of tumor malignant or benign. It can easily extend to multiple classes (multinomial regression) and makes no assumptions about distributions of classes in feature space. However, non-linear problems cannot be solved with logistic regression because it has a linear decision surface [Walker and Duncan, 1967].

Neural Network Machine Learning

Neural network models in contrast to some classical classification approaches such as those described above can capture complex non-linear relationships. Earlier neural network models commonly use a feed-forward approach, which processes the words of text input in a sequential manner with one word followed by the next word (including for their representations within the layers of network). Examples of such neural networks are recurrent neural network (RNN) and long short-term memory (LSTM) which have been extensively used in various text classification tasks [Xiao and Cho, 2016], including social network-related classification [Huang et al., 2019, Gambäck and Sikdar, 2017, Poria et al., 2016]. Though they process one word at a time in sequence they do often include methods to retain, at each stage, knowledge of other words in the input sequence. However such models can struggle to capture effectively these long term dependencies as doing so depends typically on a back-propagation training process that involves calculating gradients where those gradients can become unmanageable due to being either too high or too low (referred to as exploding and vanishing gradients respectively). The LSTM architecture mitigates this somewhat with a gating unit which allows it to selectively determine what to remember over long spans reducing the number of successive gradient calculations. Despite, this improvement, these neural models can still fail at providing more context-specific representations and tend to be computationally expensive [Merity et al., 2018, Yang et al., 2018b].

An example of a widely used neural network classification model is the fastText pipeline classifier which addresses this problem with an approach based on word embeddings (see also section 2.3.3) that represent the meaning of words with multi-dimensional vectors based, in the case of fastText, on parts of words [Joulin et al., 2017]. The approach enables good prediction accuracy in classification tasks where some classes have very few examples. The fastText classification pipeline is referred to as a shallow neural network as it consists of a single layer of neurons and it is also referred to as a linear classifier (in contrast to multi-layer neural networks). The classification pipeline initially represents each word in a sentence with its corresponding embedding. These word representations are then averaged to create a sentence representation, which is fed into the classifier layer. The fastText classification pipeline has given a strong performance in many classification tasks [Joulin et al., 2017]. However, it has not received much attention in ecology studies or social media-related research.

The limitations of the early feed-forward neural network architectures are addressed in the transformer architecture in which the representation of each word is directly connected to the representation of every other word [Merkx and Frank, 2020]. These connections use attention methods (typically a form of dot product) that update one representation as a function of other connected representations. The non-sequential manner in which data is processed enables capturing more relationships between words and thus provides better contextual representation [Vaswani et al., 2017].

For instance, BERT (Bidirectional Encoder Representations from Transformers) [Devlin et al., 2019] is a transformer-based model that achieves state-of-the-art performance in various NLP text classification tasks. Although BERT has been used to classify Tweets, such as to infer their locations [Scherrer et al., 2021], we are not aware of previous work to date in applying such transformer models to wildlife observation. The BERT model, similar to other transformer-based models, is created in two phases. In the pre-training phase, word representations are trained from scratch using masked language modelling with only unlabelled data. In the fine-tuning phase, a pre-trained model is adapted to a particular downstream task. For example, for the classification task, the embedding of the token called [CLS] is extracted from the last hidden layer of the BERT neural network representation. The output corresponding to that token can be considered as an embedding for the entire input sentence. This token is passed to a separate neural network layer (sequential classifier) in order to predict labels for unseen instances. In this phase, labelled training data is required.

2.3.3 Feature Extraction

Feature extraction consists of selecting the tokens which will participate into training of the classification model and building numerical representations for the selected tokens. We distinguish between three main types of feature extraction techniques, i.e. a simple frequency-based feature representation, word embeddings consisting of multi-dimensional vectors that represent the semantics of words and capture semantic relationships between words [Mikolov et al., 2013b, Pennington et al., 2014, Bojanowski et al., 2017], and transformer language models (also referred to as contextualised word embeddings) that have a neural network architecture.

Frequency-based Representation

Traditional feature representation techniques represent words simply as indices in a vocabulary. An example is the n-gram model, often used in combination with classical machine learning approaches [Peng and Dean, 2007, Mikolov et al., 2013b], described in Section 2.3.2 where the input features consist of a vector representing the presence or frequency of each word from an entire text collection (with most elements therefore being zero). Such approaches that represent words directly do not capture the meanings of words and will fail to take account of out-of-vocabulary words encountered when applying the classifier to unseen data [Peng and Dean, 2007, Mikolov et al., 2013b].

Word Embedding Representation

As mentioned earlier, word embeddings represent words as low-dimensional vectors intended to capture the semantics of the respective words. Words that are similar in meaning will tend to occur close to each other in the vector space, enabling measurement of similarity between individual words or of analogy between pairs of words. Common techniques for generating word embeddings are Continuous Bag-of-Words (CBOW) and skip-gram [Mikolov et al., 2013a]. The skip-gram approach learns to predict a target word based on a nearby word. On the other hand, the CBOW model predicts the target word according to its context. Unlike most of the previously used architectures for learning word vectors, training of the skip-gram or CBOW model does not involve dense matrix multiplications which makes the training more efficient [Mikolov et al., 2013a]. One of the first widely used models using skip-gram approach for building term representations is Word2Vec [Mikolov et al., 2013a]. Word2Vec embeddings [Mikolov et al., 2013a] are generated with a two-layer neural model where the output of the model is an embedding matrix, where each term (single or multi-token) from the corpus vocabulary is represented as an n-dimensional vector. A limitation of Word2Vec is that it ignores the morphology of words by assigning a distinct vector to each word. This Word2Vec limitation is addressed in the fastText approach

[Bojanowski et al., 2017] where each word is represented as a bag of character n-grams. A vector representation is associated with each character n-gram and words are represented as the sum of these representations. This enables the construction of vectors for rare or misspelled words. Glove embeddings of words [Pennington et al., 2014] are generated from a matrix of the co-occurrence of pairs of words such that the learnt embeddings have the property that the dot product of pairs of word embeddings reflects the log of the probability of the co-occurrence of the respective words. Word embedding models, pre-trained on large corpora of unlabelled data such as news corpora, are widely used in solving NLP problems by adapting the pre-trained models to the specific task or domain.

Language Model Representation

A limitation of the word embedding models described above is that they produce a single vector of a word independent of the context in which it appears. Language models, built using transformer-based principles described in Section 2.3.2, address this limitation by computing dynamic representations for words based on the context in which they are used [Peters et al., 2018, Devlin et al., 2019].

One of the first state-of-the-art transformer-based models is BERT [Devlin et al., 2019]. It is built using using transformer-based Masked Language Model (MLM) which randomly masks some of the tokens from the input, and the objective is to predict the original vocabulary id of the masked word based only on its context. Unlike left-to-right language model pre-training, the MLM objective enables the representation to incorporate the left and the right context, which allows more context-based representations. The pre-trained BERT model can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of tasks, such as question answering and language inference, without substantial task-specific architecture modifications [Devlin et al., 2019]. We described the usage of BERT for classification in Section 2.3.2.

Transfer Learning

Neural network models usually require large volumes of annotated data to perform well [Tan et al., 2018]. However, obtaining such large volumes of labelled datasets is a time- and resource- consuming process [Tan et al., 2018]. Transfer learning is a widely applied technique in neural network machine learning where representations learned for one task can be applied or adapted for a different task using less volumes of training data, compared to training neural models from scratch [Bailey and Chopra, 2018].

Some neural network architectures, such as transformer models, employ transfer learning in which, for NLP applications, the model is trained initially on large generic text corpora, referred to as pre-training, which can be very time consuming. To apply the model to a specific task, it can be fine-tuned on a smaller set of application specific data that allows the model to adapt to the particular application [McCann et al., 2017, Howard and Ruder, 2018] (as already mentioned in previous sections). The pre-training can be expected to have exposed the model to a much wider vocabulary and range of language uses than in the task-specific training dataset. However, the word representations obtained through pre-training might be similar to those of related words that do appear in the task-specific dataset, which allows the model to generalise better when it is applied to unseen data [Goldberg, 2016]. Such language models and sets of word embeddings could also be learned from scratch using the application dataset, resulting in the case of conventional word embeddings in corpus-trained embeddings.

2.3.4 Feature Integration

The feature integration step involves building a representation of the entire sentence using the token feature vectors, obtained during feature extraction. A simple but widely used method for building text representations is bag-of-words (BOW) approach which is based on the statistics of the n-gram occurrences, counts or tf-idf, of the 1-grams and 2-grams in the given text. Further, many feature integration approaches are based on applying dimensionality reduction techniques over the word embedding representations in order to obtain sentence representations. We conduct experiments with different dimensionality reduction techniques in Chapter 4.

2.3.5 Text Classification for Social Media

Text classification approaches have been extensively used for various social media-related applications. These include disaster management [Reynard and Shirgaokar, 2019, Huang et al., 2019, Yu et al., 2019], e-commerce [Swamy and Gorabal, 2020], healthcare-related studies [Liu and Chen, 2019, Sarker et al., 2018, Al-Garadi et al., 2021, Tokala et al., 2018], detection of misinformation, hate-speech, and sarcasm detection [Gambäck and Sikdar, 2017, Poria et al., 2016].

Many of the presented approaches use early neural networks such as Convolutional Neural Network (CNN) for building classifiers or classical machine learning algorithms. For instance, the authors of Reynard and Shirgaokar [2019] present a combination of geospatial and machine learning techniques to categorize geolocated Tweets about Hurricane Irma in Florida. The authors employed sentiment analysis to classify tweets about damage and/or transportation. They used a multinomial logit model to examine which features of the Tweet, Tweeter, or location were likely to be associated with negative or positive sentiments. [Huang et al., 2019] identified disaster related social media content experimenting with two CNNs, Inception-V3 CNN and word embedded CNN. The neural network models are used to extract visual and textual features which are then concatenated to form a combined feature representation which is fed to the classification model. Results showed that combining the image and textual features has benefits in classification. Similarly research by Gambäck and Sikdar [2017] used a CNN framework but for identifying hate-speech text. Four classification models were trained on respectively character 4-grams, Word2Vec word embeddings, randomly generated word embedding vectors, and word embedding vectors combined with character n-grams. The authors used as a baseline Logistic Regression model. Results showed that CNN-based classifier achieve higher F1-measure while Logistic Regression with character n-grams model gives higher recall. A similar approach is proposed by Poria et al. [2016] who use CNN for identifying sarcasm features. Tokala et al. [2018] present a deep learning approach to distinguish Tweets that present personal medication intake, possible medication intake and non-intake. They performed extensive experiments with classical machine learning algorithms such as Logistic Regression, Random Forest, SVMs, Gradient Boosted Decision Trees (GBDT) and neural network architectures such LSTM and CNN.

The aforementioned approaches all assume the presence of large amounts of training data (i.e., more than 20,000 training instances). Further, the described work do not consider more recent state-of-the-art transformer-based neural network models.

A more generalised classification model for filtering crisis Tweets is proposed in Li et al. [2018]. The method is based on the use of pre-trained and specialised corpus-trained word embeddings for representing the Tweet's vocabulary. The research presents two approaches for building Tweets embedding vectors. The first approach is based on calculating either the mean of all word embeddings in a Tweet, the TF-IDF weighted average (of each dimension) of the word embeddings, or concatenating min, max and average of the embeddings of each word in a sentence along each dimension. The second approach uses sentence encoding techniques of respectively SIF [Arora et al., 2017], InferSent [Conneau et al., 2017] and tfSentEncoder [Cer et al., 2018]. The performance of the different embedding methods is evaluated with Naive Bayes, Random Forest, K-nearest Neighbours and SVM classifiers. The authors provide extensive analysis on how different word embedding models affect classification performance. However, the research does not consider the use of multiple classification methods and state-of-the-art language models. Additionally, it does not compare performance of approaches for domains with a limited amount of training data. In the verification approach presented in Chapter 4, we use similar approaches for building Tweets representations. However, the work here differs from and extends theirs in using different classification models, settings, and pre-processing techniques.

Recent research [Al-Garadi et al., 2021, Guo et al., 2020, Liu et al., 2021, Lopez-Lopez et al., 2021] on text classification for social media is using transformer-based methods. For instance, the study by Guo et al. [2020] compares the performances of different variants of pre-trained transformer-based models, RoBERTa, BERTweet and Clinical-BioBERT, on a wide range of social media text classification datasets. Results showed that RoBERTa and BERTweet perform comparably on most datasets, and considerably better than Clinical-BioBERT, even on health-related datasets. Similarly, the authors of Liu et al. [2021] use BERT architecture to develop CrisisBERT which is fitted for two crisis classification tasks, namely crisis detection and crisis recognition. They compared their model to the classification models of Logistic Regreession, SVM and Naive Bayes, as well as CNN and LSTM. Similarly, Lopez-Lopez et al. [2021] use BERT and roBERTa to detect sexism on social media and compared them with traditional machine learning approaches, such as SVM, Logistic Regression, SGD-based classifier and XGBoost.

Finally, Al-Garadi et al. [2021] provide extensive comparison between multiple transformer-based language models, which utilize tweet-level representations that enable transfer learning (e.g., BERT, RoBERTa, XLNet, AlBERT, and DistilBERT). Further, they propose fusion-based approaches, and compared them with several traditional machine learning models. The authors used the proposed approach for the automatic detection of non-medical prescription medication use from social media. Results suggested that transformer-based models are more stable and require less annotated data compared to the other models.

2.3.6 Text Classification for Wildlife Data

Relevant research on proposing classification approaches for identifying genuine wildlife occurrences on social media is very limited. There are, however, a number of studies that apply machine learning to detect various aspects of the environment and to detect postings that relate to particular environmental topics. Some of these exploit data from both images and text as in Leung and Newsam [2012] who use Flickr images and the tags describing the images to perform land-use classification with an SVM classifier. The approach was evaluated on two university campuses and three land-use classes were considered: Academic, Residential, and Sports. The study showed that the text entries accompanying photos are informative for geographic discovery. In other examples of classifying aspects of the environment such as Jeawak et al. [2017], SVM classifiers take as input a bag-of words feature vector combining text from Flickr postings with environmental data. They found for all experiments, including predicting species distribution, scenicness, land cover and climate factors, that the use of the social media data always improved the results relative to only using the environmental data input. An associated study by the same authors [Jeawak et al., 2018] used Flickr data and focused on bird species distribution. They demonstrated the benefit of a meta-classifier approach that combines prior predictions with machine learning features that represent the presence of the species name in postings in the vicinity of the predicted location. In other related studies for similar prediction tasks, the same authors presented methods for creating embeddings (i.e. vector space representations) of geographic locations using methods based on the GloVe word embedding technique [Jeawak et al., 2019]. The geographic embeddings were extended to spatio-temporal embeddings in Jeawak et al. [2020]. In both cases the embeddings were used as input features to SVM, and with spatio-temporal embeddings also to a MLP (multi-layer perceptron, a basic form of neural network) classifier, and were demonstrated to provide improvement relative to the simpler feature vector-based (bag-of-words) approaches. The use of MLP did not provide significant benefit relative to SVM.

Work by[Monkman et al., 2018] presents a text and data mining (TDM) approach applied to social media from specialised forums to gather spatio-temporal information on wildlife recreation activity relating to fishing a particular species, European seabass, that is subject to legal controls on overfishing. NLP-based software was used in a ruled based system to classify sentences based on their inclusion of terms from a manually constructed lexicon. Stringham et al. [2021] present research on categorising online wildlife trade data. They test the ability of a suite of text classifiers to extract relevant advertisements from wildlife trade occurring on the Internet. The authors use a collection from Australian classified websits where people can post advertisements of their pet birds. The authors compare three classical machine learning algorithms, Logistic Regression, Multinomial Naive Bayes, and Random Forest. The conclusions from this work show that text classification is a suitable method for categorising online wildlife trade data, however the approaches might be context-dependent.

Overall, the aforementioned research is limited in scale and classification approaches are mainly based on using statistical classification models, without fully exploring different feature selection and classification methods.

2.3.7 Summary

Text classification approaches are widely adopted in various research in social media mining with a prominence of applications in the medical domain, crisis identification and emergency response, as well as fake news discovery. Most of the presented work is based on using deep learning such as early neural networks, including CNN and LSTM, or using state-of-the-art transformer models such as BERT and roBERTa. The majority of the approaches have been evaluated with large amounts of training data and they lack extensive comparison between different feature extraction and feature integration approaches and how these affect performance of classifiers. Additionally, state-of-the-art performance of transformer-based models have not been fully explored in ecology-related studies where there is a limitation on amounts of labelled data available. Further, most of the research on text classification for wildlife is limited to using classical machine learning models such as SVM or semi-automatic methods, and limited in scale. We address these research gaps with research question **RQ 2**.

2.4 Geo-referencing Micro-blogs

2.4.1 Definition and Applications

Geo-referencing social media data refers to the process of assigning coordinates (latitude and longitude values) to a social media posting. Social media data associated with coordinates have been a valuable source of information for many studies such as health, disaster management, tourism, environmental monitoring, crime, civil unrest and marketing [Stock, 2018b, Zheng et al., 2018, Gelernter et al., 2013, Castillo, 2016]. In particular, as mentioned in Section 2.2.3, geo-referenced social media data collections can be used to facilitate studies of wildlife distribution patterns which in turn are increasingly important for alerting rapid ecosystem changes such as climate change, diseases spread, and invasive species occurrences [Amano et al., 2016, Barve, 2014].

Many social media platforms use location information provided by the user mobile devices. However, many users deactivate the location sharing abilities and thus a large proportion of the social media content do not include spatial coordinates [Middleton et al., 2018, Di Rocco et al., 2021]. This problem is not as vivid for image-sharing platforms such as Flickr as it is for the micro-blogging social platforms such as Twitter [Middleton et al., 2018, Stock, 2018a]. For instance, recent research by Tuxworth et al. [2021] showed that from 87,894,019 Tweets collected for identifying emergency events only 0.34% of them were associated with coordinates. Different research also estimated that the rate of geotagged messages out of all messages vary from 0.8% [Musaev et al., 2015] to 6% [Chen et al., 2014].

Despite the large potential of Twitter to provide a valuable source of wildlife observational data [Daume, 2016], its usage in such studies is limited due to the lack of coordinate data. Therefore, in the thesis we focus specifically on the problem of geo-referencing Twitter postings using only the message content.

In the rest of this section, we review and discuss main approaches used for geo-referencing posts. Existing approaches can be split into three main groups: gazetteer-based methods (see Section 2.4.2), language modelling-based methods (see Section 2.4.3), and regression-based methods (see Section 2.4.4).

2.4.2 Gazetteer-based Methods

Gazetteer-based methods involve extracting location names from text and then mapping these place names to coordinates [Stock, 2018a]. The approach can be broadly divided into two main steps. The first step involves identifying place names within text using Named Entity Recognition (NER) in combination with gazetteers to identify whether the extracted named entity is a genuine location and to extract associated co-ordinates. Some of the most widely used NER tools for location names extraction include the Stanford NER tool [Bassi et al., 2016, Li et al., 2015], GATE [Jaiswal et al., 2013], spaCy [Honnibal and Montani, 2017] and AllenNLP [Gardner et al., 2018]. More recent approaches are based on using deep learning neural network methods such as CNN [Kumar and Singh, 2019] and BERT [Davari et al., 2020]. Commonly used gazetteers include GeoNames [Inkpen, 2016, Zhang and Gelernter, 2014, Ikawa et al., 2013] and OpenStreetMap [Daly et al., 2013, Di Rocco et al., 2016].

The second stage consists of location name disambiguation. Ambiguity of location names can refer to the presence of multiple location names within a message text as well as when a single location name is associated with multiple coordinates. A number of methods are used to try to disambiguate place names, including weighting by population, geographic feature types, geographical proximity and other place names that are found nearby in the text [Zhang and Gelernter, 2014, Inkpen, 2016]. Location names disambiguation is an important step and it can be particularly challenging [Middleton et al., 2018]. Rule-based approach is commonly employed to assist in identification of place names following NER and sometimes gazetteer use, by looking for patterns of language within which place names frequently occur [Gu et al., 2016, Zhang et al., 2017].

A simple and popular gazeteer-based method is the google-geocoder approach which uses NER to identify possible locations in text and then sends it to the Google Geocoder API to get a location reference. Another similar approach by Zhang and Gelernter [2014] uses GeoLocator and it is based on extracting entities and verifying them using GeoNames gazetteer. A more recent location name extraction tool called GeoTxt [Karimzadeh et al., 2019] was developed for the extraction and geolocation of place names in unstructured text. The tool offers six NER algorithms for location name extraction and multiple gazetteers for toponym identification. These approaches perform location extraction without considering disambiguation or matching of location names to coordinates.

In contrast, Middleton et al. [2018] create a geoparsing library performs extraction of location names as well as location disambiguation. It uses a local OpenStreetMap database. The location name extraction step is based on NER where named entities are matched to OpenStreetMap (OSM) locations. After that, location name disambiguation is performed using rules such as token subsumption (rejecting smaller phrases over larger ones (e.g. New York will prefer [New York, USA] to [New York, UK])) and more. The authors also propose a hybrid approach for location names extraction using language modelling and gazetteers which proved beneficial when compared to approaches based only on using gazetteers. However, the method aims at building hybrid approaches for assigning most suitable location name to given pair of coordinates. In the thesis, we focus on using gazetteers as part of hybrid approaches for facilitating more accurate geo-referencing of social media posts.

Gazetteer-based methods are simple and do not require training data. Further, these approaches can lead to satisfactory results as long as the social media data is rich in location names. However, location disambiguation still remains a challenging problem especially considering the short length of the social media posts and the large number of misspellings and jargon used. Further, in addition to location names which refer to administration regions such as cities and towns there are many geographic names that are 'dynamic' (e.g., shop names, emerging new hip areas). As a result, maintaining comprehensive and up-to-date gazetteers of geographic names is a very challenging task [Kordopatis-Zilos et al., 2017]. Further, these approaches usually fail at predicting locations on fine-granularity and suffer with location disambiguation problems Stock [2018b].

2.4.3 Language modelling Methods

Much of the work on geocoding social media posts is based on using language modelling methods [Di Rocco et al., 2021, De Rouck et al., 2011, Häberle et al., 2019, Kumar and Singh, 2019, Paule et al., 2019, Rahimi et al., 2017b]. Language modelling is used to reflect the entire range of words used in messages, with the idea that locations are characterised by all the words used to refer to them, not just the toponyms [Stock, 2018b]. Location-based terminology may include location-specific words such as venue names, dialect or language style typical for particular locations [Stock, 2018a].

This approach usually consists of four basic steps. First, messages are grouped within a set of regions. Then, feature selection methods are applied to identify location-relevant terminology. After that, most approaches use a form of text classification to identify which area is most likely to contain the true location of the resource. The final step consists of finding the most appropriate location for the given unlabelled instance within the identified region which could be for example the centroid of the region or the location of an existing post that is most similar to the one being georeferenced [Van Laere et al., 2013]. The use of machine learning approaches allows the extraction of hidden patterns which can be much more valuable for providing geographic information than the information derived using gazetteers [De Rouck et al., 2011].

Previous work considers three main language modelling approaches [Fornaciari and Hovy, 2019a]. These are based on i.e., using fixed cell sized geodesic grids [Serdyukov et al., 2009, Wing and Baldridge, 2011, 2014, Hulden et al., 2015, Melo and Martins, 2015, Eisenstein et al., 2010a, Cheng et al., 2010, Kinsella et al., 2011, Backstrom et al., 2010, Cheng et al., 2010, Rahimi et al., 2015], clustering approaches [Van Canneyt et al., 2013, Laere et al., 2014b], adaptive grids [Kordopatis-Zilos et al., 2017, Di Rocco et al., 2021, Roller et al., 2012], or predefined administrative regions where coordinates are mapped to the closest administrative area [Fornaciari and Hovy, 2019a]. However, creating these grids or clusters usually requires large amounts of training data especially when predictions are performed on sub-city level [Van Laere et al., 2013]. Some recent research on geo-referencing using language modelling has focused on developing approaches suitable for scenarios with a limited amount of training data. For instance, Di Rocco et al. [2021] present a method for fine granularity location prediction with an algorithm (Sherloc) that uses a gazetteer and associated ontology of feature types to create a metric space language model that is specific to a particular geographic region or grid cell. The model is an embedding (based on a dimensionality reduction procedure) of the place name knowledge for the respective region, where toponyms are reduced to their individual tokens (words). Sherloc extracts the toponyms from the social media posts and matches the embeddings of their components to the elements of the embedding space. The semantically closest toponyms to a message are found and clustered, taking the centroid of the smallest cluster as the inferred location. While Sherloc requires no prior training, and it can infer the location at sub-city level with high accuracy, the approach can identify locations at sub-city level only when the parent region (e.g. a city or grid cell), referred to as the reference area, is given, and only when the message contains at least one toponym. In the thesis, we also experiment with a strategy employing NER and gazetteer methods, the purpose of which in our case is to improve the precision of our regression-based language models. However, our approach does not require prior knowledge of the local region and does not require that a toponym is present in the text to be georeferenced.

Alternative approaches address the lack of geo-referenced training data by exploring the transferability of language models built using different social media sites. For instance, Laere et al. [2014a], building on De Rouck et al. [2011], investigate the application of probabilistic language models trained on Flickr and Twitter to assign coordinates to Wikipedia articles. The results showed that language modelling approaches trained with Flickr data or a combination of Flickr, Wikipedia and Twitter data outperformed language models trained only with a Wikipedia dataset or classical gazetteer-based methods (using Yahoo! Placemaker and Geonames). The authors use classical machine learning models for building classifiers which requires pre-processing the heterogeneous data sources before building feature vectors. Further, classical machine learning models still require large amounts of training data. The need of large amounts of training data still remain a major problem associated with language modelling approaches [De Rouck et al., 2011, Stock, 2018b, Di Rocco et al., 2021]. This is particularly problematic for Twitter where many posts are not associated with coordinates and the collection of large amounts of data is restricted by the Twitter API limitations. It is even more challenging when data need to be collected for a very specific purpose, region, or time frame. Further, the primary output of language modelling approaches is an area, within which a particular location needs to be inferred. Also, the partitioning of the training data into a finite set of areas superimposes a certain factor of scale to the results where, depending on the information available, such a partitioning can be too coarse for one resource or too fine-grained for another resource, although some multi-scale methods attempt to address this issue [Kulkarni et al., 2020].

2.4.4 Regression-based Methods

Regression-based methods tackle the problems associated with language modeling approaches by creating models which can predict coordinates for a given instance without the need to split a given region into distinct areas. These methods are similar to language modelling in that they use supervised machine learning algorithms for predicting the coordinates of unlabelled instances. However, in the language modelling-based methods, the classes are geographic areas while in the regression-based approaches, models directly predict coordinates. These approaches have received less attention compared to language modelling, but have the potential advantage of not needing to specify grid cell sizes or cluster numbers or determine locations within such regions. Therefore, we want to explore these methods and analyse their potential in geo-referencing wildlife-related micro-blogs.

The research by Eisenstein et al. [2010b] is one of the first to formulate the problem as a regression task predicting the coordinate values as numerical values. Priedhorsky et al. [2014] use Gaussian Mixture Model (GMM) algorithms to predict the coordinates of a given Tweet. They learn a mixture of bi-variate Gaussian distributions for each individual n-gram in the training set. During prediction, they add the Gaussian mixture of each n-gram in the input text, resulting in a new Gaussian mixture which can be used to predict a coordinate with associated uncertainty. The approach does nor require gazetteers or other supplementary data. Also, because the approach predicts geographic coordinates directly, there is no need to pre-specify regions of interest. The approach outperformed other regression and classification methods and proved to perform well for relatively small amount of training data (30,000 Tweets). Rahimi et al. [2017a] extend on the work by Priedhorsky et al. [2014] by using neural network models which incorporates mixtures of Gaussian distributions in order to predict coordinates for Twitter data.

Alternatively, Fornaciari and Hovy [2019b] combine language modelling and regression-based methods for coordinate prediction. They build a multi-task learning CNN model by combining label classification with regression for predicting geo-coordinates for Twitter data. The authors found that the two methods complement each other especially when using more fine-grained labels where regression help improve precision.

Recent work on using regression for geo-referencing social media data experimented with transformer-based models for building multivariate regression approaches. Thus Scherrer et al. [2020], adapt BERT sentence classification architecture for the regression task in order to predict coordinates for social media posts. The datasets used in that study are a generic Twitter dataset and two Jodel datasets collected in different languages. They perform experiments with various pre-trained language models and hyper parameter settings. In particular, they perform experiments with three BERT language models — a BERT model trained on the task training data, a pre-trained multilingual BERT model, and language specific pre-trained BERT models. Results show that using language-specific BERT model which is then fine-tuned for multivariate regression leads to significant improvements over the other language models and a traditional machine learning approach based on Support Vector Regression (SVR) with TF-IDF character n-grams. Scherrer et al. [2021] build on this research where they compare the regression BERT model (geoBERT) to classification models and experiment with two different vocabulary sizes. Results confirmed findings from Scherrer et al. [2020] where regression-based model outperformed language models.

2.4.5 Summary

Geo-referencing social media posts is a widely researched area where most approaches use language modeling where data points with similar coordinates are grouped into discrete set of classes. Most of the available work is based on using large amounts of training data (millions of training instances). Further, language modelling require an additional step for inferring coordinates based on the predicted region for a given instance. Further, choosing the most optimal way of clustering training data is a challenging task. Another group of work, based on regression, tackles these problems by creating models which can predict coordinates for a given instance without the need to split a given region into distinct areas. However, regression-based approaches are not fully explored in literature. Further, the majority of research in geo-referencing social media data is based on using classical machine learning algorithms or early neural networks. Despite the state-of-the-art performance of transformer-based models for various text analysis tasks, these models have not been well explored in ecology-related studies or for geo-referencing social media posts. We aim to explore further the potential of such models for georeferencing wildlife-related Tweets. Finally, hybrid approaches based on gazetteers and language modelling methods have been successfully used in previous research. However, such methods have not been explored in combination with regression. The aforementioned research gaps are addressed by research question **RQ 3**.

2.5 Identification of Movement Patterns Over Time and Space

Object tracking finds applications in many research problems. For example, when making inferences concerning future weather conditions, it is necessary to track weather phenomena such as a snow storm [Atluri et al., 2018, Corcoran and Jones, 2018]. Large volumes of spatio-temporal data are increasingly collected and studied in diverse domains, including climate science, social sciences, neuroscience, epidemiology, transportation, mobile health, and Earth sciences [Atluri et al., 2018]. Another application is tracking migration data and identifying migration patterns of wildlife, as well as studying swarm behaviour where a swarm is defined as a set of agents (animals, people, robots) moving in close spatial proximity to each other. However, a big challenge of tracking objects with dynamic topology is that the object's topological properties can change over time. For instance, splitting an object into multiple objects or the merging of multiple objects into a single object [Corcoran and Jones, 2018]. Therefore, identifying movement patterns of objects can be a challenging task. Many of the existing works on tracking objects focus on modelling topological relations between regions and detecting changes in topological features. However, these approaches do not compute when components first appeared and subsequently disappeared. It also does not determine if the connected components at different times are in fact the same or different connected components. The research presented by Corcoran and Jones [2017] addresses these limitations using spatio-temporal analysis to keep track of the appearance and disappearance of individual objects. The model presented by Corcoran and Jones [2017] encodes the spatio-temporal characteristics of topological features of objects, such as holes and connected components. The authors identify objects that persist between successive time slices and records the start and end duration of each object across the times slices. The persistence of topological features with respect to time is computed using zig-zag persistent homology. Zig-zag homology gives a set of intervals representing the periods of existence of the topological features in question [Carlsson and De Silva, 2010]. In order to facilitate statistical and data mining techniques the set of intervals are converted into a persistence landscape. A persistence landscape is a vector space representation of topological features, which makes it easy to be combined with statistical and machine learning tools [Bubenik, 2015]. Bubenik introduces a set of different algorithms for calculating persistence landscapes in Bubenik and Dłotko [2017]. A limitation of the standard method employed for computing these persistence intervals is that it is only capable of inferring the appearance and subsequent disappearance of objects but does not maintain their identities relative to their respective regions (components) in the source image. Corcoran and Jones [2018] extend these persistent homology methods to attach unique identifiers to objects with dynamic topology, from their creation to disappearance, keeping track of the image locations of the objects. When one object is merged with another, one of the objects will lose their original identity, while, when an object splits, one or more separate identities will be attached to the newly spawned objects (depending on how many there are). It may be noted that earlier work that implemented methods for tracking topological change, notably [Worboys and Duckham, 2006], employed a rule-based approach that was acknowledged as not being complete with regard to all possible change situations.

The approach by Corcoran and Jones [2017] and further developed in [Corcoran and Jones, 2018] provides a novel tracking model capable of tracking objects whose topological properties change over time. It has been successfully used for identifying moving objects related to animals (fish) and environmental phenomena (cloud movement). The methods also facilitate the application of statistical and data mining techniques. It is demonstrated that the proposed model can be used to perform retrieval and clustering of swarm behaviour in terms of topological features. This makes these methods suitable to be further used for trajectory extraction for moving objects. However, this area has not been explored yet.

2.5.1 Data Mining Techniques for Extracting Trajectories of Movement

A trajectory is a sequence of geo-locations with corresponding timestamps in spatio-temporal space [Feng and Zhu, 2016]. Analyzing the trajectories of moving objects is of interest in many fields in order to understand the dynamics and behavior of those objects [Izakian et al., 2020]. Some examples of applications for extracting and analysing trajectory data include path optimization of logistics companies, improvement in public security management, personalized location-based services, or the migration patterns of animals traveling for better access to food, water, and shelter [Farine et al., 2016, Su et al., 2020].

The widespread use of location-aware devices has led to an increasing availability of trajectory data. As a result, researchers have devoted efforts to developing methodologies including different data mining methods for trajectories [Yuan et al., 2017, Mazimpaka and Timpf, 2016]. Representative works include the design of effective trajectory indexing structures [Su et al., 2020], built to manage trajectories and support high-performance trajectory queries [Su et al., 2020]. Data mining methods are applied to trajectories to detect points of interest (POI), find popular routes from a source to a destination, predict traffic conditions, discover significant patterns, and perform data compression [Su et al., 2020]. Data mining techniques depend on the type of objects whose trajectory is in focus (e.g. people, animals, weather data) and the application (e.g. hot-spot discovery, extraction of mobility profile, discovery of interaction between animals) [Mazimpaka and Timpf, 2016]. Main types of analysis are classification, clustering, frequent pattern mining and group pattern mining [Mazimpaka and Timpf, 2016]. Clustering is a popular method for analyzing trajectories because it provides useful insight into data without the need for a training set [Mazimpaka and Timpf, 2016, Ansari et al., 2020, Wang et al., 2021]. Trajectory clustering aims at finding trajectories that are of the same (or similar) pattern, or distinguishing some undesired behaviours, such as outliers [Yuan et al., 2017, Ansari et al., 2020, Wang et al., 2021]. A main challenge with trajectory clustering is that algorithms need to account for spatial and temporal characteristics of the data where data points need to be processed in a sequential manner (where each point represent the object at a given time). Therefore, predominant algorithms such as DBSCAN [Ester et al., 1996] or K-means [Von Luxburg, 2007] are unsuitable for trajectory data. A drawback of K-means is its tendency to form spherical clusters, which is inappropriate for clustering streamline data [Blazquez-Herranz et al., 2021]. A widely used clustering algorithm for spatio-temporal data, such as trajectories is ST-DBSCAN [Birant and Kut, 2007]. ST-DBSCAN is a density-based clustering algorithm, which originated from DBSCAN. In contrast to the existing density-based clustering algorithms, ST-DBSCAN algorithm has the ability of discovering clusters according to non-spatial, spatial or temporal values of the objects. The algorithm has been applied mainly to road-based data.

2.5.2 Summary

Spatio-temporal analysis based on zig-zag homology helps identify the persistence of objects over time for case studies of swarms and weather imagery data. These methods also facilitate the application of statistical and data mining techniques which makes them suitable for further studies into identifying trajectories of movement patterns.

However, such studies have not been conducted yet. Further, existing work on trajectory data mining is mainly concerned with using pre-processing, classifying and clustering methods related to road optimisation problems. However, there is a lack of research into data mining methods suitable for studying movement patterns related to weather phenomena. Further, suitable algorithms need to be explored for trajectory clustering which takes into account both spatial and temporal characteristics of the data. We address these research gaps with our final research question **RQ 4**.

2.6 Conclusions

The motivation for the research questions declared in the previous chapter lies in the need for developing methodologies which enable the use of untapped social media data for ecology studies and to help enrich official citizen science data portals.

Relevant research on assessing the value of social media data as a supplement to citizen science data collections has been conducted on a very small scale (including only few species) where data verification is performed manually or semi-automatically. Further, these studies lack broad analyses on how social media data can facilitate official data portals considering only statistical or spatial measures, and excluding temporal analyses.

An important aspect of preparing social media data for use in ecology studies is identifying posts related to genuine wildlife observations. Most of the existing verification approaches are based on using domain-trained image recognition tools or involve manual processing. However, obtaining large volumes of training data for training the image recognition tools might be unfeasible. Further, most of the research is limited in scale and it is applicable to only certain species or taxonomic groups.

Verification for text-based messages can be performed using text classification algorithms which have been extensively used for various purposes in social media mining. Most of the existing classification approaches are based on neural network models such as CNN, RNN, and the most recently developed transformer-based models such as BERT. However, most of the research presents big data solutions assuming large amounts of training data. Further, there is lack of comparison between different classification algorithms and analysis into how different feature extraction methods affect the performance of classifiers. Further, using text classification for verifying social media posts relevant to wildlife observations can be challenging because of the more specialised language used (species Latin names and other biology-related terminology). However, this is an unexplored area of research.

Another important step of preparing social media data for use in ecology studies is to ensure posts are geo-referenced to support object tracking and learning of movement patterns. The majority of geo-referencing methods for social media data are based on language modelling where data points are clustered into location-specific regions and classification is applied to identify regions for unseen data points. Further, language modelling requires an additional step for inferring coordinates based on the predicted region for a given instance. Also, splitting regions within clusters or grid cells can be a challenging and data-specific task which requires large amounts of data, even for fine-grained locations. Regression models partially resolve some of the problems associated with language modelling methods as they do not require clustering or additional steps for inferring coordinates. Instead, regression algorithms assign coordinates directly to data points. However, these methods are understudied, especially in combination with state-of-the-art neural network models and transfer learning techniques. Most of the available research is using classical machine learning algorithms and do not provide extensive analysis on strategies suitable for building less data consuming geo-referencing models.

Finally, in order to enable studies of movement patterns, there is a need to establish techniques for extracting trajectories of objects. Recently developed spatio-temporal analysis based on zig-zag homology has been successful in identifying object locations that persist over time. These methods also facilitate the use of data mining techniques.

However, there is no further investigation into how these methodologies can be used to facilitate trajectory extraction.

The research gaps summarised above provided a motivation for large scale study, involving statistical, topological and temporal analysis looking at the potential of social media data to be used as a supplement to citizen science datasets (addressed by **RQ 1**). Further, two approaches for verifying image-based and text-based social media posts have been proposed, both being suitable for validating large and diverse collections, regardless of the species considered at hand (addressed by **RQ 2**). We perform extensive analysis into the use of state-of-the-art transformer models, transfer learning techniques, and regression for building less data-consuming geo-referencing models suitable for wildlife-related posts. We also research two hybrid approaches combining rule-based approaches and transformer-based multivariate regression in order to build more precise geo-referencing models (addressed by **RQ 3**). Finally, we show how zig-zag homology methods can be used to support trajectory extraction where trajectories are further normalised using clustering in order to identify movement patterns (addressed by **RQ 4**).

In the following chapter, we present extensive analysis into the potential of social media data to be used to enhance citizen science data portals. This work helps identify what are the benefits of using social media data for studying wildlife, for what types of species and what patterns (spatial or temporal) social media data is useful. Further, we present a novel verification method for image-based data which is suitable for large and diverse data collections.

Chapter 3

Suitability of Image-based Social Media as a Supplement to Citizen Science Portals

In this chapter, we present a large scale study exploring the potential of social media datasets to supplement species distribution data, and in doing so to serve as a form of passive citizen science. We assess the value of species distribution data gathered from the Flickr photo-sharing website relative to existing content on the public source of biodiversity data, UK National Biodiversity Network (NBN) portal. As described in Section 2.1.1, NBN Atlas[1] portal holds the most extensive collection of biodiversity information within the UK with over 220 million species occurrences. NBN datasets have previously proved useful in studying distribution patterns of UK species [Leivesley et al., 2021, Blight et al., 2009]. We focus specifically on Flickr for collecting species distribution data as it hosts one of the most extensive and easily accessible collections of geo-referenced photos of its kind and, because it is photo-based, it enables the possibility to validate observations by comparing the asserted species name, as provided in a tag or caption, with the content of the image. We conducted analyses with two case studies, one being the 1500 species that were most frequently recorded on NBN and the other being invasive species in the UK that have records on NBN. Studying invasive species can help understand if Flickr is suitable for real time iden-

[1] https://nbn.org.uk

tification of any sudden changes in the habitat of these species and thus help establish on-time protection mechanisms for the native species and for the balance of the ecosystem.

As described in Section 2.2.3, there are several limitations of previous research on using social media to augment traditional biodiversity portals in that the analyses have been performed on very small numbers of species, the methods for accessing the social media are either manual or only partly automated, and the results are limited in the degree of verification (Daume [2016], ElQadi et al. [2017], Barve [2014]). Recent work by August et al. [2020] investigates whether an image classifier for identifying plants could facilitate the discovery of unexploited biodiversity data from Flickr. However, this approach is focused purely on species occurrence on Flickr and thus does not provide a clear evaluation of the role of social sites observations compared to more traditional approaches.

In this chapter, we address these gaps by performing a large scale study evaluating Flickr as a resource of wildlife data against the NBN collection. In comparing species distributions from Flickr with those of the NBN we quantify the value of social media acquired distribution data on the largest number of species considered to date in such studies. This research addresses question **RQ 1: Can social media data serve as a useful supplement to citizen science data portals in representing the spatial and temporal distribution of bio-diversity data?** from the research questions presented in Section 1.2. More specifically contributions include extensive comparison between the image-sharing social media platform and the citizen science data portal. We have performed statistical, spatial, and temporal analysis considering different spatial and temporal settings. Furthermore, we develop a novel method of validating Flickr species images with the Google Cloud Vision API based on automatic matching of the assigned categories to the content of a hierarchical species taxonomy.

Our validation approach is similar to that used in ElQadi et al. [2017] to verify Flickr data using image content recognition with the Google Reverse Image Search. The authors use the Google Reverse Image Search in order to return labels per photo which best describe the content of the given photo. All labels per species are ordered in descending order of frequency. Then, the species-relevant tags and species-irrelevant tags are manually identified among the most frequent ones which indicate that the given photo is a true representation of the given species. This manual selection of relevant tags per species we considered an ‘exact match’ between the specific species and what would be the most relevant tags for this specific species. Despite the benefits of such approach especially when photos need to be evaluated only for a couple of species, we consider it unsuitable for larger collections where the manual upload of photos and manual selection of relevant tags per species can be a time-consuming process. Therefore, we propose a fully automated image verification approach suitable for verifying large and diverse species collections. Specifically, we deploy Google Cloud Vision API which allows fully automatic image verification. Further, we reduce the incidence of missed matches by employing a species taxonomy that supports matching between alternative names for a species as well as generic matches between terms in the relevant species hierarchy that were not used in the Flickr tags.

The structure of this chapter is as follows. Section 3.1 explains the methods used for performing the analyses and developing our image validation approach. Section 3.1.1 explains the data collection process and describes the Flickr and NBN datasets used for performing analysis. Section 3.2 presents the results while Section 3.3 discusses the findings from this chapter and Section 3.4 concludes it.

3.1 Methods

We perform three types of analysis to compare species occurrence between the NBN and Flickr, consisting of a summary statistical analysis and spatial and temporal analyses. The statistical analysis compares the frequency of occurrence of species between the two data collections, performed on different taxonomic levels of species and class.

The spatial analysis determines whether Flickr species observations match by location the NBN species observations. Because many species have variable distributions and abundances throughout the year we also use a temporal analysis to compare the time patterns of the NBN and Flickr data collections. We compared the locations of data occurrences for the two data sources for a time span of 3 months, 6 months and 12 months. We verify Flickr species identification through an image content verification approach using the Google Cloud Vision API to identify objects that appear in a given photo. The Google Cloud Vision API labels images with multiple taxonomic categories (i.e. labels) ranging from general to specific. Our image validation approach is based on coarse matching between all species names following down from the class of a species and the labels returned by Google Cloud API. In this way, we avoid a potentially high number of false negatives for less common species that are less likely to be identified on the API at the species level but might be identified at higher taxonomic levels. An outline of the methodology is depicted in Figure 3.1 and each step is detailed below.

3.1.1 Data Collection

NBN Data Collection The NBN was selected as the biodiversity data portal for our study because it holds the most extensive collection of biodiversity information within the UK. We collected the names and the number of occurrences for the top 1500 species on NBN using the NBN Atlas Occurrence Facet Search. We performed our search over all collections within the NBN and limited it for the territory of the UK.

For each of the species retrieved from the NBN we obtained, via a search on the NBN, all the alternative names associated with the species (scientific name and common names), the NBN species ID, and its taxonomic classification hierarchy. The names associated with each of the species were used for downloading data from Flickr. The taxonomic classification hierarchy is used for the verification of the Flickr data collection in combination with the Google Cloud Vision API. The NBN service does not

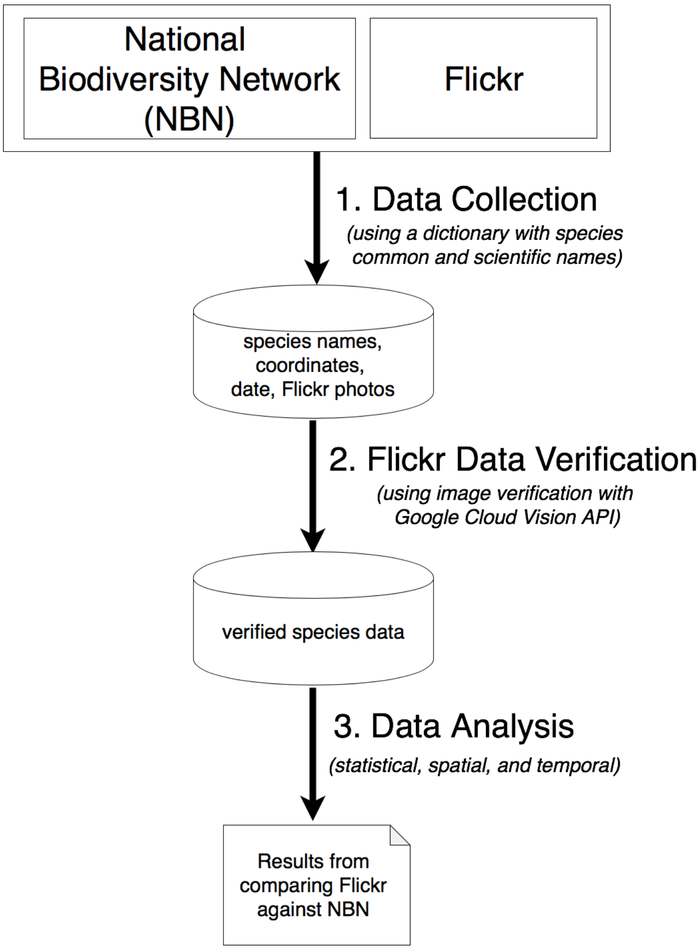

National Biodiversity Network (NBN)
Flickr
1. Data Collection
(using a dictionary with species common and scientific names)
species names, coordinates, date, Flickr photos
2. Flickr Data Verification
(using image verification with Google Cloud Vision API)
verified species data
3. Data Analysis
(statistical, spatial, and temporal)
Results from comparing Flickr against NBN

Figure 3.1: A diagram of the proposed methodology which contains three steps

support exact match between a given search term and the species name given to a record. Instead, the NBN service does partial matching between the search term and the species record names. For instance, the search term ‘brown squirrel’ can return results for other types of squirrels such as red squirrels, because there is a partial match between the search term *‘brown squirrel’* and the record name *‘red squirrel’*, i.e, the word ‘squirrel’. Thus, downloaded records can sometimes include species which are irrelevant to the search term. To resolve this problem, we remove the species records irrelevant to the search term. We perform this by automatically matching the search term to the species names given in the NBN records. If there is no match, records are considered irrelevant and thus are filtered out.

Further to that, some records are incomplete, lacking temporal or geo-information. To address this we filtered out records with missing information. For inclusion in our dataset each record constituted record ID, geo-coordinates of the occurrence, date of the occurrence, NBN species ID.

Flickr Data Collection Using the Flickr API interface we used both the scientific and common names, and limited our search to geo-tagged posts within the UK. Our search was therefore based on downloading posts with tags matching at least one of the alternative names given for a species in NBN. We downloaded the following types of information from Flickr: image coordinates, 'taken date' and 'posted date' of the post, post id, the image associated with the post, title, and all the tags associated with the post. 'Taken date' refers to the time at which the photo was taken while 'posted date' represents the time at which the photo was uploaded to Flickr. For performing the temporal analysis, we use the 'taken date'. However, early observations showed that 'posted date' and 'taken date' do not differ more than 3 months for the majority of Flickr posts. Additionally, we consider only posts where coordinates are extracted from a GPS-enabled device which is either the device used to upload the photo, or the device used to take the photo. Manual observation of the collected records shows that in most cases it is a single device used for taking and uploading the photo. We ignore posts associated only with user-provided locations.

3.1.2 Flickr Data Validation

For describing the image validation method we heavily rely on three notions, clarified in Table 3.1. Flickr images needed to be validated because the content of the photos uploaded with associated tags might not match the species name tags given by the Flickr users. However, existing image verification approaches lack the ability to scale to large collections of species and they often need manual or semi-automatic verification. Further, a common limitation of automated image verification methods is the inability to

Concept	**Description**
Flickr Tags	The tags Flickr users have given to the photos they have uploaded on Flickr
Google Cloud Vision API labels	These are the labels returned by the image recognition API for the objects recognized in each Flickr photo. We use the Google Cloud Vision API to verify whether the species on the Flickr photos represent the given species
NBN species names	The list of species names extracted from NBN species classification taxonomy

Table 3.1: Main concepts used to describe the image verification method

accurately distinguish between species with similar visual characteristics. Therefore, performing an exact match between species names and image recognition model labels could result in many possible valid photos being regarded as invalid representation of the species, which limits the coverage of methods. We aim to address this issue and provide an approach for verifying large and diverse image-related species data fully automatically by using a Bag-of-Words (BOW) approach. Specifically, we use Google Cloud Vision API to coarse match between all names following down from NBN species taxonomic class and the labels returned by Google Cloud Vision API. A potential problem of the BOW approach is that on lower levels it might not be able to distinguish between species belonging to the same class such as two different types of grass. However, we hypothesise that for diverse species collections such coarse match-based approach will help improve the coverage of automated image verification methods (recall measure) without significantly affecting their accuracy and precision. We perform an evaluation and error analysis for the BOW approach against a fully automated 'exact match' approach inspired by ElQadi et al. [2017] where matching between Google Cloud Vision API labels and species names is performed on species-level. The 'exact match approach' has the potential to be more accurate and precise. Further, we expand on the analysis by performing evaluation at the genus-level. In this way, we compare three approaches for evaluating species-related image data, i.e. exact match, class-level match, and genus-level match. This comparison allows us to verify which approach is the most suitable for verifying large collections of species.

The main steps involved in our image verification approach are illustrated in Figure 3.2. The methodology consists of the following steps: First, for each Flickr image, we download all species names from NBN following down from the species taxonomic class (for class-level match) or genus (for genus-level match). We perform image verification using the Google Cloud Vision API model and store the labels returned by the model. Then, we apply a coarse match between the NBN species names and the Google Cloud Vision API labels. If there is a match between the labels returned by the image verification model and the NBN species-related names, then the image is considered a correct representation of the species given by the Flickr user.

Google Cloud Vision API is however not trained on wildlife data and thus some of the less well-known species names might not be returned as labels, for instance, 10-spot Ladybird (*Adalia decempunctata*) and 22-spot Ladybird (*Psyllobora vigintiduopunctata*). Also, species belonging to the same class (e.g. 'cuckoo' and 'sparrowhawk') might have very similar visual appearance and thus the Google Cloud Vision API cannot be assumed to distinguish between the two species. Therefore, using exact matching between species names and Google labels will lead to a high number of false negatives.

An example of a Google Cloud Vision API result for a single photo correctly tagged on Flickr as Adder, gives the following categories: *Reptile (98%), Snake (98%), Scaled reptile (93%), Viper (91%), Serpent (89%), Terrestrial Animal (87%), Rattle Snake (84%), Sidewinder (70%), Adaptation (67%), Colubridae (65%), Eastern Diamondback Rattlesnake (56%), Elapidae (53%).* The higher the score, the more confident the assignment of the category is for the given image, where the score is given in brackets next to the tags.

The photo labels returned by Google Cloud Vision API can be organised as a taxonomy

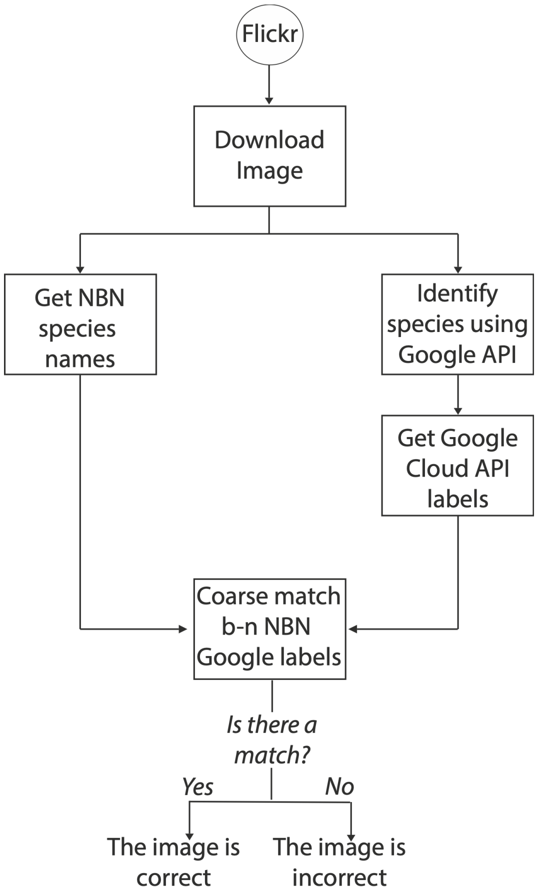


Figure 3.2: Image Validation Approach Overview

that matches the species taxonomy returned by NBN.

Figure 3.3 displays the NBN classification for Adder and the labels returned by Google Cloud Vision API for this photo. We use the NBN taxonomic classification for the species to choose relevant species names to match the labels returned by Google Cloud Vision API. A BOW approach is adopted where we treat the names in the classification hierarchy for a species as a list of names ignoring the hierarchical and semantic relations between these names. We consider all names in the classification hierarchy following down from class, and we match these terms to the labels given by Google Cloud API. In the example, given in Figure 3.3, the use of the classification finds an exact match between the species name ‘Viper’ (an alternative name for ‘Adder’) and the Google Cloud Vision API term ‘Viper’. Using exact matching on the Flickr label of

Adder would not have found any match, resulting in a false negative for this observation. Another example is for species *Phleum pratense (Timothy Grass)*, which is from class Magnoliopsida and family *Poacae (Grass)*. Google Cloud Vision API returns for images with this species the label ‘grass’, rather than ‘timothy grass’ and thus coarse match would be successful in this case.

Note that we use both scientific and common names for matching, as both can occur within the NBN derived taxonomy and the labels returned by the Google Cloud Vision API. We performed manual verification of the Google results for 50 randomly selected species where we randomly select 40 images per species. We used these 2000 images to evaluate the performance of our image verification method.

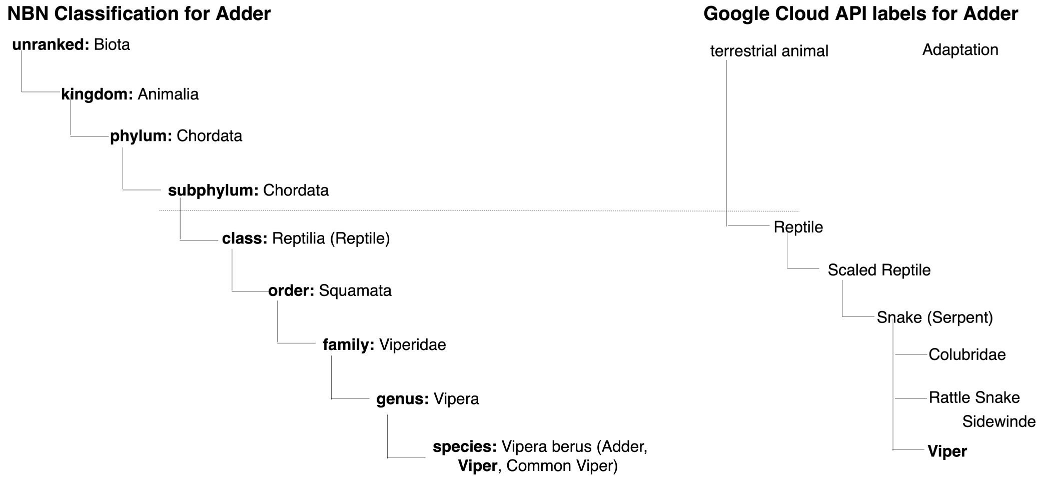


Figure 3.3: Google Cloud Vision API label taxonomy and NBN classification for Adder.

3.1.3 Data Analysis

The data analyses are based on two case studies: the 1500 most frequently recorded species on NBN and the invasive species in the UK that appear in both data collections. Spatial comparison between the NBN and Flickr datasets was performed using spatial grid modelling, in which geographic space is divided into regular grid cells. The cells were classified according to whether they contained observations from one or other or both of the two sources. The classification was further refined according to time windows to support both a spatial and a temporal analysis. By varying grid cell sizes, and cell aggregation (i.e. one by one vs three by three), as well as the time window, we performed a number of scale-variant spatio-temporal analyses.

There were two main methods of performing spatial analysis:

1. One by one cell comparison: We compare Flickr and NBN species occurrence data per 10km, 20km, and 40km grid square. We performed experiments with these cell sizes because we are interested in identifying the most fine-grained level for which there is a high number of matches between species observation on Flickr and NBN without affecting the precision of the method. We regarded cell sizes beyond this size of 40km as being of more limited value for studying species distribution and migration We calculate a confusion matrix, which is used to describe the performance of a classifier on a test data set for which the true values are known, where Flickr is the test data set and NBN the ground truth values. The cells of the confusion matrix are defined as follows:

    - ‘True Positive’ (TP): a cell has both NBN and Flickr data points for the species
    - ‘True Negative’ (TN): a cell does not have occurrences from either of the sources
    - ‘False Negative’ (FN): a cell has no Flickr data for the species, but it does have NBN data for the species
    - ‘False Positive’ (FP): a cell has Flickr data for the species but no NBN data

2. Three by three cell comparison: We compare Flickr and NBN using a three by three analysis centred on every cell. In this approach, we count a true positive if there is a Flickr posting in a cell and if there are NBN records within either the cell itself or in any of the adjacent eight cells. A false negative would be

declared if a set of nine cells had at least one NBN record but no Flickr record. A false positive indicates if there is a Flickr posting in a cell, but there are no NBN records within either the cell itself or in any of the adjacent eight cells. A 'True Negative' would be no Flickr postings and no NBN records in any of the nine cells.

Based on the measures above we compute precision, recall, and F1-measure.

We look at temporal accuracy of Flickr on seasonal (3 months), half yearly (6 months) and yearly patterns (12 months). The 3 and 6 month-based analysis are performed ignoring the year. This allows identification of seasonal patterns that are usually unaffected by yearly changes such as seasonal migrations.

3.2 Results

3.2.1 Statistical Analysis

NBN and Flickr Datasets Comparison

Across the 1500 most numerous species on NBN Atlas, 90% were found on Flickr and 100% of species in the Flickr dataset were found on NBN Atlas. The NBN Atlas records, as expected, far outnumber those on Flickr, being 93,656,179 and 791,059 respectively. It is worth noting that NBN data used here covers the entire collection period; 1800-2018 while Flickr data covers only 2006-2018. It was found that 35% of the species counted on Flickr have more than 100 occurrences. Table 3.2 lists the top 10 most frequently recorded species on Flickr (mostly with more than 10000 occurrences).

The best represented species on Flickr (see Table 3.2,can be split into three main categories: pretty, i.e. photogenic, flowers (Bluebell, Daisy, Dandelion), sessile green plant species (Ivy, Beech, Bracken) and garden and aquatic birds, which are also diurnal (Continental Robin, Mallard). Notably all are easily accessible. These same patterns were mirrored at the class level with the highest number of returns for Flickr being *Magnoliopsida*, a class of flowering plants, and the second highest was *Aves*. *Phleum pratense (Timothy Grass)* as a well documented species in Flickr is an interesting observation as, compared to the other commonly observed species (see Table 3.2), it is not a well known species that is readily identified, suggesting that it was incidental in many images and Flickr may be good at picking up species that appear as a background in a photo. Another example of such a species is *Hedera helix (Ivy)*. NBN and Flickr datasets are similar in the diversity of classes they represent with the ten best represented classes in both collections being the same , with the same three most common classes of *Insecta* (Insects), *Magnoliopsida* (plants), and *Aves* (birds). Both data collections are representing species from a small number of classes. This suggests that the same observer bias in photos also occurs in NBN data collections.

Scientific name	**Common name**	**Flickr count**	**NBN count**
Hyacinthoides non-scripta	*Bluebell*	20,940	54,893
Bellis perennis	*Daisy*	20,656	28,748
Erithacus rubecula	*Continental Robin*	19,248	3,938,616
Morus bassanus	*Gannet*	17985	14252
Fagus sylvatica	*Beech*	15,842	24,973
Hedera helix	*Ivy*	14,474	27,211
Anas platyrhynchos	*Mallard*	13,500	834,039
Taraxacum officinale agg.	*Dandelion*	13,443	27,269
Pteridium aquilinum	*Bracken*	12,708	30,741
Phleum pratense	*Timothy Grass*	9,000	11,903

Table 3.2: The top 10 species on Flickr with the highest number of records

The top 10 species on NBN are garden birds (see Table 3.3), and they are represented well in the Flickr dataset with occurrences in most cases above a thousand.

Scientific name	**Common name**	**NBN count**	**Flickr count**
Turdus merula	*Blackbird*	4,609,821	3,234
Cyanistes caeruleus	*Blue Tit*	4,164,338	3,491
Erithacus rubecula	*Continental Robin*	3,938,616	19,248
Columba palumbus	*Woodpigeon*	3,584,436	1,660
Prunella modularis	*Dunnock*	3,513,651	2,179
Parus major	*Great Tit*	3,507,350	2,670
Fringilla coelebs	*Chaffinch*	3,444,776	3,474
Passer domesticus	*House Sparrow*	3,184,175	2,312
Streptopelia decaocto	*Collared Dove*	3,094,475	929
Chloris chloris	*Greenfinch*	2,900,214	2,030

Table 3.3: The top 10 species on NBN with the highest number of records

NBN and Flickr Datasets Comparison for Invasive Species in the UK

There are 82 invasive species for UK that also have occurrence records on NBN. The total count of records of invasive species on NBN is 1,485,744. The total number of Flickr posts for the invasive species that are also recorded on NBN is 27,187. The number of species with occurrences above 100 for both NBN and Flickr data collections is 19 (of 82), which is 23% of the number of invasive species on NBN. The invasive species with more than 100 occurrences for both NBN and Flickr are diurnal mammals, birds (more than 50%) along with a few ‘pretty’ flower species (see Table 3.4).

The best represented invasive species on Flickr are *Branta canadensis (Canada Goose)*, *Scirurus carolinensis (Grey Squirrel)*, *Gallinago gallinago (Snipe)*, *Oryctolagus cuniculus (Rabbit)*, *Rhododenron ponticum (Rhododendron)*, *Aix galericulata (Mandarin Duck)*, and *Cygnus atratus (Black Swan)*. The species for which NBN and Flickr have a similar number of records are *Sus scrofa (Wild boar)* and *Bubo bubo (Eurasian Eagle*

Owl).

Scientific name	**Common name**	**NBN species count**	**Flickr species count**
Branta canadensis	*Canada Goose*	377,111	3,328
Sciurus carolinensis	*Grey squirrel*	350,113	3,249
Gallinago gallinago	*Snipe*	325,210	1,619
Oryctolagus cuniculus	*Rabbit*	96,093	7,994
Alopochen aegyptiacus	*Egyptian Goose*	31,591	862
Rhododenron ponticum	*Rhododendron*	30,803	3,489
Branta leucopsis	*Barnacle Goose*	24,269	289
Aix galericulata	*Mandarin Duck*	19,693	1,500
Muntiacus reevesi	*Reeve's muntjac*	16,428	489
Cygnus atratus	*Black Swan*	8,761	1,148
Buddleja davidii	*Buddleia*	5,654	443
Heracleum mantegazzianum	*Giant Hogweed*	5,348	190
Anser caerulescens	*Snow Goose*	5,085	177
Anser indicus	*Bar-Headed Goose*	3475	164
Aix sponsa	*Wood Duck*	2,688	290
Cervus nippon	*Sika Deer*	2,442	226
Chrysolophus pictus	*Golden Pheasant*	1,745	167
Sus scrofa	*Wild boar*	441	373
Bubo bubo	*Eurasian Eagle Owl*	395	537

Table 3.4: Species occurrences for NBN and Flickr for invasive species with number of occurrences above 100.

3.2.2 Flickr Data Verification

In initial exploratory work, we performed tests with the tags returned by the Google Cloud Vision API. We found that the tags with a score above 60% are more likely to imply the correct species displayed on the photos. The tags with a score lower than 60% usually describe either less relevant objects of the photo, e.g. parts of the background (*'leaf'*), characteristics of the animal ('fawn'), or are names of species that are irrelevant to the photo ('Diamondback Rattlesnake' when the species is Adder). Therefore, we used only tags with a score higher than 60%.

In the rest of the section, we evaluate our image verification approach (class level BOW approach) against two other image verification approaches. The first one is a fully automated 'exact match' approach, inspired by ElQadi et al. [2017] where we perform an exact match between the NBN species names associated with a given species and the Flickr posts names. The second approach is a genus level BOW approach where we consider names following down from the genus of the species. Evaluating approaches considering different levels of coarse match (species-level, genus-level, and class -level) allows us to judge which approach is more suitable for evaluating diverse collections of species without affecting the precision.

The average results, presented in Table 3.5, show that class-level BOW approach outperforms the other two evaluation approaches by a significant margin with F1 = 0.79 versus F1 (baseline) = 0.20 and F1 (BOW (genus level)) = 0.27. Specifically, the class-level BOW method achieves best results, compared to the other approaches, for 45 out of the 50 species. For 14 of these species, the class-level BOW method has equal or slightly lower precision than the other two methods, however the recall is much higher in these cases leading to a better performance overall. For instance, for 'Passer domesticus (House Sparrow)', the baseline and BOW (genus level) approach have a higher precision of 1.0 while BOW (class level) has a precision of 0.97, however the recall (0.85) is more than double compared to the recall of the other approaches. This shows that class-level BOW method is a more suitable for evaluating wider range of species than the other two approaches without significantly affecting the precision.

However, the presence of some species for which none of the approaches were successful, i.e., 'Erica cinerea (Bell Heather)', 'Stachys officinalis (Betony)', 'Solanum dulcamara (Bittersweet)', 'Hyacinthoides non-scripta (Bluebell)', and 'Acer campestre (Field Maple)', shows that there are species for which BOW approach is unsuitable and other methods need to be considered such as the inclusion of species characteristics (color, shape), attributes (feather, beak, etc.) combined with a higher taxonomy level for the species (i.e. family).

The most common causes of false positives for BOW are photos that include an artificial representation of a species, such as a boat with a figure of a goose (Figure 3.4), and hence do not represent a living species. Common cases of false negatives for BOW are photos which include the species but the focus of display is another object. In the example given in Figure 3.4, the main object in the photo is a building, and thus Google Cloud Vision API returns labels associated with the building and the characteristics of the building, rather than the plant (i.e. Hedera helix (Ivy)).

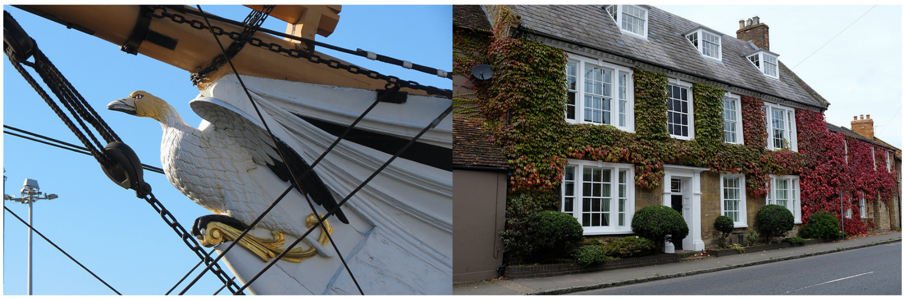


Figure 3.4: Common cases of false positive and false negative: False Positive for *Marus bassanus (Gannet)*(left) tags: bird, goose, vehicle, tall ship and False Negative for *Hedera helix (Ivy)* (right): tags: property, house, home, building, residential area, cottage, real estate, neighbourhood.

species	**baseline**			**BOW (genus level)**			**BOW (class level)**		
	p	r	F1	p	r	F1	p	r	F1
Coccinella septempunctata (7-spot Ladybird)	0.00	0.00	0.00	0.00	0.00	0.00	**0.97**	**0.95**	**0.96**
Propylea quattuordecimpunctata (14-spot Ladybird)	0.00	0.00	0.00	0.00	0.00	0.00	**1.00**	**1.00**	**1.00**
Vipera berus (Adder)	**1.00**	0.65	0.79	**1.00**	0.65	0.79	**1.00**	**0.76**	**0.87**
Tyto alba (Barn Owl)	0.00	0.00	0.00	0.00	0.00	0.00	**0.89**	**0.50**	**0.64**
Ophrys apifera (Bee Orchid)	0.00	0.00	0.00	**1.00**	0.82	**0.90**	**1.00**	**0.83**	**0.90**
Erica cinerea (Bell Heather)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Stachys officinalis (Betony)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Solanum dulcamara (Bittersweet)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Turdus merula (Blackbird)	0.96	0.76	0.85	0.95	0.75	0.84	**0.91**	**1.00**	**0.95**
Hygrocybe conica (Blackening Waxcap)	0.00	0.00	0.00	0.00	0.00	0.00	**1.00**	**0.60**	**0.75**
Cyanistes caeruleus (Blue Tit)	0.00	0.00	0.00	0.00	0.00	0.00	**0.97**	**0.89**	**0.93**
Hyacinthoides non-scripta (Bluebell)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Buteo buteo (Buzzard)	**1.00**	0.52	0.68	**1.00**	0.52	0.68	**1.00**	**0.81**	**0.89**
Corvus corone (Carrion Crow)	0.00	0.00	0.00	0.00	0.00	0.00	**1.00**	**0.95**	**0.97**
Fringilla coelebs (Chaffinch)	0.00	0.00	0.00	0.00	0.00	0.00	**1.00**	**0.89**	**0.95**
Periparus ater (Coal Tit)	0.00	0.00	0.00	0.00	0.00	0.00	**0.97**	**0.92**	**0.95**
Streptopelia decaocto (Collared Dove)	0.00	0.00	0.00	0.00	0.00	0.00	**0.92**	**0.92**	**0.92**
Bombus pascuorum (Common Carder Bee)	0.00	0.00	0.00	**1.00**	0.75	0.85	**1.00**	**0.87**	**0.93**
Zootoca vivipara (Common Lizard)	0.00	0.00	0.00	0.00	0.00	0.00	**1.00**	**0.93**	**0.96**
Erithacus rubecula (Continental Robin)	**1.00**	0.82	0.90	**1.00**	0.82	0.90	**1.00**	**0.98**	**0.99**
Anthriscus sylvestris (Cow Parsley)	**1.00**	0.27	0.43	**1.00**	0.27	0.43	**1.00**	**0.28**	**0.44**
Prunella modularis (Dunnock)	0.00	0.00	0.00	0.00	0.00	0.00	**0.95**	**1.00**	**0.97**
Acer campestre (Field Maple)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Carduelis carduelis (Goldfinch)	**1.00**	0.27	0.43	**1.00**	0.27	0.43	0.92	**0.92**	**0.92**
Dendrocopos major (Great Spotted Woodpecker)	0.00	0.00	0.00	0.00	0.00	0.00	**1.00**	**0.97**	**0.99**
Parus major(Great Tit)	0.00	0.00	0.00	0.00	0.00	0.00	**0.97**	**0.97**	**0.97**
Chloris chloris (Greenfinch)	0.00	0.00	0.00	0.00	0.00	0.00	**0.95**	**1.00**	**0.97**
Passer domesticus (House Sparrow)	**1.00**	0.35	0.52	**1.00**	0.43	0.60	0.97	**0.85**	**0.90**
Corvus monedula (Jackdaw)	0.00	0.00	0.00	0.00	0.00	0.00	**1.00**	**0.94**	**0.97**
Pyrrhosoma nymphula (Large Red Damselfly)	0.00	0.00	0.00	0.00	0.00	0.00	**1.00**	**0.95**	**0.97**
Aegithalos caudatus (Long-Tailed Tit)	0.00	0.00	0.00	0.00	0.00	0.00	**1.00**	**0.95**	**0.97**
Pica pica (Magpie)	1.00	0.44	0.61	1.00	0.44	0.61	**0.88**	**0.85**	**0.87**
Phoxinus phoxinus (Minnow)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Lutra lutra (Otter)	1.00	0.84	0.91	1.00	0.84	0.91	**0.97**	**0.90**	**0.93**
Boloria euphrosyne (Pearl Bordered Fritillary)	0.00	0.00	0.00	0.00	0.00	0.00	**1.00**	**1.00**	**1.00**
Alca torda (Razorbill)	0.00	0.00	0.00	0.00	0.00	0.00	**0.93**	**0.76**	**0.84**
Riparia riparia (Sand Martin)	0.00	0.00	0.00	0.00	0.00	0.00	**0.88**	**0.69**	**0.78**
Argynnis paphia (Silver-Washed Fritillary)	0.00	0.00	0.00	0.00	0.00	0.00	**1.00**	**1.00**	**1.00**
Turdus philomelos (Song Thrush)	0.00	0.00	0.00	1.00	0.02	0.05	**0.94**	**0.94**	**0.94**
Sturnus vulgaris (Starling)	**1.00**	0.08	0.15	**1.00**	0.08	0.15	0.97	**0.86**	**0.91**
Passer montanus (Tree Sparrow)	0.00	0.00	0.00	0.96	0.71	0.82	**0.97**	**0.92**	**0.95**
Bombus lucorum (White-Tailed Bumble Bee)	0.00	0.00	0.00	**1.00**	0.79	0.88	**1.00**	**0.89**	**0.95**
Columba palumbus (Woodpigeon)	0.00	0.00	0.00	0.00	0.00	0.00	**0.90**	**0.91**	**0.94**
Troglodytes troglodytes (Wren)	**1.00**	0.78	0.88	**1.00**	0.78	0.88	**1.00**	**0.95**	**0.97**
Hedera helix (Ivy)	**1.00**	**0.18**	**0.31**	**1.00**	**0.18**	**0.31**	**1.00**	**0.18**	**0.31**
Sciurus carolinensis (Grey Squirrel)	**1.00**	0.81	0.89	**1.00**	0.83	0.91	**1.00**	**0.92**	**0.96**
Amanita rubescens (Blusher)	0.00	0.00	0.00	0.00	0.00	0.00	**1.00**	**0.94**	**0.97**
Cygnus atratus (Black Swan)	**1.00**	0.77	0.87	0.96	0.84	0.89	0.90	**0.90**	**0.90**
Morus bassanus (Gannet)	**1.00**	0.69	0.82	**1.00**	0.69	0.82	0.97	**0.94**	**0.95**
Branta leucopsis (Barnacle Goose)	0.00	0.00	0.00	0.00	0.00	0.00	**1.00**	**1.00**	**1.00**
AVERAGE	0.29	0.16	0.20	0.39	0.23	0.27	**0.86**	**0.76**	**0.79**

Table 3.5: Comparison between class-level and genus-level BOW image verification approaches and the baseline approach, based on exact-match approach.

3.2.3 Spatial and Temporal Analysis

The Top 1500 Species on NBN

The average precision and recall across all species for each type of spatial and temporal constraint for one by one grid cell analysis is 0.38 (38%) for precision and 0.20 (20%) for recall. The recall score shows that on average 20% of all NBN data was also reflected by the Flickr data. The precision score shows that on average 38% of the Flickr cell-based identifications of a species were also reflected in the NBN data (see Figure 3.5).

The average precision and recall across all species for each type of spatial and temporal constraint for three by three analysis is 0.60 (60%) for precision and 0.10 (10%) for recall (see Figure 3.5). In comparison to one by one analysis, the average precision for three by three analysis is higher ranging from 0.27 (27%) to 0.78 (78%) for the different cell sizes while recall tends to be lower and does not vary much for the different cell sizes. The number of false negatives is significantly higher for analysis performed using a grid of size 3 by 3 and thus the recall value is lower. The reason for this can be attributed to the wider range of species recorded within the NBN in comparison to the Flickr records. Furthermore, according to the conditions for three by three analysis comparisons, false negatives occur when a set of nine cells have at least one NBN record in the absence of any Flickr record for the given species. Therefore, for species where the number of NBN records is high and the number of Flickr records is low it is very likely that cells with no Flickr occurrences will be associated with cells containing NBN records (note however that for species that are well represented on Flickr this is less likely to be the case).

The highest precision and recall scores across both types of analysis are achieved for experiments performed with cell size 40km and no temporal constraints. The lowest results are achieved for experiments performed with a time constraint of twelve months which we attribute to lack of data on Flickr.

Precision tends to be higher than recall. This higher precision reflects the fact that most locations with Flickr occurrences also contain NBN occurrences. The low recall indicates that there are many locations with NBN observations but with no Flickr observations, leading as indicated above to false negatives. However, the recall value increases significantly as the cell size increases. Also, precision increases for bigger cell sizes for the converse reason of taking account of NBN occurrences over a wider region relative to a Flickr observation. This indicates that a grid split, consisting of 40km cells provides a better balance between precision and recall measures and thus can be regarded as more suitable for validating social network observations.

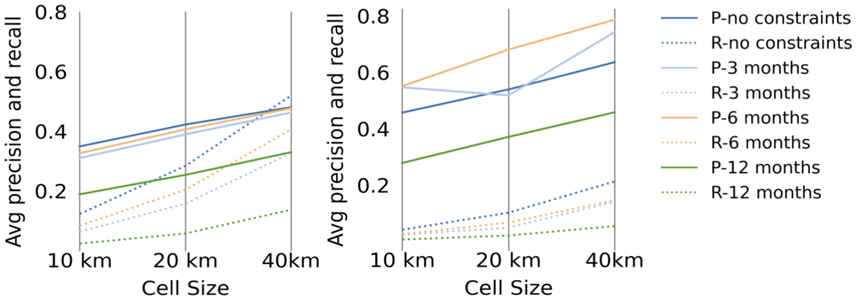


Figure 3.5: Average Precision and Recall comparison per cell size and temporal restriction: One by one analysis (left), Three by three analysis (right) where 'P' refers to precision and 'R' refer to recall. In the figure 'no constraints' refers to the analyses performed with no temporal constraints.

We calculated the best, worst and average F1 performance (see Figure 3.6), where best and worst were based on the average F1 scores for individual species for a particular cell size, while average F1 performance was across all species for the particular cell size. The average F1 measurement does not exceed 0.50. Best performing species have poorer F1 scores for cell sizes 10km and 20km and F1 score of 0.70 for the analysis performed on cell size 40km. F1-measure on average is higher when the analysis is performed with no temporal constraints. Further, the average F1 scores for the analysis conducted using a 12-month window are the lowest.

The results (see Figure 3.5 and Figure 3.6) show that the Flickr dataset best reflects

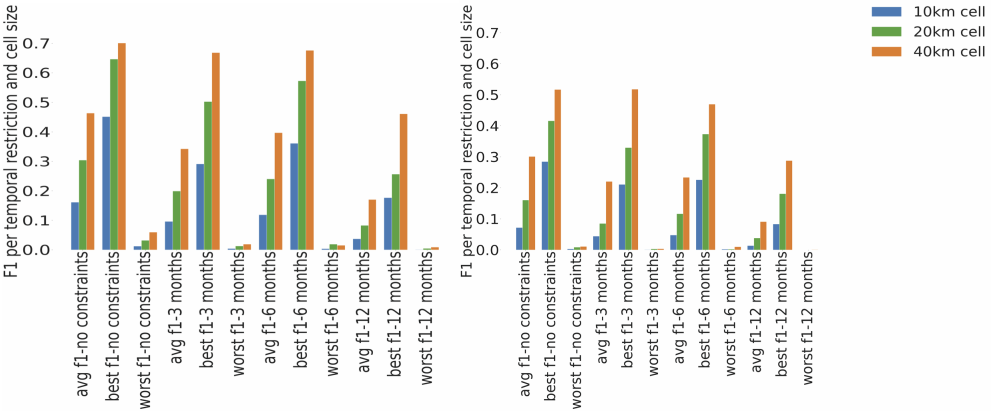


Figure 3.6: Comparison of average, best, worst F1 measure values per temporal restriction and cell size: One by One analysis (left),Three by three analysis(right).

the NBN dataset on a purely spatial analysis with no time constraints. The comparison with a constraint that observations are within 12 months of each other gives the lowest results on all measures.

As indicated above, the overall comparison between the two datasets is notable for the highly unbalanced precision and recall scores. As these scores are averaged across all considered species, we investigated those species with precision and recall both being above 0.50, and we found 134 distinct such species. We found the average F1 score for the top 10 of these species with a 40km grid size to be 0.68 (see Table 3.6). As before the best results were obtained with no temporal constraints, though with a couple of exceptions for a 6 month temporal window. The best represented species on Flickr in comparison to NBN as represented in Table 3.6 are, with one exception, birds, most but not all of which are diurnal.

Species name	Analysis type	Cell size	Precision	Recall	F1-measure
Thymelicus sylvestris (Small Skipper)	no constraints	40	0.64	0.77	0.70
Strix aluco (Tawny Owl)	no constraints	40	0.65	0.76	0.70
Sitta europaea (Nuthatch)	no constraints	40	0.6	0.82	0.69
Primula veris (Cowslip)	no constraints	40	0.61	0.79	0.69
Aegithalos caudatus (Long-Tailed Tit)	no constraints	40	0.56	0.88	0.68
Botaurus stellaris(Bittern)	no constraints	40	0.61	0.76	0.68
Libellula depressa (Broad-Bodied Chaser)	no constraints	40	0.63	0.73	0.68
Sitta europaea (Nuthatch)	6 months	40	0.62	0.74	0.68
Aegithalos caudatus (Long-Tailed Tit)	6 months	40	0.59	0.78	0.67
Certhia familiaris (Treecreeper)	no constraints	40	0.56	0.83	0.67

Table 3.6: The top ten results with the highest F1-measure across all species

Invasive Species for UK

The average results for the invasive species demonstrate the same spatial and temporal patterns as the average results for the top 1500 species, presented in the previous section, i.e. best performance is for spatial analysis performed with 40 km grid cell size with no time constraints (Figures 3.7 and 3.8).

The average precision and recall across all species for each type of spatial and temporal constraint for one by one analysis is 0.40 (40%) for precision and 0.20 (20%) for recall (see Figure 3.7). The average precision and recall across all species for each type of spatial and temporal constraint for three by three analysis is 0.60 (60%) for precision and 0.10 (1%) for recall (see Figure 3.7).

The best represented invasive species on Flickr in comparison to NBN, with precision and recall both being above 0.50, are given in Table 3.7. Results are promising for these species as the F1-measure is on average 61.2%, specifically for representing spatial patterns on 40km cell size with no time constraints (see Table 3.7). There are six distinct species with the best performance among the invasive species (note that in Table 3.7 some species have multiple rows with different conditions of analysis). The species - *Branta canadensis (Canada Goose)* and *Sciurus carolinensis (Grey squirrel)*

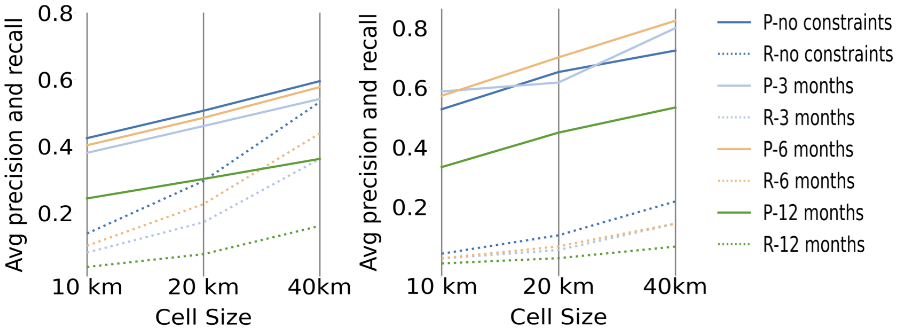


Figure 3.7: Average Precision and Recall comparison per cell size and temporal restriction for invasive species: One by one analysis on the left, three by three analysis on the right, where 'P' refers to precision and 'R' refers to recall. 'no constraints' refers to analyses performed with no temporal constraints.

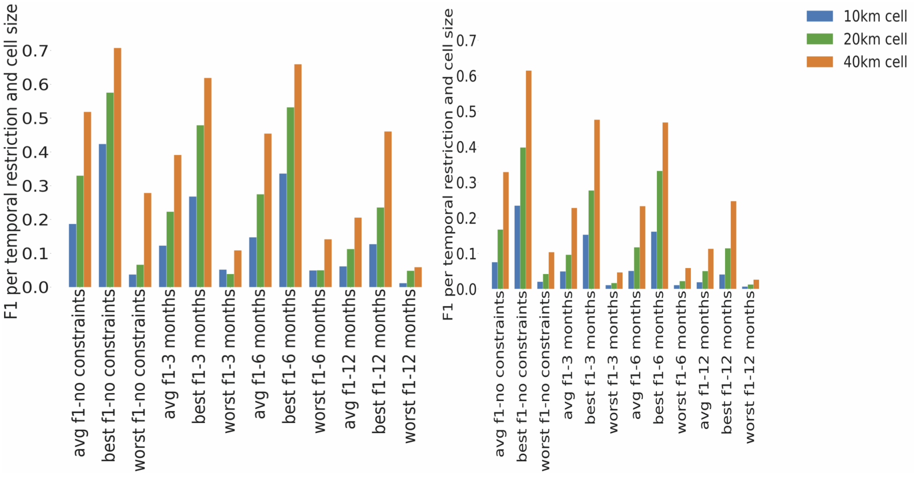


Figure 3.8: Comparison of average, best, worst F1 measure values per temporal restriction and cell size for invasive species: The one by one analysis is on the left, the three by three analysis on the right.

appear across the multiple categories of no temporal constraints, three months constraints and six months constraints. They are the best performing species in terms of having both precision and recall above 0.50 for multiple spatial and temporal restrictions and are the only species which have both precision and recall above 0.50 for cell size 20km. The best performance in terms of highest precision and highest F1 measure has been achieved for *Bubo bubo (Eurasian Eagle Owl)* with F1 = 0.71 and precision =

0.64 (with recall 0.79). These results are achieved with 40km cell size and no temporal constraints. Similarly to the results for all 1500 species, the best results for the invasive species are achieved for purely spatial analysis with cell size 40km and diurnal bird species. An exception is the mammal *Sciurus carolinensis (Grey squirrel).*

Species name	Analysis type	Cell size	Precision	Recall	F1-measure
Branta canadensis (Canada Goose)	no constraints	40km	0.55	0.80	0.65
Branta canadensis (Canada Goose)	no constraints	20km	0.55	0.53	0.54
Cygnus atratus (Black Swan)	no constraints	40km	0.59	0.79	0.68
Sciurus carolinensis (Grey squirrel)	no constraints	20km	0.57	0.58	0.58
Sciurus carolinensis (Grey squirrel)	no constraints	40km	0.59	0.80	0.68
Buddleja davidii (Buddleia)	no constraints	40km	0.51	0.51	0.51
Bubo bubo (Eurasian Eagle Owl)	no constraints	40km	0.64	0.79	0.71
Aix galericulata (Mandarin Duck)	no constraints	40km	0.51	0.60	0.55
Cygnus atratus (Black Swan)	3 months	40km	0.55	0.54	0.55
Branta canadensis (Canada Goose)	3 months	40km	0.62	0.62	0.62
Sciurus carolinensis (Grey squirrel)	3 months	40km	0.59	0.63	0.62
Cygnus atratus (Black Swan)	6 months	40km	0.59	0.71	0.65
Branta canadensis (Canada Goose)	6 months	40km	0.59	0.69	0.64
Sciurus carolinensis (Grey squirrel)	6 months	40km	0.60	0.73	0.66
Bubo bubo (Eurasian Eagle Owl)	6 months	40km	0.55	0.75	0.64
Aix galericulata (Mandarin Duck)	6 months	40km	0.54	0.50	0.52

Table 3.7: Results for the invasive species where precision and recall are both above 0.5.

3.3 Discussion

3.3.1 Data Analysis and Image Verification Approach

We conducted analyses using two case studies, one being the 1500 species that were most frequently recorded on NBN and the other being invasive species in the UK that have records on NBN. Further, we performed three types of analysis:

1. Statistical analysis — We focus on comparing the frequency of occurrence of species between the two data collections, performed on different taxonomic levels of species and class. This helps identify trends in the type of species that are best/worst represented on Flickr compared to NBN collection.

2. Spatial analysis — This analysis determines whether the Flickr and NBN species observations match by location. We perform analysis using a grid-based approach where we conducted experiments using 10 km, 20 km, and 40 km cell size.

3. Temporal Analysis — We compare the time patterns of the NBN and Flickr data collections using time spans of 3 months, 6 months, and 12 months. This helps identify how well Flickr dataset represent the seasonal, half yearly, and yearly patterns for the species on NBN.

The large data collection and extensive analysis make the research presented in this chapter the most extensive work conducted to date on the suitability of social media platforms to supplement citizen science data collections, to the best of our knowledge. Related research [August et al., 2020, Daume, 2016, ElQadi et al., 2017, Barve, 2014] is limited in scale and type of analysis performed.

We also proposed a fully automated image verification approach suitable for verifying large and diverse species collections. We used Google Cloud Vision API which allows fully automatic image verification. The approach is based on coarse matching between all species names following down from the class of a species and the labels returned by Google Cloud API. In this way, we avoid a potentially high number of false negatives for less common species that are less likely to be identified on the API at the species level but might be identified at higher taxonomic levels. We evaluated our approach using 50 randomly chosen species, each associated with 40 images which gives us in total 2000 images. Additionally, we compared the proposed approach (i.e., class-level BOW approach) against two other image verification approaches, i.e., species-level BOW and genus-level BOW. This helps identify which method is more suitable for evaluating diverse collections of species without affecting the precision. Previous image verification methods are suitable for smaller species collections, and usually involve manual or semi-automatic verification [Daume, 2016, ElQadi et al., 2017, Barve, 2014], or require large amounts of labelled images for training image classifiers [Skreta et al., 2020].

3.3.2 Findings

Below we present a summary of main findings from this chapter:

1. Of the top 1500 most numerous species on NBN 90% were also found on Flickr, confirming that social media data can represent a wide range of species. An overall comparison between the NBN and Flickr datasets indicates that they mostly represent the same classes of species. The best represented classes in both collections are the same with the top three being Insecta (Insects), Magnoliopsida (Plant class), and Aves (birds). Flickr has a good representation of flowering plants and garden and sea birds. Many Flickr uploads represent species that look attractive on photos and are easier to photograph (i.e. they are diurnal, and/or are sessile) as well as being relatively common species. Examples of well represented species on Flickr include invasive species such as *Sciurus carolinensis (Grey squirrel)* and *Branta canadensis (Canada Goose)*, and the birds *Thymelicus sylvestris (Small Skipper)*, *Strix aluco (Tawny Owl)*, *Sitta europaea (Nuthatch).*

2. Our image verification approach proved to work well on a large collection of species. The approach by ElQadi et al. [2017] of exact match between the Google tags and species names may work for a small collection of well-known species for which the Google species labels tend to be more reliable, but not for a more extensive collection, including less well-known species, for which the Google label is liable to be more generic (i.e. providing the class or genus rather than

the actual species name). In our approach, we use the taxonomy structure of the species to select relevant tags. Thus we verify images as genuine wildlife by matching the provided Flickr species name against the Google-provided class or the genus of the image content and all the tags lower down the classification hierarchy.

3. The spatial and temporal analyses for both case studies show that the Flickr dataset reflects the NBN dataset patterns best for experiments performed with cell size 40 km with no temporal constraints. The poorer results from the analysis performed with temporal constraints suggest that the Flickr dataset does not represent the temporal patterns for the species on NBN well. This is especially true for the yearly comparison between the two datasets (i.e. 12 month window).

4. The results of the precision calculations showed that there are 93 species for which precision is higher than 60%, for cell sizes 20km and 40km. This observation suggests that Flickr posts do present a potentially useful source of wildlife observations. However, the low recall value indicates that the Flickr data collection is less able to represent the full range of wildlife species in comparison to NBN. This is emphasised in the three by three analysis that gives the highest precision values, but provides the poorest recall. It should be remarked here that our scores for precision depend upon the quality of the NBN ground data, a dataset collected through citizen science campaigns by non-professionals (as discussed in 'Introduction'). Therefore, it is quite possible that some of the Flickr observations classed here as false positive could actually be correct due to the absence of existing citizen science observations at the respective location.

3.3.3 Limitations

A main limitation of this research is the lack of analysis on diverse set of social media platforms. It will be beneficial to conduct a similar larger scale study of the potential (beyond the relatively limited studies conducted to date) of other social networks such as Twitter to determine whether they can also supplement traditional biodiversity data sources.

Another problem that has not been addressed in this work is that collecting social network data on a larger scale is a challenging task because most of the networks have restrictions on data access with thresholds on the amount of data that can be downloaded. A solution to this might be to look at how data from multiple social network sources can be combined for extracting wildlife data. It is also a strong motivation to apply and if possible improve upon methods for geocoding the many accessible social media posts that do not have GPS coordinates [Stock, 2018a] (we address this problem in Chapter 5).

The image verification method can also be improved by looking at using a combination of inclusive and exclusive tags (i.e. tags used to consider a photo irrelevant) and through the development of more sophisticated computer vision methods for automated identification of individual species.

Despite these limitations, our analysis was conclusive in that Flick can serve as a source of wildlife observational data for some species. Further, the image verification approach proved suitable for validating large and diverse wildlife-related collections of images.

3.4 Conclusions

This chapter presented a large scale study exploring the potential of social media data to supplement citizen science datasets. In particular, we evaluated species distributions on Flickr relative to those submitted to the largest citizen science portal for the UK, the National Biodiversity Network (NBN) Atlas. Our study included the 1500 best represented species on NBN, and common invasive species within UK.

We performed three types of analysis comparing the statistical, spatial, and temporal distribution of species on Flickr compared to NBN. Question **RQ 2** from the initial hypothesis presented in Section 1.2 has been answered in order to show that social network related data could offer a rich source of observation data for certain taxonomic groups, and/or as a repository for dedicated projects. In particular, spatial and temporal analysis suggest that the Flickr dataset best reflects the NBN dataset when considering a purely spatial distribution with no time constraints. The best represented species on Flickr in comparison to NBN are diurnal garden birds, as around 70% of the Flickr posts for them are valid observations relative to the NBN. Additionally, we presented a fully automated image verification method for identifying genuine species observations on Flickr, suitable for verifying large and diverse collections of species. The approach is based on the Google Cloud Vision API in combination with species taxonomic data to determine the likelihood that a mention of a species on Flickr represents a given species.

In this chapter, we focused on developing verification methods suitable for image-based social network platforms. However, many widely used social network platforms such as Twitter are text-based which requires the development of verification methods suitable for text data. Therefore, we focus on developing automated text-based verification methods suitable in Chapter 4.

Chapter 4

Text Classification for Verifying Social Media Relevant to Wildlife

In the previous chapter, we presented an automated verification approach suitable for verifying images tags related to wildlife. While image verification techniques are undoubtedly very useful, there are many social media posts (i.e., Tweets) mentioning species names that do not include images. Further, an image-based verification approach does not in itself provide a fine-grained distinction between wildlife-related posts and posts that are actual wildlife observations. In this chapter, we present an automated verification method for identifying Tweets (text-based posts) related to wildlife observations using text classification approaches.

A problem with using social media such as Twitter to identify wildlife is that postings frequently use the common names of wildlife species in contexts that are totally unrelated to making a wildlife observation. For example, the keyword *'bluebird'* can refer to a species but it can also refer to a rugby team, as in the Tweet *'Come on blue birds #bluebirds'*. Another example is the keyword *'snipe'* which can refer to the bird Snipe but it can also be used in the sense of shooting, and is widely used terminology in video games, e.g. *'Im LIVE right now come watch me trying to snipe !...'*. Common names of wildlife species can also be used to refer to a restaurant or a brand, such as *'The Swan'*. A further issue with data quality arises with regard to the reliability of species identification in those message postings that are intentional observations. An associated challenge is that of distinguishing between wildlife-related Tweets that are direct observations and Tweets that mention wildlife but are not observations. For instance, the Tweet *'Unfortunately predators invasive alien species IAS like grey squirrels contributing decline native #wildlife red squirrels #ias like must also controlled'* discusses a wildlife topic rather than being indicative for the presence of species. In comparison, the Tweet *'Mine always big fans coolest greylag #goose never forget spotted #bird question observing #mandarin #duck taking stroll park #greylaggoose #mandarinduck #aixgalericulata #anser'*, indicates observations of a duck and a goose. In this regard, literature is sparse in presenting solutions for validating social media postings that may be useful biodiversity observations.

Sections 2.3.5 and 2.3.6 of the Background chapter presented text classification methods for social media and wildlife data, respectively. In summary, related research to wildlife data [Stringham et al., 2021, Monkman et al., 2018, Jeawak et al., 2018, 2020, Leung and Newsam, 2012] is limited in scale and involves the need for manual work. Additionally, studied classification approaches are mainly based on using statistical classification models, without fully exploring different feature selection and classification methods. Recent research [Al-Garadi et al., 2021, Guo et al., 2020, Liu et al., 2021, Lopez-Lopez et al., 2021] on text classification for social media is using transformer-based deep learning methods. However, most of the approaches assume a large number of training instances, and lack detailed comparison between different feature extraction and feature integration approaches and consideration of how these affect the performance of classifiers. Additionally, transformer-based models have not been fully explored in ecology-related studies.

We address these gaps by proposing a text classification-based solution for identifying Tweets which include posts for genuine wildlife observations regardless of the species observed. Three classification approaches are compared, in particular logistic regression classification with various forms of input features; the word embeddings based fastText pipeline; and the contextual word embedding model of BERT. We perform experiments with pre-trained and corpus-trained embeddings as well as different methods for building feature vectors. Species distribution data were obtained from Twitter, because of its wide usage and its real-time nature. The data we have obtained relate to 37 species, including invasive species in the UK. We also look at language in the Tweets (including specific hashtags and other text) that is indicative for wildlife occurrences. This can help the creation of targeted campaigns that influence social media trends in order to produce higher quality data.

The remainder of this chapter addresses question **RQ 2: What are the most efficient text classification approaches for verifying that social media postings are genuine wildlife observations?** from the research questions identified in Section 1.2. More specifically, contributions include:

1. A fully automated text classification approach for identifying genuine wildlife observations on Twitter - not restricted to species types or geo-tagged Tweets. Our approach takes a Tweet as an input and produces a class label for this Tweet with no human interaction.
2. An analysis of the relative effectiveness of different approaches to extracting and integrating features (i.e. data items) that serve as the input to several alternative forms of text classification, given a relatively small corpus of data for training the classifiers.
3. An investigation into the specific components of Tweets, including hashtags and URL links, that are indicative for genuine wildlife observations on social media

The rest of the chapter is structured as follows. Section 4.1 describes the methods we use for feature extraction and integration as well as the classification approaches we compare. We further explain the method of collecting and pre-processing the Tweets in order to be used in the classification pipeline. Section 4.2 presents classification results, findings from an analysis on the features indicative for wildlife, and error analysis. Section 4.3 reflects on the methods used and the findings, while Section 4.4 concludes the chapter.

4.1 Methods

In this section, we explain the development of a text classifier model, for identifying wildlife observations on social media sites. The methodology we follow consists of five main steps, *Tweets collection*, *Pre-processing*, *Feature Extraction*, *Feature Integration*, and finally training a *Wildlife Observation Classifier*.

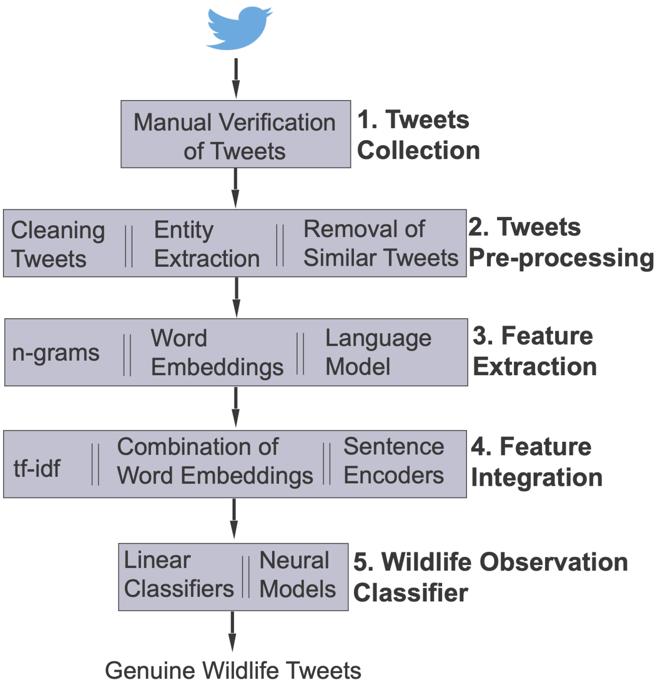


Figure 4.1: Overview of the methodology followed to build a classifier including main steps ('Tweets collection', 'Tweets pre-processing', 'Feature Extraction', 'Feature Integration', 'Wildlife Observation Classifier') as well as the different methods we experimented with during each of these steps.

See Figure 4.1 for overall flow of the methodology and Figure 4.2 for an example of a Tweet being processed using the classification methodology. During the collection and pre-processing steps (see Sections 4.1.1 and 4.1.2) we gather Tweets related to wildlife, from which stop words are removed, tokens normalised, and duplicates are removed. During *Feature Extraction* (see Section 4.1.3) we build word feature vectors for the

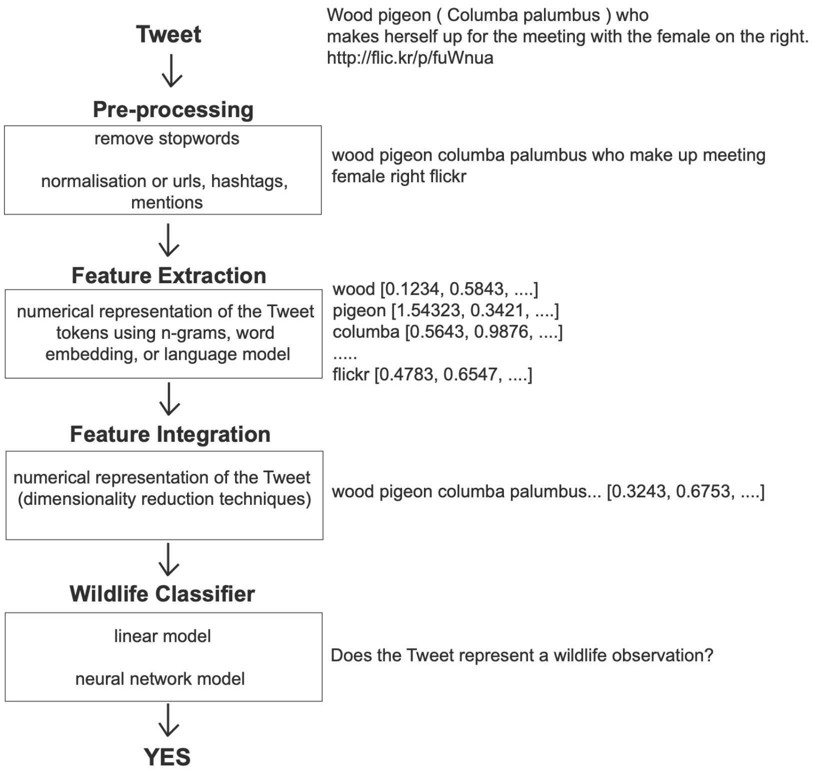

Tweet
Wood pigeon (Columba palumbus) who makes herself up for the meeting with the female on the right. http://flic.kr/p/fuWnua
Pre-processing
remove stopwords
normalisation or urls, hashtags, mentions
wood pigeon columba palumbus who make up meeting female right flickr
Feature Extraction
numerical representation of the Tweet tokens using n-grams, word embedding, or language model
wood [0.1234, 0.5843, ...]
pigeon [1.54323, 0.3421, ...]
columba [0.5643, 0.9876, ...]
.....
flickr [0.4783, 0.6547, ...]
Feature Integration
numerical representation of the Tweet (dimensionality reduction techniques)
wood pigeon columba palumbus... [0.3243, 0.6753, ...]
Wildlife Classifier
linear model
neural network model
Does the Tweet represent a wildlife observation?
YES

Figure 4.2: Step by step guide of the methodology using the example of a Tweet (left side — describes steps while right side — gives a relevant Tweet representation for each step).

corpus using techniques based on the feature representation approaches, described in Section 2.3.3. In the *Feature Integration* step, we combine word feature vectors, using dimensionality reduction techniques, into a single feature vector representing the entire Tweet (see Section 4.1.4). Finally, we experiment with three classification algorithms for building a *Wildlife Observation Classifier* (Section 4.1.5). These are based on the three main types of supervised machine learning approaches explained in Section 2.3.2.

4.1.1 Tweets 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 ten most numerous species on NBN, and the ten most numerous species on Flickr, some of which overlap. We used the species names of the most numerous species on Flickr and NBN (obtained from analysis presented in Chapter 3) as search terms for collecting data from Twitter. This facilitates future comparisons between Twitter and Flickr for wildlife observation studies. Further, numerous species on Flickr can also have a higher number of records on Twitter. We searched Twitter for 38 species and found posts for 37 species in total (we provide more information on data distribution per species in Section 4.1.6, Table 4.6). The Tweets have been collected regardless of whether they are geo-tagged. The reason for this is that the majority of Tweets are not geo-tagged, even though some of these could be geo-tagged if they contain geographic references. It is also the case that for some of the UK invasive species the number of geo-tagged Tweets is relatively low. We collect Tweets for the period 2007 – 2019 using the historic Twitter API. For each Tweet, we downloaded the following information: date when the Tweet was posted, username, any hashtags, mentions (i.e. Twitter usernames preceded by the @ symbol), and links associated with the Tweet. Additionally, we only downloaded Tweets written in English. The collection of Tweets is used for training a text classifier.

4.1.2 Tweets Pre-processing

Cleaning Tweets Stanford NLP Core [Manning et al., 2014] is used for pre-processing the dataset, in particular for POS tagging. Stanford NLP Core is a set of tools for performing various NLP tasks such as tokenisation, POS tagging, sentiment analysis, NER, etc. The library has been used in various research achieving satisfactory results. Stop words were removed using the Natural Language Toolkit (NLTK) [Bird et al., 2009] stop word list. NLTK is a python library for performing NLP analysis. Following tokenisation of the Tweets we identify hashtags, mentions, external links, and pictures within the Tweets text. External links within the Tweets were normalised in order to identify the main website source and disregard other parameters associated with the link such as queries and fragments. For example, the url *'https://youtube/uJZh5Ou1WNUa0'* after normalisation is *'youtube'*.

Entity Extraction We extract named entities in order to identify noun phrases using Stanford NLP Core. We use the noun phrases and named entities to identify terms (e.g. *'blue tit'*, *'audiology house'*) that could assist in classification. These terms are used to build feature representations with the BOW approach rather than only using tokens (single words).

Removal of Similar Tweets A problem with the Tweets collection is the high number of duplicates, some of which are Re-Tweets, due to one person Tweeting an existing Tweet. Duplicate Tweets and Re-Tweets have identical or very similar vocabulary to the original Tweets. The presence of high numbers of duplicates causes uniformity of the dataset vocabulary and thus classifiers may overfit to the given duplicates and fail to give accurate predictions when Tweets with diverse language are given. To avoid overfitting, we remove duplicates using Levenshtein distance [Levenshtein, 1966]. Levenshtein distance is a string metric for measuring the difference between two word sequences where the distance between two words is the minimum number of single-character edits (insertions, deletions or substitutions) required to change one word into the other. A threshold of 0.97 similarity was defined for Tweets to be considered duplicates. The selected threshold was found following experiments with different values of the threshold (0.65, 0.80, 0.90, 0.97, 0.99). A higher value did not capture insignificant differences, such as misspellings and single character insertions between the Tweets, while a lower threshold was inappropriate as it returns Tweets that are not duplicates. Re-Tweets were removed using regular expression matching for Tweets starting with *'rt'*. We also removed single word Tweets.

The collection also contains a large number of similar Tweets, many of which are produced by spam accounts. Examples of such Tweets are: *'#Forkknife #Snipe #blackout #Ps4 #Callofduty #Ttv #Live #Twitch #Share #funko #Rage #Supportsmallstreamers* *live at ...'* and *'#Callofduty #blackout #Ps4 #Supportsmallstreamers #Snipe #Support #Live #funko #Forkknife #wack #Share live at ...'*. They share ten tokens *'#snipe', 'live', '#forkknife', '#blackout', '#callofduty', '#share', '#supportsmallstreamers', '#funko', '#ps4', '#live'*, which is the majority of the tokens in both Tweets. Thus, we consider these similar. The method we use for removing similar Tweets is based on finding the number of tokens that appear in both Tweets and it is performed in the following steps:

1. Convert Tweets to BOW representations
2. Given two Tweets, intersect their BOWs to find their common tokens:
3. If the length of the list containing the common tokens is above the *threshold* of 90% of the number of tokens contained in a Tweet, for one of the Tweets, then the two Tweets are considered similar and the Tweet with the flagged up threshold is removed.

4.1.3 Feature Extraction

We performed experiments with three main types of feature extraction techniques, as described in Table 4.1. They are reflective of the main existing approaches for building feature representations, identified in Section 2.3.3, i.e. simple n-gram representation, word embedding models, and language models. In particular, the n-grams are a combination of the 1-grams and 2-grams in the Tweet texts. The research presented in Section 2.3.3 showed that the two most efficient and well established approaches for building word embedding models are CBOW and skip-gram. Further, Word2Vec [Mikolov et al., 2013a] and fastText [Bojanowski et al., 2017] are the word embedding models based on these methods and have been successfully applied in many domains. Therefore, we have performed experiments with fastText and Word2Vec pre-trained word embedding models. A limitation of Word2Vec is that it ignores the morphology of words by assigning a distinct vector to each word. This Word2Vec limitation is addressed in the fastText approach [Bojanowski et al., 2017] where each word is represented as a bag of character n-grams which enables the construction of vectors for rare or misspelled words. Additionally, we use the GloVe word embedding model which uses a matrix of the co-occurrence of pairs of words to build word representations (explained in Section 2.3.3). We have included Glove pre-trained embeddings as it has been trained on Twitter data. In addition to the pre-trained fastText embeddings we use the wildlife-related Tweets collection to train a corpus-specific word embedding model using the fastText architecture. This uses the skip-gram method to build word embeddings with 300 dimensions.

Finally, we also perform experiments with the language model BERT [Devlin et al., 2019] introduced in Section 2.3.3. As explained in Section 2.3.3, BERT takes into account the context of each word and hence offers an advantage over word embedding models where words have fixed representations regardless of the context within which the word appears. In this chapter, we take an advantage of both steps in the BERT architecture (see Section 2.3.3):pre-training and fine-tuning. In this initial step of the methodology, we use the base pre-trained BERT model to create a sentence encoding (see next section) that will be used as input to a Logistic Regression classifier, before using the fine-trained version, with BERT's built-in classifier, in subsequent experiments.

Approach	Approach Description	Model	Model Description
n-grams	A continuous sequence of n tokens from a given text	1,2-grams	Represent Tweet as a sequence of 1 and 2 grams
Word Embeddings	Neural models that use uni-directional approach for learning word representations and thus they produce single vector of a word irrespective of the context in which it appears	Word2Vec pre-trained [Mikolov et al., 2013a]	A two-layer neural model that uses skip-gram to learn word embeddings from raw text.
		fastText pre-trained [Bojanowski et al., 2017]	Vector representations are generated for each character n-gram and words are represented as the sum of these representations.
		Glove pre-trained [Pennington et al., 2014]	A matrix of the co-occurrence of pairs of words is used to learn embeddings for which the dot product of pairs of word embeddings is equivalent to the log of the probability of the co-occurrence of the respective words.
		fastText corpus-based	We use Tweets to train a corpus-specific word embedding model using fastText. The skip-gram method is used to create word embeddings with 300 dimensions.
Language Model	Pre-train deep bidirectional representations by jointly conditioning on both left and right context in all layers.	base BERT [Devlin et al., 2019]	We use the base pre-trained BERT language model, which has been trained on Books Corpus and English Wikipedia

Table 4.1: Feature Extraction Step — A summary of main methods deployed during this step.

4.1.4 Feature Integration

In this step, we generate Tweet classification feature vectors using the three main approaches outlined for building feature vectors in Section 2.3.4 from Chapter 2. One is simply based on the statistics of the n-gram occurrences, specifically the counts of the 1-grams and 2-grams in the Tweet text. As an alternative to counts of word occurrences we experimented with using tf-idf values of words but this did not provide an improvement in performance. This can be regarded as BOW method. The second approach uses various combinations of word embeddings as features, one being the average of the embeddings of the words in a tweet and the other being a *tf-idf* weighted average of the embeddings where the *tf-idf* values are those of the respective words.

The third approach is based on sentence encoding methods. The first of these uses the uSIF (unsupervised smoothed inverse frequency) method of [Ethayarajh, 2018] that creates a weighted average of word embeddings where a lower weight is placed on more frequent words. The method introduces a weighting scheme that improves on the approach of [Arora et al., 2017]. To form a sentence embedding they subtract from the weighted average a weighted projection of the weighted average onto the first m principal components of all weighted average sentence embeddings, i.e. the *common discourse vectors* (where m equals 5 rather than only subtracting the first principal component as in Arora et al. [2017]). This is referred to as piecewise common component removal. The second sentence encoding method uses the pre-trained BERT base language model to extract the embedding of the token called [CLS], i.e. for classification, from the last hidden layer of the BERT neural network representation. The output corresponding to that token can be considered as an embedding for the entire input sentence. Note that the input to the first layer of the BERT model is a sequence of the embeddings of each word of the sentence where those initial pre-trained embeddings are modified in subsequent layers to adapt to their context. A summary of Feature Integration techniques is given in Table 4.2.

Approach	Approach Description	Model	Model Description
statistical approach	statistic-based approach for representing words in a sentence	count	We assign frequency weights to the 1,2 grams in a given Tweet
Combination of Word Embeddings	It uses simple average or tf-idf weighting of word embeddings in the sentence	Mean	Average the embeddings of each word in a Tweet along each dimension
		TF-IDF	We assign TF-IDF weights to the words in a Tweet, and calculate the weighted average of the word embeddings along each dimension (where the contribution of a word is proportional to its TF-IDF weight)
Sentence Encoders	It employs more specialised sentence encoding to adapt the word embeddings	uSIF [Ethayarajh, 2018]	Based on calculating the weighted average of word embeddings, with a lower weight placed on more frequent words. From each weighted average vector is subtracted the projection on their first principal components.
		BERT sentence encoder	A sentence embedding is represented by the embedding of the “classification” token [CLS] extracted from the last hidden layer of the BERT representation.

Table 4.2: Feature Integration Step — A summary of main methods deployed during this step.

4.1.5 Wildlife Observation Classifier

We use three types of classifier where each classifier represents one of the main text classification methods outlined in Section 2.3.2. In this way we want to ensure a coverage of the main existing approaches including the state-of-the-art. These are classical machine learning models, the fastText pipeline, and fine-tuned BERT. We experimented with a classical machine learning model based on frequency-based features and a suite of classification algorithms available in the Scikit-Learn library [Pedregosa et al., 2011], namely Gaussian Naive Bayes (GNB), Logistic Regression and Support Vector Machines (SVM). Of the three, the best results were achieved using Logistic Regression. We use Logistic Regression for the n-gram baseline and for the classifiers in which the features were those described above in the previous section. Thus these features include sentence representations based on average pre-trained Word2Vec, fastText and GloVe embeddings, corpus-trained fastText embeddings, as well as the uSIF and base BERT sentence representations. In addition to using pre-trained fastText embeddings with Logistic Regression, we used the fastText pipeline which has its own classifier. Our final form of classifier was the fine-tuned BERT model where an additional final layer of the model serves as a binary classifier. A summary of classification techniques is given in Table 4.3

Approach	**Approach Description**	**Model**	**Model Description**
linear model	It can represent linear relationships	Logistic Regression (LG)	A strong baseline for many text classification tasks [Joachims, 1998]; [McCallum et al., 1998]; [Fan et al., 2008], even more recently on noisy corpora such as social media text [Mohammad et al., 2018, Çöltekin and Rama, 2018]; however it tends to struggle with OOV words, fine-grained distinctions and unbalanced datasets
		fastText pipeline [Joulin et al., 2017]	It partially addresses issues associated with LG by integrating a linear model with a rank constraint, allowing sharing parameters among features and classes, and integrates word embeddings that are then averaged into a text representation
Neural Model	can learn non-linear and complex relationships	fine-tuned BERT [Devlin et al., 2019]	We use pre-trained BERT word representation model and add a final sequence classification layer

Table 4.3: Wildlife Observation Classifier — A summary of classification approaches used to build a classification model.

4.1.6 Dataset

We selected a subset of the initial Tweets collected using search phrases of common and scientific species names (introduced in Section 4.1.1). The subset was chosen randomly to ensure the subset is representative of the distribution of all Tweets among the different species search names. We manually annotated Tweets as either a genuine wildlife observation or a false wildlife observation. The main annotation was done by a single annotator. To verify the quality of the annotation two other people annotated a sample of 100 Tweets. In both cases a high level of agreement was found with the first annotation, with a Cohen-kappa value of 0.98 in both cases. Note that the annotation process involved following links within Tweets and examining the content of images, and paying attention to the nature of hashtags, where genuine wildlife Tweets were characterised by the common use of photos of the observation and of wildlife community tags, or of Latin names of species, thus allowing for the possibility of fairly reliable manual tagging as was found here (see Section 4.2.3 for a discussion of indicative features of genuine wildlife observations). Further, we balance the datasets among the two classes (genuine wildlife observation versus no wildlife observation). After removing Re-Tweets and similar Tweets, we were left with 2798 manually annotated Tweets.

We used all collected Tweets (i.e. 1,769,384) for producing the corpus-trained word embedding model excluding the Tweets which we manually annotated and are used for classification. The main features and statistics of the dataset used for training the word embedding model are summarized in Table 4.4.[1]. An overview of the manually labelled subset of the Tweets collection which was used for training the classifier is presented in Table 4.5.

#Tweets	1769384
#Tokens	31780390
Avg Length	18

Table 4.4: Twitter collection used for building corpus-trained word embeddings, consisting of unlabeled data. '#Tweets' refers to the number of Tweets used for training the model, '#Tokens' refers to the number of tokens within the collection, 'Avg Length' refers to the average number of tokens per Tweet.

Analysis into the distribution of Tweets per species (see Table 4.6) showed that the best represented species on Twitter can be split into three main categories: pretty, i.e. photogenic flowers (Bluebell, Daisy, Dandelion), sessile green plant species (Ivy, Beech, Bracken) and garden and aquatic birds, which are also diurnal (Blue Tit, Great Tit,

[1] *# Split* (e.g. *# Tweets*) in the table indicates the number of instances in the given dataset.

	Verified as True (wildlife occurrence	**Verified as False (no wildlife occurrence)**	**Total**
#Tweets	1,257	1541	2,798
#Tweets with hashtags	679	693	1,372
#Tweets with mentions	247	452	699
#Tweets with pictures	323	322	645
#Tweets with links	976	1,369	2,345

Table 4.5: A subset of the Twitter collection, manually labelled and used for training classification models ('#Tweets' refers to the number of Tweets labelled per class, i.e. verified as true wildlife observation or false wildlife observation).

Mallard).

Scientific Name	**Common Name**	**#Tweets**
Fagus sylvatica	Beech	298,542
Gallinago gallinago	Snipe	239,719
Parus major	Great Tit	132,798
Pteridium aquilinum	Bracken	116,591
Cyanistes caeruleus	Blue Tit	110,780
Hedera helix	Ivy	91,383
Bellis perennis	Daisy	87,471
Turdus merula	Blackbird	74,857
Scirurus carolinensis	Grey squirrel	65,300
Fringilla coelebs	Chaffinch	57,960
Passer domesticus	House Sparrow	43,135
Anas platyrhynchos	Mallard	46,135
Columba palumbus	Woodpigeon	44,851
Chloris chloris	Greenfinch	37,839
Prunella modularis	Dunnock	32,791
Taraxacum officinale agg.	Dandelion	31,948
Heracleum mantegazzianum	Giant Hogweed	31,570
Hyacinthoides non-scripta	Bluebell	30,282
Branta canadensis	Canada Goose	27,094
Aix sponsa	Wood Duck	27,403

Table 4.6: Tweets distribution per species — limited to the 20 best represented species on Twitter ('#Tweets' refers to number of Tweets per species).

4.2 Results

4.2.1 Evaluation Experiments

As mentioned in Section 4.1, our evaluation is focused on a mix of features, mostly employing various forms of word embeddings along with Logistic Regression, fastText [Bojanowski et al., 2017] and BERT [Devlin et al., 2019] classifiers. In addition to embedding-based features we include a Logistic Regression classifier based on frequencies of n-grams reflected by their counts of words as a baseline. We used the 1000 most frequent n-grams to form feature vectors for the baseline classifier (i.e. a bag-of-words approach).

The pre-trained and application corpus-trained word embeddings were fed as input to a fastText pipeline where we used default parameters and 'softmax' as the loss function. However, for the fastText classifier we present only results based on corpus-trained embeddings due to the poorer results produced with pre-trained embeddings. For the BERT classifier, we fine-tuned it for the classification task using a sequence classifier, a learning rate of 2e-5 and 4 epochs. In particular, we made use of the BERT's Hugging Face default transformers implementation for classifying sentences [Wolf et al., 2019]. The results of the classifier experiments were quantified with precision, recall, F1-measure and accuracy. We also used 10-fold cross validation. This helps to avoid any bias in the test and training sets with regard to particular species.

4.2.2 Classification Results

Classifier	Feature Extraction	Feature Integration	p	r	F1	Accuracy
Logistic Regression baseline	1,2 grams	count	93.33% (2.08%)	95.07% (2.43%)	94.16% (1.56%)	94.71% (1.41%)
LogisticRegression	terms	count	93.52% (1.75%)	94.59% (2.16%)	94.02% (1.19%)	94.60% (1.06%)
	Word2Vec pre-trained	mean	93.14% (1.85%)	90.79% (2.16%)	91.93% (1.39%)	92.85% (1.22%)
		TF-IDF	86.50% (2.36%)	69.42% (3.55%)	77.00% (2.97%)	81.43% (2.17%)
		uSIF	94.34% (2.10%)	91.57% (3.31%)	92.88% (1.63%)	93.71% (1.38%)
	fastText pre-trained	mean	92.44% (2.12%)	82.57% (2.78%)	87.19% (1.82%)	89.14% (1.46%)
		TF-IDF	91.61% (2.94%)	61.87% (5.07%)	73.75% (4.06%)	80.35% (2.53%)
		uSIF	91.53% (2.47%)	91.81% (3.05%)	91.62% (1.66%)	92.46% (1.46%)
	Glove pre-trained	mean	77.51% (6.46%)	87.96% (5.91%)	82.33% (5.73%)	76.34% (7.72%)
		TF-IDF	70.48% (3.61%)	92.13% (3.89%)	79.80% (3.09%)	70.85% (4.76%)
		uSIF	63.31% (1.25%)	95.81% (2.72%)	76.23% (1.37%)	62.67% (2.17%)
	fastText corpus-based	mean	92.57% (2.69%)	94.91% (3.17%)	93.67% (2.03%)	94.24% (1.86%)
		uSIF	92.27% (2.24%)	94.35% (2.83%)	93.27% (1.86%)	93.89% (1.68%)
	base BERT	BERT sentence encoder	92.31% (1.06%)	95.39% (1.58%)	93.82% (1.03%)	94.35% (0.93%)
fastText pipeline	fastText pipeline	fastText pipeline	93.44% (1.37%)	96.10% (2.15%)	94.74% (1.38%)	95.21% (1.23%)
fine-tune BERT	base BERT	BERT sentence encoder	**96.0%** (1.03%)	**96.1%** (1.45%)	**96.0%** (1.23%)	**96.0%** (1.84%)

Table 4.7: Results per classification approach (‘p’ refers to precision, ‘r’ refers to recall).

The baseline classifier based on frequency scores of n-grams as features provided remarkably good precision 93.33% and recall 95.07% (see Table 4.7). The feature extraction method based on noun phrase and named entity terms rather than 2-gram representation does not lead to significant improvement over the baseline. The classification model based on using Word2Vec pre-trained word embeddings is the best performing model using pre-trained embeddings. It performs better than classification models using Glove pre-trained embeddings. A potential reason for Glove to perform worse, even though it was trained with Twitter data, is that wildlife Tweets include a lot of common and Latin names for species which are not widely used in general Tweets.

The use of fastText corpus-trained embeddings led to further improvements over the pre-trained models with a 1% increase in F1-measure. Further to that, a simple linear classifier model coupled with corpus-trained fastText embeddings performed quite similarly to a linear (logistic regression) classifier coupled with the BERT sentence encoding resulting from the [CLS] token of the base BERT language model. The use of the uSIF sentence encoding method was usually found to be better than alternatives of a simple mean of word embeddings or a *tf-idf* weighted mean, but in some cases the improvement was relatively minor and in the case of the GloVe pre-trained embeddings it was inferior to the simpler alternatives.

Notably the fine-tuned BERT model gives the best results with precision, recall, F1-measure and accuracy all being 96%. The fastText pipeline is the second best performing classifier with precision 93.44%, recall 96.10% and an F1-score of 94.70%.

4.2.3 Indicative Features Analysis

We performed analysis on the features indicative for wildlife using the manually annotated Tweets. The results in Figures 4.3-4.5 show that there are trends across the usage of hashtags, mentions, and links distinguishable between the genuine wildlife Tweets and the non-genuine wildlife Tweets. For instance, the majority of the genuine wildlife Tweets have hashtags related to birds and wildlife, mentions of wildlife and nature groups such as *'@bbcspringwatch'* and *'wildlife uk'*. Further, the genuine wildlife observations include more links to pictures. In contrast, the false wildlife Tweets contain hashtags and mentions related to gaming groups.

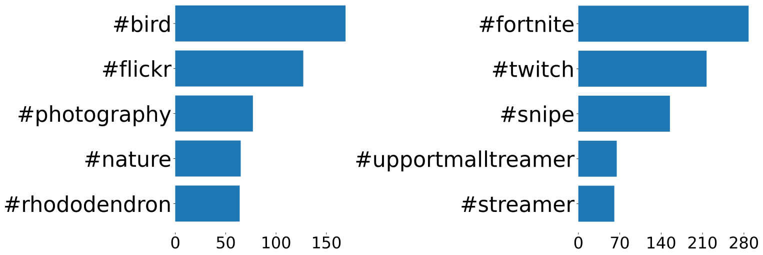


Figure 4.3: The ten most frequent hashtags per class label, Tweets with genuine wildlife observations (left), Tweets with false wildlife Tweets (right).

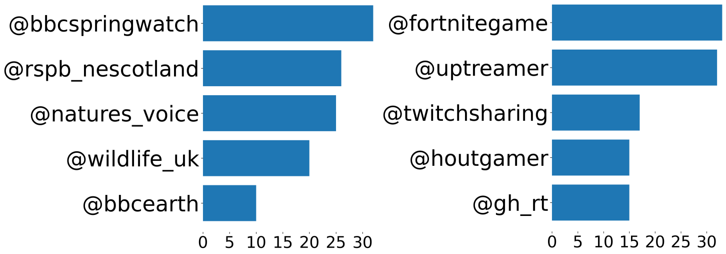


Figure 4.4: The ten most frequent mentions per class label, Tweets with genuine wildlife observations (left), Tweets with false wildlife Tweets (right).

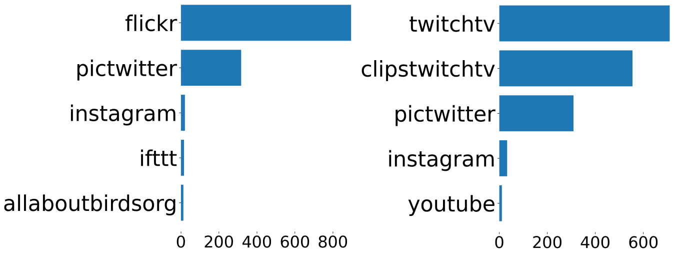


Figure 4.5: The ten most frequent URL links class label, Tweets with genuine wildlife observations (left), Tweets with false wildlife Tweets(right).

In order to identify whether hashtags, mentions, and URLs can be used as a way of distinguishing the genuine wildlife species we performed a statistical analysis looking at the number of non-genuine wildlife Tweets containing the most indicative features, displayed in Figures 4.3 to 4.5. Experiments showed that none of the top 5 most frequent wildlife-related mentions are present in the non-genuine wildlife Tweets. There are two false wildlife Tweets including wildlife indicative hashtags, i.e., *'#bird'* and *'#wildlife'* with examples respectively, *'Blue Tit Bird Painting Blue Yellow White http://dld.bz/fj5W5 #birds #wildlife #painting'* and *'Unfortunately predators invasive alien species IAS like grey squirrels contributing decline native #wildlife red squirrels #ias like must also controlled'*. The first Tweet is about a painting of a bird rather than an actual wildlife observation and while the second example is relevant to wildlife it is not about a wildlife observation. There is a single false wildlife Tweet with a wildlife indicative URL (i.e., *'instagram'*).

The main conclusions from this analysis are:

1. The presence of mentions such as *'@bbcspringwatch', '@rspb_nescotland', '@natures_voice', '@wildlife_uk', '@bbcearth'* are strong indications that a Tweet is a true wildlife observation since they are mentions of official campaigns for wildlife observations. However, these kind of mentions appear in less than a 100 Tweets. This suggests that crowdsourcing of wildlife observations could be improved by promoting such groups.

2. Hashtags such as *'#wildlife'* and *'#bird'* can be used for distinguishing between wildlife-related and false wildlife Tweets, but they are not indicative of Tweets with genuine wildlife observations. Photography related hashtags (*'flickr'*, *'photography'*) and nature-related tags have however been used exclusively in genuine wildlife observation Tweets. This suggests that there is a trend towards the usage of wildlife hashtags in Twitter wildlife observations that are not related to official campaigns. It might also explain why the baseline that uses only n-grams rather than embeddings as features, performs very well, albeit not as well as the BERT model.

4.2.4 Error Analysis

We compare the performance of the two best performing classifiers — fine-tuned BERT and fastText pipeline using a test set of 559 Tweets from the 2798 manually annotated Tweets, which corresponds to one test fold from the 10 fold cross validation (described in Section 4.2.1). A confusion matrix of the performance of the classification models is given in Table 4.8.

	‘Wildlife’	**‘Not Wildlife’**
Predicted as ‘Wildlife’	239	15
Predicted as not ‘Wildlife’	8	297

	‘Wildlife’	**‘Not Wildlife’**
Predicted as ‘Wildlife’	233	17
Predicted as not ‘Wildlife’	14	295

Table 4.8: Confusion matrix for fine-tuned BERT classification model (left) and fastText classification pipeline (right), where ‘Wildlife’ signifies genuine wildlife observation and ‘Not Wildlife’ signifies Tweets that are not genuine wildlife observations.

Error Analysis comparing the false positives and false negatives between the two classifiers showed that BERT performs better for Tweets which mention species name in a different context than wildlife. An example of false positive for fastText where BERT correctly classifies the Tweet as ‘not Wildlife’ is *’looking buyer 8 woodduck #littleeggharbor #nj #realestate http://tour.circlepix.com/’*. This Tweet is about buying property with the name *8 woodduck* rather than talking about the species. BERT also performs better for Tweets containing the Latin names of the species. A false negative example of the latter for fastText, where BERT correctly classifies the Tweet as ‘Wildlife’ observation, is: *’spotted branta canadensis canada goose in our garden pic.twitter.com/’*.

Experiments comparing the best performing classifiers for different Tweets lengths showed that BERT performs better than the baseline for any length. Further, fine-tuned BERT gives better results than fastText pipeline and the baseline for shorter Tweets.

For long sentences fine-tuned BERT and fastText pipeline have very similar performance with a difference less than 1% (see Figure 4.6). BERT outperforms fastText for sequences shorter than 10 tokens while for longer sentences, especially when containing more than 20 tokens, the performance between the two classifiers is very similar. This shows that BERT is the most suitable model for classifying shorter sequences such as social media posts compared to fastText and the Naive Bayes classifier.

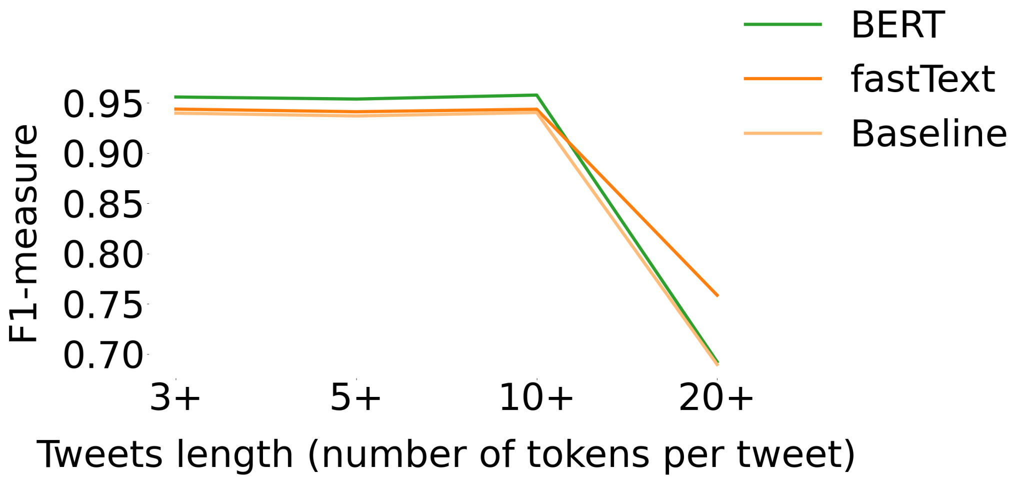


Figure 4.6: Comparison between the performance of baseline, fastText, and BERT classifiers for different length of Tweets.

4.3 Discussion

4.3.1 Classification Methodology

In order to perform a thorough analysis into existing classification techniques, including state-of-the-art, we used multiple approaches for each main stage of the text classification process - feature extraction, feature integration, and classification algorithm, as illustrated in Figure 2.3, part of Section 2.3.1. We performed experiments with three classifiers, representative of the main types of classification algorithms: a classical (linear) Logistic Regression, the fastText pipeline and the fine-tuned BERT transformer-based model classifier. These methods were used variously in association with features that consisted of simply counts of the actual words (as 1- and 2-grams) in the Tweets, which was treated as a baseline, and various forms of embeddings of the sequence of words in a Tweet. These latter sentence embedding methods included simple and *tf-idf* weighted averaging of the embeddings of each word, along with the uSIF sentence embedding method and the sentence embedding obtained from the CLS token of the last layer of the basic BERT language model. Various word embedding methods were employed, namely pre-trained GloVe, Word2Vec and fastText, corpus trained fastText embeddings, along with the contextually generated BERT embeddings.

Developing this classification methodology extends on related work, presented in Sections 2.3.5 and 2.3.6 by providing in-depth analysis of existing feature extraction, feature integration, and classification techniques. This helped identify algorithms for building automated verification models suitable for diverse wildlife-related textual data even when the labelled corpus is limited.

4.3.2 Findings

The main findings from this chapter are:

1. Classification results and error analysis presented in Sections 4.2.2 and 4.2.4 showed that, despite the relatively small amount of labelled data, features based on the corpus-trained embeddings from fastText produced better results than pre-trained embedding models including the GloVe embedding model, trained on generic Twitter data. The latter performance advantage can be attributed to the fact that genuine wildlife observations can use Latin species names which might be relatively insignificant in use in the pre-trained GloVe embeddings. It is this occurrence of distinctive vocabulary that might also explain why the baseline Logistic Regression classifier, in which the features were either simply the count of words or of n-grams, outperformed all other classifiers except the fastText

pipeline and the fine-tuned BERT classifier. Regarding the specialised sentence embedding method of uSIF, in the case of pre-trained Word2Vec and fastText embeddings it was found to be superior to mean and tf-idf weighted methods, but for GloVe the opposite was the case. Also for fastText corpus-trained embeddings uSIF was slightly inferior to using the mean. These findings indicate that pre-trained embedding models, trained on large but generic corpora are less beneficial for classification of text with more specialised terminology (i.e. the species-related data), compared to corpus-trained word embeddings which are trained on a smaller but more task-specific dataset. Notably the BERT sentence encoding method with its contextually adaptive embeddings achieved a similar performance to the fastText corpus trained method. In contrast, fine-tuning the BERT sentence encoding model for the classification task outperforms fastText classifier and the same BERT sentence encoding method coupled with Logistic Regression. This shows that fine-tuning the transformer-based models to the task is highly beneficial for the performance of the models versus using the pre-trained models as an input the statistical classification algorithms.

2. The best performing fine-tuned BERT classifier performed well even for Tweets with more specialised language (i.e. Latin names, use of non-English words). Further, it correctly classified non-genuine wildlife observations Tweets that used the common names of wildlife species in contexts that are totally unrelated to making a wildlife observation. This indicates that deep learning transformer models can perform well even for small amounts of labelled data, especially when more contextual knowledge is needed. Further, this BERT model performed better than linear models for very short Tweets while for longer Tweets, deep learning performed similarly to linear models. The high performance of the fine-tuned BERT classifier (i.e., 96% accuracy) shows the potential of state-of-the-art deep learning models to be used for developing automated tools for identifying valuable ecology data among informal social network sources automatically and on a larger scale, independent of the species observed at hand.

Therefore, this research addresses many of the gaps associated with previous work on text classification for wildlife data, presented in Section 2.3.5 where some solutions involve manual processing, the use of linear classification models or analysis limited to a few species. Additionally, our analyses address the suitability of different classification approaches for smaller wildlife-related datasets, compared to previous research presented in Section 2.3.5

3. Analysis on the use of hashtags and mentions across genuine wildlife observation Tweets showed that hashtags such as *'#wildlife'* and *'#bird'* can be used for distinguishing between wildlife-related and false wildlife Tweets, however, they are not indicative of Tweets with genuine wildlife observations. Photography related hashtags (*'flickr'*, *'photography'*) and nature-related tags have however been used exclusively in genuine wildlife observation Tweets. This suggests that there is a trend towards the consistent usage of hashtags related to wildlife observations which are not related to official campaigns. In future, such hashtags could be used by informal social network campaigns to encourage people to indicate when they are posting about wildlife. However, the presence of some of these hashtags cannot be alone considered adequate in itself for identifying wildlife observations. A reason for this is that the list of indicative features may expand as new species names are used or Tweets are collected for different time spans, regions, and languages. Additionally, Tweets often contain misspellings which can affect the representation of indicative features. The use of more sophisticated methods such as language models with contextual word embeddings allows us to identify semantic relationships between terminology used and the given class. More specifically, terms with similar meaning will have similar representations which can help accurate classification despite the diverse spelling or diversity of terminology. Therefore, classification models can be improved by creating feature selection techniques which assign higher importance to such indicative features. This would allow these methods to be applied to a wider range of species, geographical regions and even different languages.

4. The statistical analysis, presented in Table 4.6 show trends in the best represented species on Twitter, which can be split into three categories, i.e. pretty (photogenic) flowers, sessile green plant species, and garden and aquatic birds. Similar species distributions have been found in Flickr, presented in Chapter 3, which suggests that there are common trends among different social networks on the type of species they represent well. A more detailed analysis into the value of Twitter for collecting species-specific data is outside the scope of this Chapter. Instead, we are interested in providing tools for identifying genuine wildlife-related data which can be applied to studying any kind of species. However, in future, the developed classification pipeline can be used to filter genuine wildlife observations which can then be used to perform more detailed analysis of spatial and temporal distribution of specific species.

4.3.3 Limitations

The work presented in this chapter could be extended by using larger transformer-based models such as RoBERTa and performing experiments with corpus-trained language models as well as experimenting with classifiers using earlier neural networks such as CNN. Despite this limitation, our analyses were conclusive in that transformer-based models when fine-tuned to the classification task can be very valuable in verifying wildlife-related observations on social network platforms without the need to collect large training corpus. Considering the small volume of data, we have performed evaluation using 10-fold cross validation to ensure that the model does not overfit the dataset. The results show a standard deviation less than 1.8%. In future, it would be of interest to experiment with a larger dataset and to include a wider range of species with a view to making the validation method more generally applicable. This will allow us to extend the quantitative analysis presented in Chapter 3 including Twitter datasets as well. The presented validation method can be regarded as limited in that it is generic, being with regard to wildlife in general. It would be of interest to develop validation methods that applied to individual species.

4.4 Conclusion

In this chapter, we have explored the problem of identifying genuine wildlife observations on Twitter using text classification approaches. This is a significant challenge as Tweets commonly mention species names without being actual observations of the named species. In preparation for developing a machine learning classifier to identify genuine observations we created a dataset of Tweets that were manually annotated according to whether or not they were classed as genuine wildlife observations. Question **RQ 2** from the hypothesis presented in Section 1.2 has been answered in order to show that a state-of-the-art language model such as BERT, when fine-tuned for classification, is valuable in classifying correctly instances with more specialised terminology even when a training set of less than 3000 instances is provided. This shows the potential of state-of-the-art neural network transfer learning techniques to facilitate the discovery of valuable wildlife related data on social networks without the need of human verification steps or officially organised citizen science campaigns. Analysis into the usage of hashtags, mentions, and URL links throughout the genuine wildlife related Tweets suggested trends into the use of hashtags that are unrelated to official citizen science campaigns. Such hashtags can therefore be exploited in automated feature selection techniques for improving classification performance, as well as used as part of more informal campaigns encouraging people to use these hashtags when wildlife observations are posted. We provided a broad analysis of the suitability of various text classification and feature extraction methods for identifying genuine wildlife observations on social media. In doing so we address the need for devising automated strategies which facilitate the discovery of valuable ecology-related data from informal online sources which can be used to expand and enrich existing citizen science data portals.

In this and the previous chapter, we focused on the problem of developing verification techniques suitable for validating social media posts related to wildlife suitable for large and diverse species data collections. In the next chapter, we focus on the problem of geo-referencing text posts in order to facilitate further the use of social media data in wildlife-related studies such as movement patterns identification.

Chapter 5

Geo-referencing Social Media Data Related to Wildlife Observations

As explained in Section 2.4 from Chapter 2 and found in our analysis presented in Section 4.1.1 from Chapter 4, it is hard to obtain geo-referenced Tweets, especially when search is limited to a certain topic or region. However, geo-referenced datasets obtained from social networks sites have the potential to facilitate studies of wildlife distribution patterns which in turn are increasingly important for alerting rapid ecosystem changes such as climate change, diseases spread, and invasive species occurrences [Amano et al., 2016, Barve, 2014] (discussed in Section 2.4, Chapter 2). Therefore, in this chapter, we address the problem of geo-referencing social media posts. Similarly to the previous chapter, we use Twitter for performing experiments as it has the potential to serve as a source of valuable wildlife-related data.

We opted for using a regression-based approach for assigning coordinates versus the more widely explored language modelling approach as regression algorithms do not require additional steps for assigning coordinates and also do not involve partitioning of the training data into clusters or grids which can be data-specific and data consuming task (discussed in Section 2.4).Geo-parsing in combination with geocoding is another method which achieves high precision for georeferencing especially when there is a limited amount of annotated data as it does not require any training data. However, as discussed in Section 2.4 this approach works only when the social media posts mention location names which are also present in the gazetteer and by itself will often fail to resolve ambiguity correctly when the same place name can refer to multiple locations.

Scherrer et al. [2021, 2020] showed that the regression approach, especially when combined with a transformer-based model (contextualised neural network word representation model) led to significant improvements over classification approaches for geo-referencing social media posts. However, there is still a lack of extensive research on how neural network models and transfer learning techniques can be utilised for predicting coordinates of social media, especially in settings with a relatively small training set (less than 150,000 training instances) and wildlife specialised posts. Previous research has also indicated the benefit of a hybrid approach that combines language models with geo-parsing [DeLozier et al., 2015]. A benefit of using regression models is that it always returns coordinate values for a given test instance independent of whether the instance include a place name or not. A benefit of geo-parsing methods is that it does not require annotated data and it can predict coordinates with high accuracy but only for instances which mention place names. Thus, the two approaches can be complementary to each other. However, regression has not previously been used in combination with geo-parsing or other rule-based methods in previous research.

In this chapter, we address these gaps by adopting a hybrid approach in our case applying state-of-the-art neural network models to the regression task with support for multivariate regression in order to predict latitude and longitude values for Twitter posts, in combination with gazetteers where place names are present. We perform experiments with various transfer learning techniques showing that transformer-based word representation models trained on the domain, and on training data enriched with multiple social media sources, leads to significant improvements in georeferencing Twitter posts. Further, investigation into two strategies for improving precision of regression models showed that a location names extraction method based on using Named Entity Recognition (NER) and gazetteers enhances the precision of geo-referencing approaches.

In the rest of the chapter, we address research question **RQ 3: Can deep learning transformer regression models provide an effective means of geo-referencing so-**

cial media posts? from Section 1.2. The contributions of this research are:

- A geo-referencing model based on state-of-the-art neural network word representations which has been adapted to multi-output regression — The model outperforms regression models based on statistical machine learning algorithms
- A domain trained neural network-based word representation which when adapted to multi-output regression does lead to improvements in geo-referencing Tweets over the publicly available pre-trained models
- A transfer learning method for enhancing regression models for geo-referencing Tweets by enriching training sets with Flickr geo-referenced posts
- Analysis of the effectiveness of two hybrid approaches, involving regression for improving the precision of geo-referencing models.
- Providing the largest collection of geo-referenced wildlife-related Twitter data, to the best of our knowledge

The rest of the chapter is structured as follows: Section 5.1 describes the methods we used and the methodologies for building the regression model and the hybrid approaches. In Sections 5.2 and 5.3 we present the results and discussion. Finally, Section 5.4 concludes the paper and presents future directions.

5.1 Methods

We focus on geo-referencing Tweets relevant to wildlife observations within the UK. For these purposes, we use a transformer-based model (RoBERTa language model), which we adapted for the regression task, similarly to geoBERT [Scherrer et al., 2021]. The methodology for building the regression model for predicting coordinates consists of two main steps (see Figure 5.1). In the first step, we pre-train the RoBERTa language model to obtain word representations. In the second step, we fine-tune the language model to the regression task in order to assign coordinate values to Tweets. Additionally, we perform experiments with two approaches for improving the precision of the geo-referencing models (see Sections 5.1.3 and 5.1.4). We compare the developed methods to a statistical regression model and BERT language model, fine-tuned for the regression task. Furthermore, we address the problem of having a small volume of training data by combining Twitter posts with text-based Flickr posts. We provide more detail in Section 5.1.5.

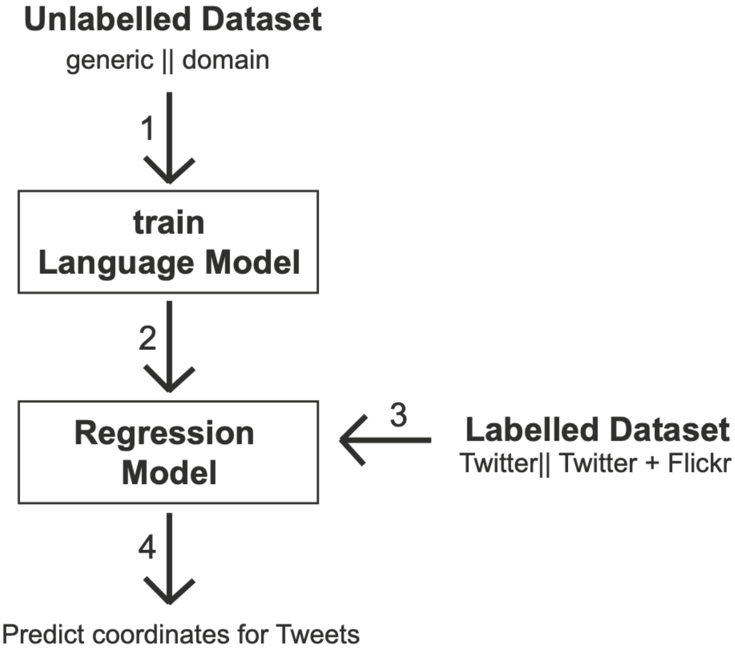


Figure 5.1: Methodology for building RoBERTa-based regression model for predicting coordinates of Tweets.

5.1.1 Pre-training Language Model

We perform experiments with two RoBERTa language models:

- Generic base RoBERTa model [Liu et al., 2019a] — The base RoBERTa model has been trained using generic English datasets. The model is case sensitive.
- Domain-trained RoBERTa model — 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 5.1.5, that are not associated with coordinates. We used the masked language modeling (MLM) technique for fine-tuning RoBERTa where, given a sentence, the model randomly masks 15% of the words in the input before predicting the masked words [Liu et al., 2019b]. This technique enables learning more contextually rich sentence representations, compared to earlier neural network models (see Section 2.3.2). Notably, the MLM technique has also been used for pre-training the base RobERTa model. The model was fine-tuned for three epochs using the Hugging Face library [Wolf et al., 2019] implementation for MLM.

As one of our baselines, we use the BERT language model pre-trained using large generic unlabelled corpora from various sources. We used the large uncased BERT model available from the Hugging Face library.

5.1.2 Regression Models

We develop a regression model by adapting RoBERTa to multivariate regression in order to be able to assign latitude and longitude values to each unlabelled Tweet. Currently, the RoBERTa architecture supports the regression task for single values using Mean Square Loss function. We adapt RoBERTa to multivariate regression (for both latitude and longitude prediction), calculating Mean Square Loss function per label.

For these purposes, we used the implementation of RoBERTa for multi-label classification provided by the Huggingface Simple Transformers library [Wolf et al., 2019]. In order to train our regression model, we used 10 epochs, a batch size of 32 and we also save only the model which performs the best for the development set.

As our baselines, we have used two other regression approaches. We used the Support Vector Regression (SVR) [Awad and Khanna, 2015] algorithm which has been used in previous research on geo-referencing social media data and is known to generalise well to unseen data with good accuracy. Our implementation used the Scikit-Learn library version of the SVR algorithm [Pedregosa et al., 2011]. It was adapted to multiple-output regression using a simple strategy consisting of fitting one regressor per target (latitude and longitude values). As input features, we used TF-IDF-weighted n-grams consisting of characters with length 3-10. We performed experiments with other character lengths, however, 3-10 led to the highest results. We have also compared the RoBERTa regression model to the BERT model fine-tuned for regression following implementation methods explained in Scherrer et al. [2021].

In the rest of this section, we will describe the hybrid approaches we used for improving the precision of geo-referencing methods. They make use of the regression models developed in the paper and techniques that require less training data, which makes them suitable for small datasets.

5.1.3 Hybrid Approach based on Location Names Extraction and Regression

We have implemented a hybrid approach that combines a location names extraction method, based on NER and geocoding, with the RoBERTa regression model (see Figure 5.2). The location names extraction approach, exploits the presence of place names within the Tweets where location names are first extracted from the Tweets and mapped to their coordinates. Location disambiguation is performed at two stages of the approach. We use the RoBERTa regression model coordinates when no location names are found in a Tweet, and for performing disambiguation when an ambiguous place name is detected with the gazetteer. The approach consists of the following steps, as shown in Figure 5.2.

Location Names Extraction: We identify location names within the Tweets using two named entity recognition (NER) methods. An entity is regarded as a potential place name is it has one of the following NER labels: 'GPE', 'FAC', 'LOC' and 'ORG'. To improve 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 [Honnibal and Montani, 2017] which has been successfully used for NER for short texts in previous research. We use the transformers pre-trained NER model, part of the library. We have also used the Flert NER model [Schweter and Akbik, 2020] trained on a large English language corpus, available from `https://huggingface.co/flair/ner-english-large`. In an initial analysis, we also used the BERT model fine-tuned for Named Entity Recognition (NER) [Devlin et al., 2018], trained on the CoNLL-2003 English news articles dataset [Tjong Kim Sang and De Meulder, 2003]. However, the results showed that the latter model does not perform well for the given dataset.

Map Locations to Coordinates: We obtain the coordinates for each location name, identified in the first step by using the geocoding library Nominatim[1]. Nominatim uses OpenStreetMap data to find the coordinates for given location names. We have limited the geocoded results to be UK-based because we analyse only wildlife observations within the UK.

Location Names Disambiguation: We perform location names disambiguation at two stages of the approach, also described in Algorithm 5.1:

[1] Nominatim: `https://nominatim.org`

Algorithm 5.1: Location Name Disambiguation Heuristic

```
Input: Tweet
Output: lat,lon
if $loc \in Tweet$ then
    if $len(loc) > 1$ then
        return finest grain location object
    if geocoding returns multiple instances loc(lat, lon)) then
        return loc(lat, lon) closest to regression(lat, lon)
else
    return regression(lat, lon)
```

- If a Tweet contains more than one location, we select the location which refers to the finer grained geographic object.

- If more than one location has the target place name, we select the location with coordinates closest to the coordinates returned by the regression model, calculating distance with the Harversine formula.

- If a Tweet does not contain location names, then we use the coordinates returned by the regression model

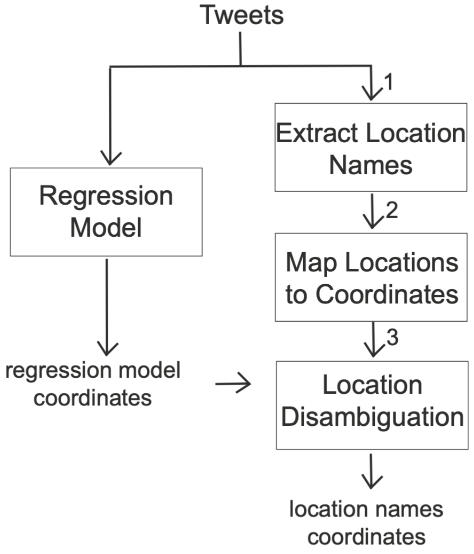

Tweets
1
Extract Location Names
2
Map Locations to Coordinates
3
Location Disambiguation
location names coordinates
Regression Model
regression model coordinates

Figure 5.2: Description of Hybrid Approach based on Location Names Extraction and RoBERTa-based regression.

5.1.4 Hybrid Approach based on Semantic Similarities and Regression

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 most similar item from the training data that is also located in the same target cell or cluster as that predicted, 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 5.3, are as follows:

Step 1: Define Regions: For each unlabelled instance, we find the training instances which are within a given radial distance from the given unlabelled Tweet based on the coordinates predicted by the RoBERTa-based regression model. We performed experiments with three radial distances, i.e. 5km, 10km, 20km.

Step 2: Find the most similar training instance: We find the most similar training instance to each unlabelled Tweet using neural network models for building sentence embedding representations of the Tweets and then calculating the cosine similarity between the embedding vectors. For each unlabelled instance, we select the most similar training Tweet within the region of the given instance. In order to obtain embedding representations for the Tweets, we experimented with two neural network architectures. We used the sentence transformer model, Sentence-BERT (SBERT) [Reimers and Gurevych, 2019] which is a modification of the pre-trained BERT network that uses siamese and triplet network structures to derive sentence embeddings that can be compared using cosine similarity. We used a pre-trained SBERT model which has been trained on a large and diverse dataset of over 1 billion training pairs, available from Hugging Face library at `https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2`. We also performed experiments with corpus-trained embeddings obtained with fastText architecture [Bojanowski et al., 2017]. However, the results were unsatisfactory.

Step 3: Re-assign coordinates to unlabelled Tweets: Finally, we give each test instance the coordinates of the most similar training Tweet that is within the region of the test instance.

Step 4: Average Coordinates: We average the coordinates obtained using the two methods, i.e., regression and semantic similarity approach. We performed experiments with and without this final step.

5.1.5 Datasets

We collected Twitter and Flickr datasets limited to UK boundaries and related to wildlife observations. For these purposes, we used search phrases relevant to common and

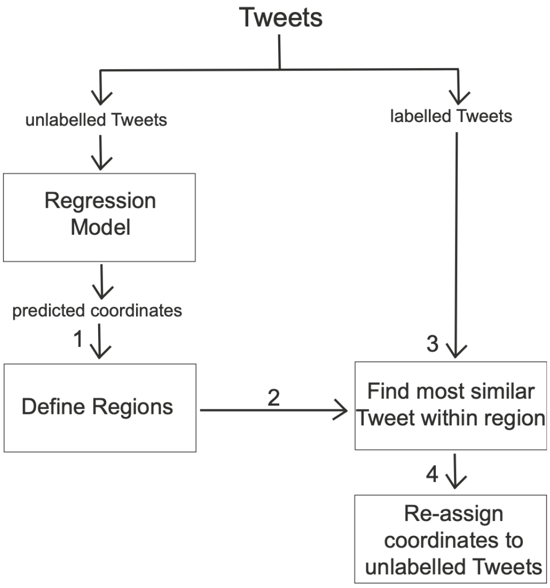


Figure 5.3: Description of Hybrid Approach based on Semantic Similarities and RoBERTa-based Regression.

scientific names of various species within the UK. The Tweets were collected regardless of whether they are geo-tagged. We collected Tweets for the period 2007 – 2019 using the historic Twitter API. We are interested in predicting coordinates referred to by the Tweet text rather than the user profile location and therefore we downloaded only the Tweet information. We do not use the user profile location information because it might not match the location at which the Tweet has been created. This is especially true for species related posts where users often travel to different locations and take a note of wildlife observations that they have encountered during their travels. Specifically, for each Tweet we downloaded the post, any hashtags, mentions, and links associated with the 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 for pre-training the BERT language model which is later fine-tuned for the regression task (see Table 5.1). We further discuss the different language models that we built in Section 5.1.1.

We downloaded Flickr data using the Flickr API interface for the period 2007 – 2019

	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

Table 5.1: 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)*).

inclusive. We limited search to geo-referenced Flickr posts because we use Flickr data only as a supplement to the fine-tuning stage where data with labelled coordinates is required. Additionally, we downloaded only the text-related data (title, description, tags) because we want to augment the Twitter dataset with additional text data, with no exploitation of associated images. We added the text data from Flickr (title, description, tags) with associated coordinates to the training set of Twitter data without any further pre-processing steps.

Training and testing data: An overview of the datasets used for training language models and regression models is given in Figure 5.1 where the ‘labelled’ instances are instances which are associated with coordinates and they are used as a training set for the regression models while the ‘unlabelled’ instances are used for pre-training the RoBERTa domain-specific language model. For evaluation purposes we used a stratified split (80/10/10) for the Twitter dataset. We have obtained in total 146,651 labelled wildlife-related Tweets which is, to the best of our knowledge, the largest collection of geo-referenced wildlife-related Twitter data available.

5.1.6 Coordinates Normalization for performing regression

Neural networks do not perform well for numerical training labels that are unequally distributed [Scherrer et al., 2021]. Therefore, coordinate values need to be normalised before being used as an input to the neural network-based regression models [Scherrer et al., 2021]. For these purposes, we use the method proposed in [Scherrer et al., 2021] where authors use joint scaling and MAE loss function for normalising the coordinate values. The advantage of this method is that standard deviation is performed jointly on both latitude and longitude values, rather than independently for the two dimensions which can lead to data distortions.

5.1.7 Evaluation Metrics

We evaluate our models using standard measures used in previous related research on predicting coordinates for social media data [Zheng et al., 2018, Gritta et al., 2019]. These are Median Error Distance (MedianED) and Mean Error Distance (MeanED). The measures are defined in terms of the Distance Error *DE(m)*. For each tweet *m*, with a known actual location $loc_r(m), DE(m)$, is defined as either the Haversine distance or Euclidean distance *d* between $loc_r(m)$ and the inferred location, $loc(m)$: $DE(m) = d(loc(m), loc_r(m))$. The MeanED is defined as the average DE for each tweet while the MedianED is the median of DE for each Tweet.

5.2 Results

5.2.1 Evaluation Experiments

As mentioned in Section 2, our evaluation is focused on comparing a regression model based on the RoBERTa language model with two other regression models employed in previous research on geo-referencing social media data, one based on statistical machine learning algorithm (Support Vector Regression) and a regression model based on BERT. We have performed experiments with the pre-trained RoBERTa model, trained on a generic dataset and a RoBERTa model which has been fine-tuned to the Twitter domain using the Tweets we have collected related to wildlife observations (see Table 5.1). The baseline SVR classifier is based on TF-IDF frequencies on character grams of length 3-10. The other baseline based on BERT uses a pre-trained BERT language model, trained on the generic dataset and then fine-tuned for the regression task. Further, we present two approaches for improving the precision of georeferencing models based on location name extraction and semantic similarity between training and unlabelled instances. The development ('dev') set is used for identifying and saving the best performing model on the development set which is then used for assigning coordinates to the test instances. As mentioned in Section 5.1.7 we use MedianED and MeanED for evaluating the approaches.

5.2.2 Regression Results

Results from the performance of the regression models (see Table 5.2) showed that transformer-based models can have a significant advantage over traditional machine learning models for geo-referencing social media content. Both, the median error distance and the mean error distance are much lower even for the baseline BERT-based regression model when compared to Linear SVR model by a margin of more than 50km for 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). Further, the transformer-based regression models perform very similarly for both the dev and test set. This shows that the models generalise well for unseen datasets while the performance of the Linear SVR model drops significantly for the test set. For instance, for the best performing Linear SVR, the MedianED of the dev set is 98.60 km versus MedianED of the test set is 156.82 km which is 50 km increase in the distance error space. In contrast, the differences in the BERT and RoBERTa models' MedianED and MeanED values for the dev set and the test set is not more than 3 km.

Results also show that using RoBERTa for building regression models for geo-referencing Tweets is more beneficial than using BERT (see Table 5.2). Even when the generic RoBERTa model is used, it still outperforms the BERT regression model where both the MedianED and the MeanED are 20 km 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 shows that using larger transformer models for regression, even when trained on generic datasets, is highly beneficial for the performance of regression models especially when the labelled dataset is sparse.

The use of a RoBERTa word representation model fine-tuned to the domain data has led to further improvements over the pre-trained RoBERTa model with 1-2 kilometers decrease in MedianED and MeanED ('RoBERTa generic' versus 'RoBERTa wildlife', Table 5.2). In Table 5.2, 'generic' refers to a pre-trained publicly available language model which has been trained using generic online datasets, 'wildlife Tweets' refers to language model which has been fine-tuned to the domain data (wildlife-related Tweets), 'wildlife Tweets+combined training set' refers to a regression model which is using a RoBERTa model, fine-tuned to the domain and a training set, consisting of Twitter and Flickr data, 'NER + RoBERTa-based regression' refers to the hybrid approach consisting of location name extraction and regression, 'semantic similarity + RoBERTa-based regression' refers to the hybrid approach consisting of semantic similarities and regression, 'best single NER model' refers to using a single, i.e., the 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 spaCy NER library and Flert NER model. Notably the best performing regression model is using a RoBERTa model fine-tuned to the domain and also a training set consisting of Twitter and Flickr posts, resulting in

Regression Model	**Method**	**Dev Set**		**Test Set**	
		MedianED	**MeanED**	**MedianED**	**MeanED**
Linear SVR	TF-IDF	98.60 km	127.03 km	156.82 km	181.32 km
BERT	generic	93.63 km	119.34 km	94.90 km	121.37 km
RoBERTa	generic	38.30 km	102.09 km	40.96 km	101.35 km
	wildlife Tweets	37.99 km	101.50 km	39.84 km	100.89 km
	wildlife Tweets+combined training set	36.81 km	101.05 km	38.04 km	100.44 km
semantic similarity + RoBERTa-based regression	5 km	-	-	38.24 km	100.36 km
	10 km	-	-	38.16 km	100.17 km
	20 km	-	-	38.78 km	100.26 km
NER + RoBERTa-based regression	best single NER model	-	-	36.68 km	98.91 km
	voting mechanism	-	-	**36.47 km**	**98.22 km**

Table 5.2: Results from regression models performance

MedianED = 38.04 km and MeanED = 100.36 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.

5.2.3 Analysis on hybrid approaches

As mentioned in Section 2, we developed two approaches for improving the precision of regression models for geo-referencing Tweets.

The approach based on radial distances and semantic similarity (described in Section 5.1.4) does lead to slightly lower MeanED values compared to the best performing regression model, when 10 km radial distance is used (see Figure 5.4 and Table 5.2), though there is no improvement in MedianED. However, as illustrated in Figure 5.5, the hybrid approach based on semantic similarity is notable for performing particularly well when compared to the RoBERTa-based regression model at the highest locational accuracy band in which the error is less than 5 km.

In contrast, the hybrid location name extraction method that uses the RoBERTa regression method for disambiguation (described in Section 5.1.3) 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

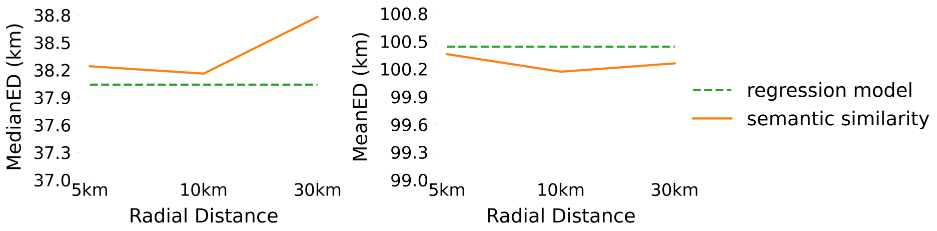


Figure 5.4: The effect of the radial distances over the performance of the Semantic Similarity+RoBERTa-based regression where the best performing RoBERTa-based regression model is used as a baseline.

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, spaCy NER library and Flert NER model, led to even further improvements with MedianED = 36.47 km and MeanED = 98.22 km. The error distribution results, presented in Figure 5.5, illustrate the fact that while the hybrid approach based on location name extraction and regression increases the precision of geo-referencing models significantly for all error distances up to 95km, the improvement is most marked within distances of 5km. Additionally, a comparison between the NER voting-based approach (which uses a gazetteer to obtain coordinates) and the RoBERTa-based regression model (see Table 5.3) showed that, just for those posts in which place names can be detected, the NER/gazetteer method using RoBERTa-based regression for location disambiguation (‘NER+regression-based disambiguation’) outperforms the purely transformer-based regression models and purely location name extraction method (‘NER+Nominatim-based disambiguation’) for geo-referencing social media posts, obtaining a median ED of 1.32 km. In Table 5.3 ‘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 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. In contrast, the regression-based disambiguation method does not lead to improvement over the disambiguation method based on restricting the Nominatim search to a specific region, i.e. UK (‘NER+Nominatim-based disambiguation with UK context’). This shows that combining location name extraction and regression approaches is beneficial for geo-referencing Tweets, especially when location searches cannot be limited to a specific context. However, using only the NER method with gazetteers coordinates has a major limitation for data sets such as the one employed here as only about 5% of the Tweets contain detectable place names (see Table 5.4). Our hybrid approach that uses coordinates both from gazetteers and predicted from regression is therefore clearly advantageous for such datasets.

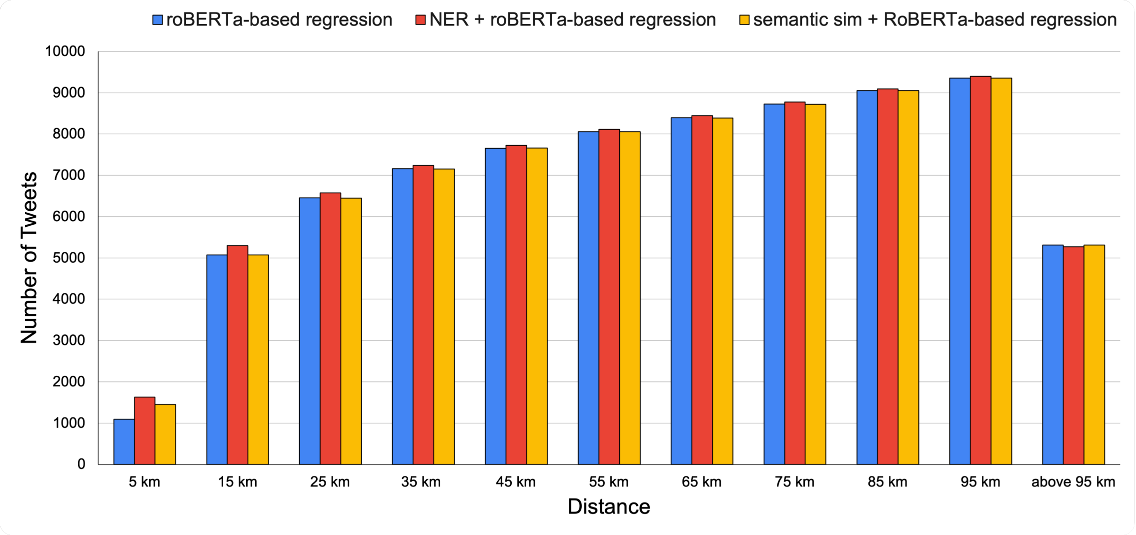


Figure 5.5: Distribution of error results showing proportion of results within 5 km, 10 km, etc for the approaches RoBERTa-based regression, NER+RoBERTa-based regression, and Semantic Similiarity + RoBERTa-based regression.

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

Table 5.3: 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.

Approach		#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

Table 5.4: Data distribution for the two hybrid approaches, i.e., semantic similarity approach , *'#Tweets'* refers to number of Tweets.

5.3 Discussion

5.3.1 Geo-referencing Methodology

The main stage of this research is the development of a multivariate regression model for predicting latitude and longitude values for a given unlabelled Tweet. For these purposes, we used the language model RoBERTa which has achieved state-of-the-art performance for many NLP-related tasks. RoBERTa's architecture already supports fine-tuning on regression, however it allows predictions only for a single value. Thus, we extended the implementation of RoBERTa's model, provided in Huggingface Simple Transformers library, to support multivariate regression by calculating Mean Square

Loss (MSL) function for each label (i.e., latitude and longitude) separately. In similar work by Scherrer et al. [2021], the authors adapt BERT to multivariate regression. We extend on this work by performing experiments with RoBERTa which is a bigger and newer language model and also outperforms BERT for many tasks. Additionally, we performed analysis with different transfer learning techniques to identify strategies suitable for geo-referencing Tweets when there is a limited amount of labelled data. These are: 1) a comparison between domain-trained and pre-trained language models over the performance of regression and 2) augmenting the training corpus using Flickr data.

We proposed two hybrid approaches, both incorporating RoBERTa-based regression model which help improve precision of geo-referencing Tweets. These approaches are based on 1) location name extraction and 2) semantic similarity and radial distances between the training and test instances.

We evaluated the proposed approaches using two baselines — a widely used statistical regression model (SVR) and a regression model based on BERT.

The multi-output regression model based on RoBERTa and the hybrid approaches incorporating regression have not been used in previous research and represent a novel approach for geo-referencing Tweets. Additionally, the domain-trained RoBERTa model proved valuable for building geo-referencing models and can be re-used for future work on classification and regression for wildlife and Twitter-related research.

5.3.2 Findings

1. Analyses presented in Sections 5.2.2 and 5.2.3 showed that transformer-based models adapted for multivariate regression have a significant advantage over a conventional SVR statistical regression model for assigning coordinate values to social media posts. Further, a comparison between two (generic) pre-trained transformer-based models, BERT and RoBERTa, showed a significant advantage

of the RoBERTa model for geo-referencing Tweets, indicating that using larger pre-trained models for geo-referencing is beneficial when the domain-specific training dataset is relatively small.

2. Analysis on transfer learning techniques showed that fine-tuning language models to the domain and then to the task helps improve the performance when compared with generic trained models even when the corpus used for fine-tuning is relatively small. Further, the domain-trained RoBERTa language model developed in this work can be used in future research on regression and classification tasks for Twitter and in wildlife-related studies. A related work by Scherrer et al. [2021] presented the geoBERT model, where BERT was adapted to the regression task for geo-referencing social media posts. In contrast, we perform analysis with the more recent and bigger language model RoBERTa, which we have adapted for multivariate regression. We also experiment with different transfer learning techniques and hybrid methods combining a rule-based NER and gazetteer approach with regression strategies to improve the precision of geo-referencing approaches. We compared our model to the BERT-based regression model which is a state-of-the-art geo-referencing approach, i.e., geo-BERT Scherrer et al. [2021]. The results presented in Table 5.2 shows a clear advantage of the RoBERTa-based model over geoBERT. Furthermore, the RoBERTa model trained on wildlife Tweets can be used for future research in georeferencing and analysing wildlife-related social media posts. A recent deep learning method that claims to be state of art and emphasises the issue of disambiguation (which is relevant here) Yan et al. [2021] reported significantly poorer median distance performance compared to our approach (97km vs 34km). The only published method that we are aware of that has better performance for georeferencing Twitter Di Rocco et al. [2021] obtains that performance by restricting the problem to a specific local geographic region and is entirely dependent upon the presence of place names in the postings, neither of which are appropriate for the problem addressed here.

3. The approach for augmenting small training sets using additional social media sources proved beneficial for training geo-referencing models. This work can be expanded in future by considering a larger number of social media platforms and using these diverse data sources for pre-training the language models. Related work by De Rouck et al. [2011], Laere et al. [2014a], analysing the benefits of using language models built with Flickr and Twitter datasets for the prediction of Wikipedia page location, used statistical machine learning methods. Their approach required a pre-processing step to normalise feature vectors used by the statistical algorithms, in contrast to our neural network-based georeferencing model for which no such pre-processing is required.

4. The hybrid approaches presented in Sections 5.1.3, 5.3 and 5.2.3 showed that using either semantic similarities or location name extraction combined with regression, which is used for disambiguation of place names when present and in isolation when they are not, can be beneficial for improving the precision of geo-referencing models. Specifically, the approach based on location name extraction and RoBERTa-based regression leads to the best results on geo-referencing Tweets and it helps enhance precision particularly in obtaining distance errors less than 5 km. A significant drawback of the NER/gazetteer approach when used in isolation is that only a small portion of Tweets include place names. In contrast, the transformer language model-based regression methods always assign coordinate values to the Tweets. An investigation of the Tweets for which NER did not return location names showed that 32% of them were predicted within 15km by the regression model and 51% were predicted within 55 km. Examples of Tweets that were geo-referenced by the regression method but not by the location name extraction method are given in Table 5.5. An obvious characteristic of Tweets that cannot be geo-referenced with NER and gazetteers is that there are no detectable place names. Examples include: *'This Herring gull was harassing returning guillemots to give up their catch. #wildlife...* `https://t.co/TFm1Whtc03`*'*, *'@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'*. They often include fine-grained locations using generic place words such as 'our bird feeders' and 'lawn' which 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 5.5) 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 miss-classified by the NER approaches, for example because of adjectives attached to the proper location names, as in *Lovely #daffodils @ Sunny Adlington* `https://t.co/p5dqjjBYVU`, where the NER methods have labelled the phrase 'Sunny Adlington' as a person. Another reason for failure of the NER/gazetteer methods is 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 5.5 example *'@Staffsbirdnews Uttoxeter Quarry: Common Tern,Common Sand, 4 Green Sand, 4 Snipe, 3 Pintail, 19Wigeon, 4 Pochard and 2 Blue Snow Geese'*. 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. The significant advantage of the regression model is that it is able to assign coordinates to such Tweets (especially those that do not mention gazetteered place names) based on learned trends from the training set. The error analysis presented in Table 5.5 shows the benefits of using regression in combination with NER methods which help improve precision with less than 5 km error in some cases. Such a hybrid approach can be very beneficial for georeferencing diverse collections of social media posts, independent of the observed species, even those that do not mention place names. Finally, the comparison between different location disambiguation methods presented in

Table 5.3, shows that using regression and NER instead of the other more widely used methods for location disambiguation can be highly beneficial for building georeferencing models.

Tweet	Dist. Error(km)
13 spoonbills and one with a avocet sitting on ones head @RSPBtitchwellmarsh http://t.co/asCe6L8UEP	4.00km
Morning all. Yes indeed, it's a marshmallow world again round here. Deep joy. And pity me poor Robin; Blackbird on their nests!	2.69km
Great Black Backed Gull spotted on 09-Jul-2013. Sent from Birds of Britain HD app by @CleverMatrix https://t.co/BBa2zrtR86	2.71km
@Staffsbirdnews Uttoxeter Quarry: Common Tern, Common Sand, 4 Green Sand, 4 Snipe, 3 Pintail, 19 Wigeon, 4 Pochard and 2 Blue Snow Geese	3.89km
What beauty, Buddleja and a Peacock butterfly! #buddleja #buddleia #butterflybush #peacockbutterfly #beauty #nature #garden #betwsycoed	4.72km
@Staffsbirdnews Uttoxeter Quarry: Redstart, Black-tailed Godwit, 3 Green Sand, 6 Common Sand, 5 LRP, Willow Tit	1.18km
Tiny bee type thingy on my pink daisy #beetypething #tinybee #pinkdaisy #daisy #pink #gardening #gardensofinstagram #lbloggers #lbloggersuk #instagarden #growyourown #plants #plantsofinstagram #gbloggersuk `https://t.co/IofJdMOyUa`	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	3.56km
#wmbirdclub #Belvide 12/10: 68 Golden; 3 Ringed Plover, Ruff, 8 Dunlin, 40 Gadwall, 27 Shoveler, 14 Wigeon, 163 Teal; 55 Pochard.	0.43km

Table 5.5: Examples of Tweets for which the regression model performed well, but the NER/gazetteer location extraction-based approach failed.

5.3.3 Limitations

The research presented in this chapter can be extended by performing further analysis using transfer learning techniques. We want to analyse whether the use of language models trained using diverse social media datasets (such as Flickr) can further improve the performance of geo-referencing Tweets.

5.4 Conclusions

This chapter presented novel work on adapting state-of-the-art transformer-based models such as RoBERTa to multivariate regression for creating geo-referencing models for social media posts. We performed analysis with various transfer learning techniques for improving the performance of regression models focusing on scenarios with a small training corpus. Question **RQ 3** from the hypothesis presented in Section 1.2 has been answered showing that domain-trained transformer models, fine-tuned for multivariate regression and using diverse social media sources for augmenting the training set with additional labelled data improve precision of geo-referencing models. Further, we provide the largest collection of geo-referenced wildlife Tweets and a domain-trained RoBERTa model which can be used in future research on geo-referencing and identifying wildlife observations on social media. Finally, we proposed a hybrid approach based on location name extraction and RoBERTa-based multivariate regression which help significantly improve the precision of geo-referencing social media posts. The work in this chapter provided useful insight into how state-of-the-art neural network models, transfer learning techniques and simpler rule-based approaches can be combined to provide less data consuming geo-referencing models.

Chapter 6

Extracting Geometric Representations Of Trajectories

In the previous chapter, we addressed the need for developing less data consuming georeferencing models for assigning coordinates to social media posts. In this chapter, we build on this research by focusing on trajectory extraction for objects whose location has already been identified. The method presented in this chapter can be applied to extract trajectory information of individual species observations from social media datasets that have been verified as genuine wildlife observations and have had coordinates attached using the methods presented in Chapters 3 and 4.

As discussed in Section 2.5, identifying patterns of movement finds applications in many domains such as climate science and studying species migration patterns. However, a main challenge of tracking objects is that their topological characteristics can change over time, such as splitting an object into multiple objects or the merging of multiple objects. The authors of Corcoran and Jones [2017, 2018] address this problem using spatio-temporal analysis based on zig-zag homology and persistence landscapes (discussed further in Section 2.5). The proposed approach helps identify moving objects over time and it has been used successfully for identifying fish swarms and cloud movement. Further, the approach facilitates the use of statistical and data mining techniques for the identified objects. This makes them suitable for further studies into identifying trajectories of movement patterns. However, such studies have not been conducted. This is also our motivation to build on the approaches presented by

Corcoran and Jones [2018] and illustrate how the topological data analysis methods can be used to extract trajectories of the identified objects. We create trajectories by calculating centroids of object's regions at each time slice and then connecting the centroids. We also perform trajectory clustering and normalisation to identify similar trajectories and facilitate the discovery of movement patterns for future work. Current research on trajectory clustering is limited at exploring algorithms (described further in Section 2.5.1) which do not take into account both spatial and temporal characteristics of the data. We address this research gap by exploring the QuickBundle (QB) algorithm [Garyfallidis et al., 2012] which has not been studied before in trajectory mining. However, it can be easily adapted to trajectory data and supports comparison for sequences, i.e., takes into account the order of the data points.

In the rest of the chapter, the final question **RQ 4: Do zig zag persistent homology methods have good potential for extracting trajectories of spatio-temporal objects?** in Section 1.2 has been answered. Contributions made as part of this research include a methodology for extracting and normalising geometric representations of trajectories for tracking spatio-temporal phenomena, and analysis into a less explored but potentially promising algorithm for clustering trajectories.

6.1 Dataset

We apply the methods to weather data, specifically tracking rainfall in radar imagery. We use the same dataset obtained by Corcoran and Jones [2018]. The images are gathered from the UK Meteorological (Met) Office[1]. The Met Office provides this data at 15 minute intervals. For a given time, the image data in question categorises the rainfall level at each location in a 500x500 regular grid over Ireland and UK. Given this data, we consider the problem of tracking objects corresponding to spatially close path-connected components of $\mathcal{R}^2$ with a rainfall level greater than a given threshold

[1] UK Met Office `https://www.metoffice.gov.uk/`

[Corcoran and Jones, 2018].

6.2 Methodology

The methodology uses the objects' locations produced by the application of Corcoran and Jones [2018] as an input. It consists of 2D array representations of cloud images, where the two indices represent the *x* and *y* pixel coordinates while the value of each array element is the unique identifier of the respective pixel at that location. All pixels that belong to the same object (i.e. cloud region) will have the same unique identifier, where that identifier serves simply to distinguish the objects from each other. Each object persists across a consecutive sequence of time slices starting at time slice T_1 and finishing at time slice T_n. These persistence intervals are provided as output from the zig-zag persistent homology procedure. An overview of the methodology is given in Figure 6.1. The methodology consists of 5 main steps. In *Step 1* we identify objects present for each time slice. We do this by finding all the pixels that belong to the same object for a given time slice. We also remove pixels with no objects in them (i.e., they have value equals '0'). At the end of this step, each object id is associated with a list of all the *x* and *y* coordinates of the pixels that belong to this object for a given time slice.

In *Step 2* we identify the location of each object for a given time slice. We approximate the location of the objects by finding their centroids. Each object is a finite set R of n elements. Thus, the centroid is the mean of the elements in the set R (see Equation 1). At the end of this step, we have each object associated with the *x* and *y* coordinates of the centroids of this object for each time slice. In *Step 3* we accumulate an object's locations across its time slices. Specifically, we accumulate the sequence of centroid coordinates for each unique object across its time slices. In *Step 4* we produce a set of trajectories and visualize them.

In *Step 5* we perform clustering of the object trajectories. Our goal is to create clusters containing similar trajectories in order to identify movement patterns. We perform trajectory clustering by adapting the QuickBundle (QB) algorithm [Garyfallidis et al., 2012], originally created for use in magnetic resonance imaging to cluster white matter fibres. Each QB cluster can be represented by a single centroid streamline, which is a sequence of points. We selected this algorithm for the following reasons:

1. The algorithm has been specifically created for data from 3D imagery of white tissues, which resembles the structure of trajectories. The algorithm was created for simplifying tractography data of white tissues where tractography data is a dataset composed of streamlines.
2. QuickBundle uses a symmetric distance function called minimum average direct-flip (MDF) distance [Garyfallidis et al., 2010, Visser et al., 2011] which takes into account the sequential nature of streamlines. This makes the algorithm suitable for clustering trajectory data where each point represents an object location at a given time and it is important to preserve the time sequence of the objects. In contrast, other distance measures, incorporated by widely used clustering algorithms such as DBSCAN and K-centroid treat streamlines as a bag of points where every point on the first streamline is to be compared with every point on the second streamline, and vice versa.
3. The algorithm has been created to deal with large datasets and return results fast which makes it efficient for analysing large volumes of ecology-related data in real time.

The number of clusters produced depends on adjusting a threshold value. High threshold values will produce less clusters with more trajectories in them while a small threshold value will produce smaller clusters. Before clustering, we defined a distance function between the trajectories using Euclidean distance [Danielsson, 1980]. We chose Euclidean distance because of its suitability for measuring distances between objects, simplicity, and lack of a threshold that needs adjusting. The MDF measure requires trajectories to have the same length. Therefore, we re-sample the trajectories to have the same number of points. This is achieved using linear interpolation.

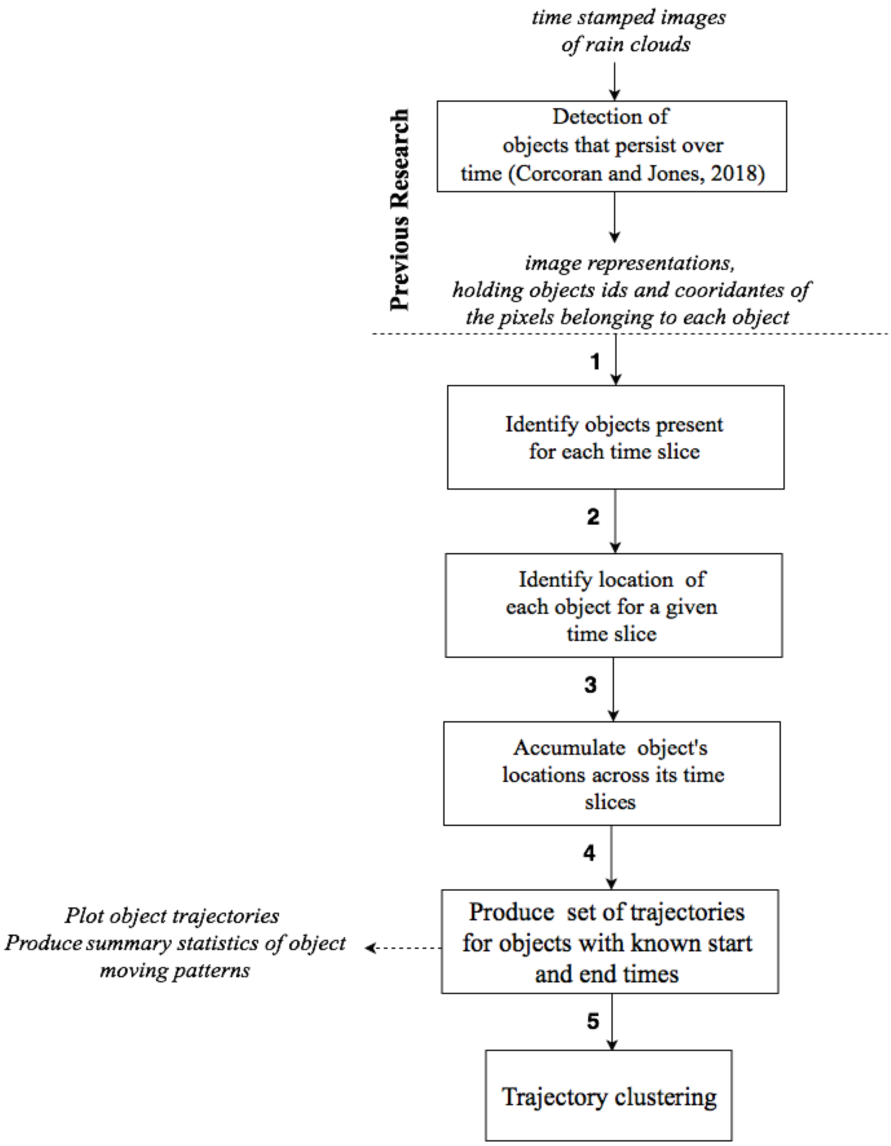


Figure 6.1: Overview of Trajectory Extraction Method

$$Object_[centroid] = \frac{\sum_{i=1}^{n} R_i}{n} \qquad \text{Equation 1}$$

6.3 Results

As mentioned in Section 6.1, the methods were applied to meteorological data collected over a 12 hour period. Cloud images were recorded at 15 minute intervals, thus we have 48 time slices. After extracting the objects from these images with respect to time, we obtained six objects. We identify objects, as indicated above, using the unique identifier of the pixels that belong to the same objects from processed imagery data. In Table 6.1 we give summary statistics of the existence of the objects over time.

Object ID	**Start time**	**End time**
Object 1	0	47
Object 2	0	47
Object 3	0	47
Object 4	11	37
Object 5	11	46
Object 6	22	22

Table 6.1: Summary statistics of objects existence over time

Object 6 appears in only one time slice (see Table 6.1). Thus, it is not visible in Figure 6.2 and Figure 6.3 except as a dot. Objects 1 - 3 persist over all 48 time slices while Objects 4 and 5 were formed at time slice 11 and are destroyed at time slices 37 and 46.

6.3.1 Trajectories representation

Figures 6.2 and 6.3 represent the trajectories of the moving objects, where Figure 6.2 provides linear representation and Figure 6.3 shows trajectory changes over time.

The trajectories of the objects (see Figure 6.3) appear to be subject to some very sudden changes of location which is unusual for weather data. This occurs here as a consequence of objects merging and splitting between time slices.

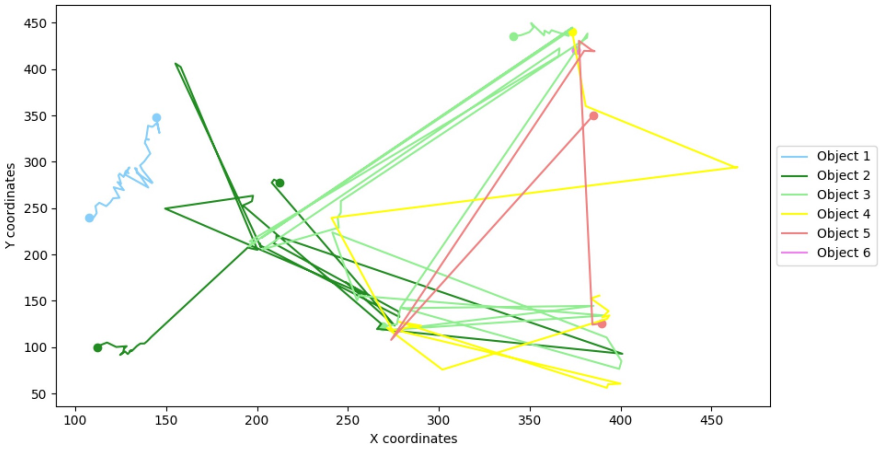


Figure 6.2: Cloud movement trajectories representation — Line representation

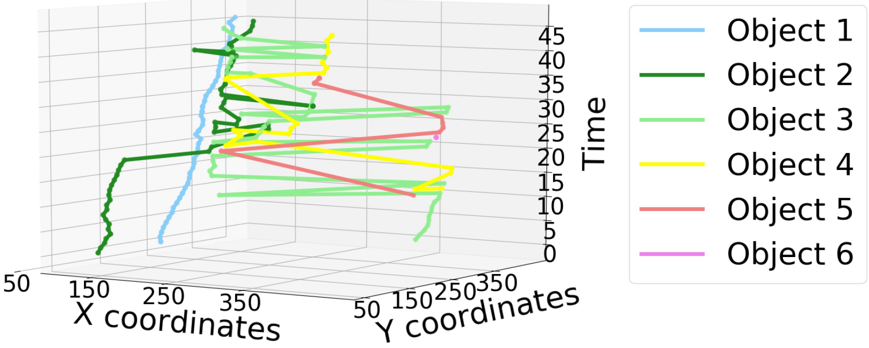


Figure 6.3: Cloud movement trajectories representation — Space time cube representation.

This is demonstrated in Figure 6.5 where we display object movements over three time slices. When large objects become connected it results in a dramatic change in the centroid of the merged object. Figure 6.5 demonstrates this with the dark purple and orange objects which merge in time slice 38. Once an object is merged into another object or it disappears in a time slice, it is destroyed. An object cannot re-appear once it has been destroyed, according to the persistence homology approach used for identifying objects over time. These sudden changes in the objects movement patterns are also illustrated in their trajectories (see Figure 6.4) where the trajectory of Object 4 (orange) has a sudden change of a direction (when it merges with the purple object) while the trajectories of Objects 1 and 2 follow almost a straight line as they do not experience drastic changes.

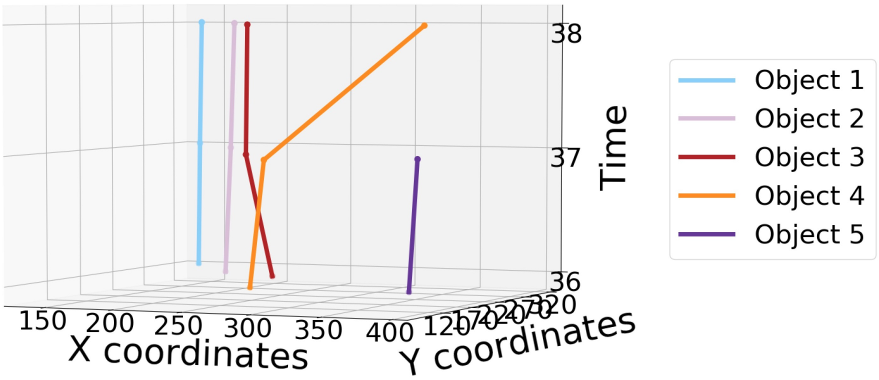


Figure 6.4: Cloud movement trajectories representation

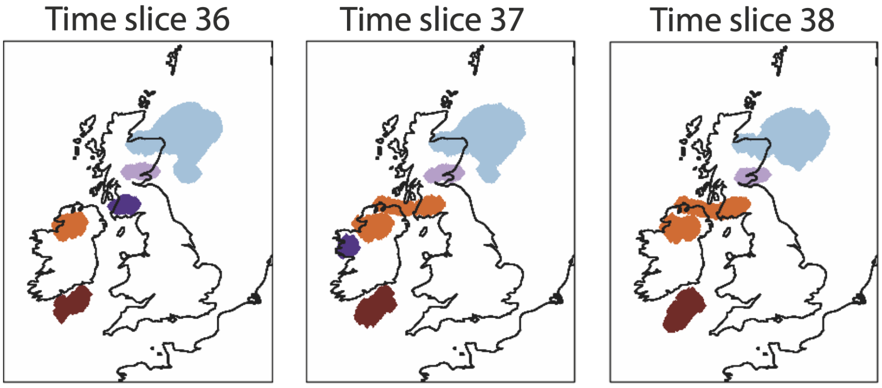


Figure 6.5: Objects movement per time slice — Time slice 36, Time slice 37, Time slice 38.

6.3.2 Trajectories clusters

In order to limit the noise in the data and identify similar movement patterns, we performed clustering of the object's trajectories (see Figure 6.6) using the QuickBundle (QB) algorithm, presented in Section 6.2. We performed experiments with different number of clusters, i.e., 2, 3, 4, and 5. However, results showed that 3 is the most optimal number of clusters for the dataset. A summary of the grouped trajectories is given in Table 6.2 and Figure 6.6 where all object trajectories that belong to the same cluster have the same colour. The cluster analysis helped identify that Objects 3, 4, 5, and 6 have similar movement patterns in contrast to Objects 1 and 2 which can be considered outliers. This analysis can help identify different types of weather phenomena.

Cluster ID	Trajectory ID
1	1
2	2
3	3,4,5,6

Table 6.2: Summary statistics of objects existence over time

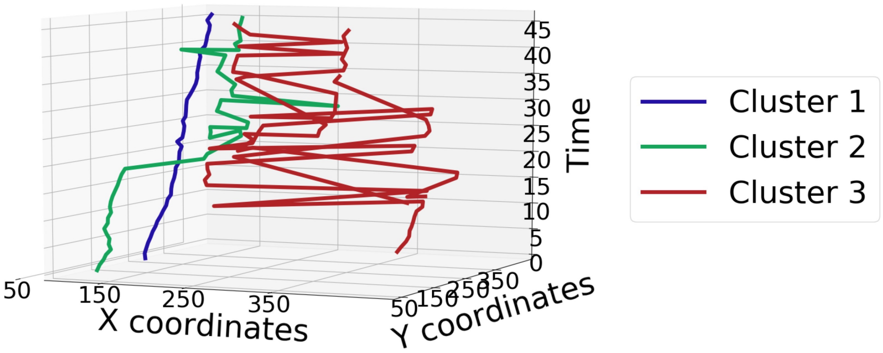


Figure 6.6: Cloud movement trajectories represented with a space time cube

Normalising Trajectories As discussed in Section 6.3.1, sudden merges or splits of the objects result in drastic changes in the trajectory locations. This can potentially cause problems when trajectories need to be used to identify direction of movements for the given objects. In order to resolve this problem, we have performed normalisation of the trajectories.

Figure 6.7 represents the normalised trajectory of Object 1 while Figure 6.8 presents the normalised trajectories for all objects. These facilitate easier identification of direction of movement which can find applications in various research in invasive species and weather phenomena tracking.

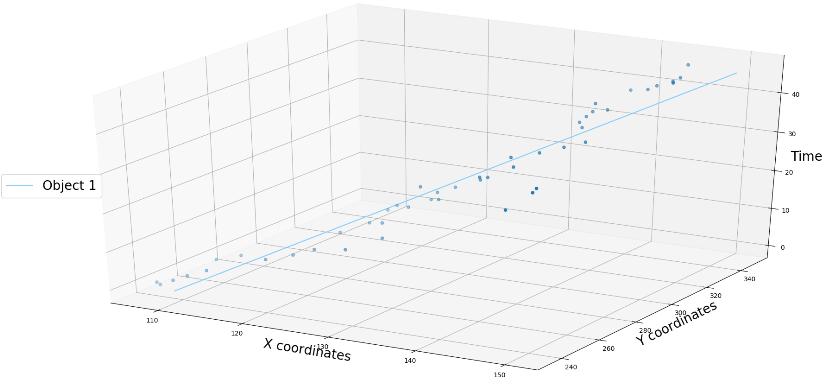


Figure 6.7: Normalized trajectory for Object 1 with original scatter points

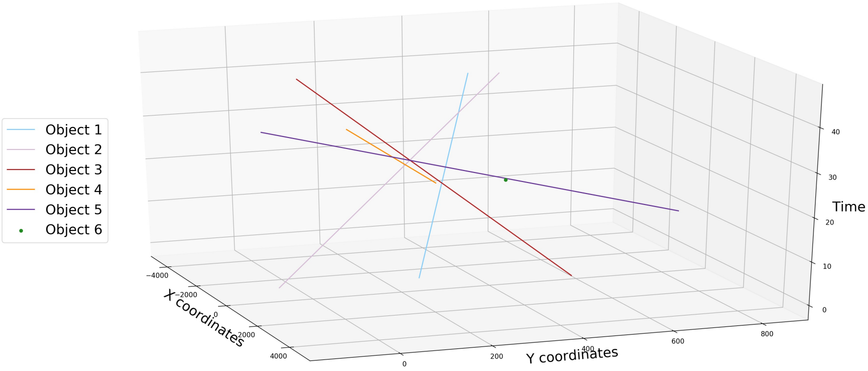


Figure 6.8: Normalized trajectory for all objects

6.4 Discussion

6.4.1 Trajectory Extraction Methodology

We developed a trajectory extraction methodology which extends on previous work for object location identification based on zig-zag homology. We create trajectories by calculating centroids of object's regions at each time slice and then connecting the centroids. The methodology also encompasses clustering and smoothing of the trajectories in order to facilitate observation of movement patterns and identification of direction of movement.

The work presents an extension to the methods built by Corcoran and Jones [2018] for identifying moving objects over time by showing how these methods can be incorporated in further analysis to support object tracking and identification of movement patterns.

The trajectory extraction and clustering methods have been evaluated for the case study of tracking rain clouds in imagery data.

6.4.2 Findings

The main findings from this chapter are summarised below:

1. Spatio-temporal methods based on zig-zag homology for identifying objects persistent over time can be extended to support trajectory extraction. The proposed trajectory extraction method can be valuable in observing and discovering object's movement patterns.

2. The clustering algorithm, QuickBundle, which was created for data from 3D imagery of white tissues is suitable for normalising trajectory data. It did help reduce noise and identify similar trajectories of movement between the objects. Similar movement patterns can suggest similar characteristics between the observed objects.

3. Trajectories from weather-related data include sudden changes in location due to splitting and merging of objects. This may introduce noise in trajectory data and require the use of smoothing techniques to help identify patterns and direction of movement.

6.5 Conclusions

This chapter presented a methodology for trajectory extraction which builds on existing topological data analysis methods for identifying objects with dynamic topology. We also performed trajectory clustering to identify similar trajectories and facilitate the discovery of movement patterns. Specifically, we explored the clustering algorithm QuickBundle which has not been used for trajectory data. However, it proved beneficial for trajectory data, even though it has been created for different purposes.

The final research question **RQ 4** from the initial hypothesis was answered in order to show that a trajectory extraction method which uses as input object location information derived from spatio-temporal analysis based on zig-zag homology can facilitate accurate depiction of objects movement patterns.

The analysis presented in this chapter has been performed for tracking rainfall in radar imagery. In future, we want to extend the work for different case studies. For instance, we want to apply the described methods to the wildlife-related social media datasets collected for the previous chapters of the thesis. Additionally, a natural step after performing normalisation of trajectories, would be to build on the developed methods in this chapter to support identification of direction of movement for the observed objects.

Chapter 7

Conclusions and Future Work

In this thesis, our aim was to identify social media mining methods that facilitate the usage of social media datasets as an unofficial source of wildlife observational data. We used Flickr and Twitter as exemplary social media platforms due to their wide usage and relatively open access for research. Further, the two social media sites differ in the nature of the data shared (i.e. images versus short texts) which gives a platform for more extensive analysis on verification methods. We provide the largest collections of geo-referenced wildlife-related Tweets and Flickr posts to the best of our knowledge. These can be used in further studies related to social media mining and wildlife. In addition, we used the NBN citizen science data portal as a gold standard for our initial analysis.

In the initial chapters, we investigated the potential of social media datasets to supplement official citizen science data portals. This study involved statistical, spatial, and temporal analysis which revealed the potential of image-sharing platforms to provide valuable wildlife data related to certain species taxonomic groups. Additionally, we presented two verification approaches for identifying genuine wildlife observations, one suitable for verifying images and the other, text. The image-based verification approach consisted of using an image recognition tool and the species taxonomic data to identify whether a species name given on Flickr represents a given species. Analysis showed that a class-level coarse match (between all species names following down from the species class and the labels returned by the image recognition tool for a given image) can be highly beneficial for identifying large and diverse species collections without affecting the precision of the method. Regarding the text-based verification, we analysed three text classification approaches and various statistical and neural network-based feature extraction and feature integration techniques in order to identify the methods which are most suitable for verifying social media posts (for example Tweets) related to wildlife observations. This research showed the potential of transformer-based models to be used for text classification tasks even when there is a limited amount of training data. In the next stage of the work, we focused on exploring geo-referencing strategies which can support accurate assignment of coordinates to social media posts even when there is a limited amount of training data. In this way, we want to support the use of social media datasets in wildlife observational studies. Motivated by the results from the text verification approach, we used transformer-based models, fine-tuned for multivariate regression to assign coordinate values to given unlabelled social media posts. Specifically, we performed analysis with state-of-the-art contextual word representations, transfer learning techniques, and rule-based methods, in order to build less data consuming georeferencing models. Findings from this research showed that adapting transformer models to the domain and then fine-tuning them for regression help improve the performance of geo-referencing models, especially when the training set has been augmented using labelled instances from multiple social media networks. Additionally, a hybrid approach, consisting of location name extraction and the neural network-based regression model led to significant improvements in the precision of geo-referencing approaches. In a later chapter, we proposed a methodology for trajectory extraction and normalisation which extends on spatio-temporal methods using zig zag homology for identifying objects locations which persist over time.

In the rest of this chapter we provide an overview and assessment of the work conducted in this thesis. We also discuss how the research presented in the thesis can be taken further in potential future projects. Finally, an overview of the thesis in terms of its contributions is described.

7.1 Analysis of Research and Results

In this thesis, the research behind establishing methodologies for analysing, verifying, and preparing social media data to be used to support wildlife-related studies has been described. In the following sections, we analyse the research that was carried out in the primary chapters of this thesis.

7.1.1 Suitability of Social Media as a Supplement to Citizen Science Portals

We conducted large scale and extensive analysis on the suitability of social media data to supplement official citizen science data portal collections. This work helped establish potential usages of social media networks for providing observational data for wildlife- and ecology- related studies. Specifically, we evaluated the species distributions on the image-sharing platform Flickr compared to the largest UK citizen science portal NBN, including the 1500 best represented species on NBN and invasive species, common for UK. This makes the research the most extensive work on suitability of social media resource for providing wildlife observational data conducted to date. We performed three types of analysis, statistical, spatial, and temporal considering different experimental settings. The analysis showed that Flickr can be a rich source of species observational data for certain taxonomic groups (diurnal garden birds and pretty flowers) and as a data source for dedicated projects. The results from the temporal analysis showed that Flickr might not be suitable for performing studies using historical data on yearly, half-yearly or seasonal species movement patterns. However, they do indicate that Flickr can be highly beneficial for providing real time analyses of movements, especially for observing invasive species.

The highlight of this work was the creation of a fully automated image verification method suitable for verifying large and diverse collections of images related to wildlife. The approach is based on using the Google Cloud Vision API in combination with the species taxonomic data to identify the likelihood that a species name on Flickr represents a given species. Specifically, the approach consists of coarse matching between all species names following down from the species class and the labels returned by Google Cloud API for a given image. The coarse match helps avoid high numbers of false negatives for less known species or similar species (12-spot and 7-spot ladybird). We compare the class-level matching approach with species-level and genus-level based approaches to identify what method is suitable for verifying large and diverse collections of image data without affecting the precision of the method. In contrast, previous image verification approaches for wildlife data [August et al., 2020, Daume, 2016, ElQadi et al., 2017, Barve, 2014, Skreta et al., 2020] involve manual or semi-automatic analysis, limited to verification of a small range of species or taxonomic groups, or require large amounts of labelled wildlife-related images in order to train an image classifier.

7.1.2 Text Classification for Verifying Social Media Relevant to Wildlife

After creating a methodology for verifying images related to wildlife, we focused on building a fully automated text classification model for identifying genuine wildlife observations on Twitter. We chose to focus on Twitter for this work as that platform remains very widely used, much more so than Flickr, and therefore has great potential (alongside Flickr) for recording wildlife observations in a timely manner. This is a challenging problem however, as often postings use the common names of species in contexts very different from wildlife observations, and Tweets, unlike Flickr postings, frequently use very informal language. Further, the literature survey on text classification for social media and wildlife-related data revealed a lack of extensive analysis into suitability of classification approaches for small collections of training data that is also related to wildlife observations and thus can have more specialised terminology. Additionally, state-of-the-art transformer-based models have not been fully explored in wildlife-related studies.

Three classification approaches are compared reflecting on some of the main types of existing machine learning algorithms. These are: logistic regression classification coupled with various forms of input features; the word embeddings based fastText pipeline; and the transformer-based model of BERT. We performed experiments with pre-trained and corpus-trained embeddings as well as different methods for building feature vectors. This research provided wider understanding of the type of feature extraction, feature integration, and machine learning algorithms suitable for building verification models for Twitter in the presence of limited amount of training data. Findings showed that the BERT model fine-tuned to the task and adding a sequence classification layer is suitable for building verification models for Twitter even in the presence of limited amounts of labelled instances. The high performance of the fine-tuned BERT classifier shows the potential of state-of-the-art deep learning models to be used for identifying valuable ecology data among informal social network sources automatically and on a larger scale, independent of the species observed at hand. Finally, analysis into the use of hashtags, mentions, and URL links in wildlife-related Tweets showed that there is a trend in the usage of hashtags related to wildlife which are unrelated to official campaigns for gathering wildlife data. In future, such hashtags could be used by informal social network campaigns to encourage people to indicate when they are posting about wildlife.

7.1.3 Geo-referencing Social Media Data Related to Wildlife Observations

After developing verification methods for identifying wildlife observations on Flickr and Twitter, we focus on the problem of geo-referencing Twitter posts because wildlife- and ecology-related studies require the presence of coordinate data to support study of species distribution. The development of effective georeferencing methods for social media wildlife observations is of considerable significance as the great majority of postings have no useful coordinates, but coordinates are essential for purposes of monitoring species occurrences. Our aim was the identification of strategies which can facilitate more accurate coordinate prediction even when training data is scarce. Thus, the conclusions from the work on verification techniques motivated further analysis into using transformer-based models and transfer learning techniques for geo-referencing Tweets. For these purposes, we adapted the transformer-based model of RoBERTa for multivariate regression for predicting latitude and longitude values. We performed experiments with various transfer learning techniques showing that adapting contextualised word models to the domain and then fine-tuning them for regression help improve the performance of geo-referencing models. Further, augmenting training data, consisting of Tweets, with Flickr textual data proved beneficial for the better performance of geo-referencing approaches. This method is similar to the one presented by Laere et al. [2014b], De Rouck et al. [2011] where the authors augment training datasets for geo-referencing models using multiple social media networks. However, they used classical machine learning algorithms which require pre-processing of the heterogeneous data sources. Instead, we use a transformer-based model which does not require pre-processing. Additionally, we proposed two hybrid approaches for improving the precision of geo-referencing models. The first approach is based on location name extraction based on NER methods and it uses the RoBERTa-based regression model to perform location name disambiguation. The second approach is based on semantic similarity-based methods where we select the most similar training instance which is within a radial distance to a given test instance based on the coordinates given by the RoBERTa regression model. Then, the test instance is given the coordinates of the most similar training post. Both approaches showed improvements in precision for geo-referencing Tweets. However, location name extraction combined with regression led to the best results on geo-referencing social media posts and it helped enhance precision for distances shorter than 5km. Additionally, a comparison between the NER-based approach (which uses a gazetteer to obtain coordinates) and the RoBERTa-based regression model (see Table 3) showed that, just for those posts in which place names can be detected, the NER/gazetteer method using RoBERTa-based regression for location disambiguation outperforms the purely transformer-based regression models and the purely location name extraction method for geo-referencing social media posts. This shows that combining location name extraction and regression approaches is beneficial for geo-referencing Tweets.

We evaluated the proposed approaches using two baselines, a widely used statistical regression model (SVR) and a regression model based on BERT, similar to the one presented by Scherrer et al. [2021]. The approaches proposed in this work outperformed the baselines by a significant margin.

Previous work on geo-referencing social media posts is mainly limited at presenting data consuming approaches or using statistical machine learning algorithms. Most of the research is also focused on using location-specific language models which tend to be data-specific and require extra steps for assigning actual coordinates to the test instances. We address these issues by using regression-based approach which do not require partitioning of the data into regions and do not require additional steps for assigning coordinates. Further, we perform extensive analysis with state-of-the-art contextual models, transfer learning techniques, and rule-based approaches in order to build less data consuming georeferencing models. The studies in this chapter showed that combining regression with NER methods can be highly beneficial for improving the precision of geo-referencing models when we have a limited amount of training data. The two approaches complement each other well. The regression models always return a pair of coordinates for a given Tweet, even when it does not contain a location name while the NER approach is very precise for Tweets including place names. The results achieved with the georeferencing model presented in Chapter 5 are comparable to other published work Scherrer et al. [2021]. However, we develop our methods using a small amount of training data which is an unexplored problem in georeferencing. All this shows that combining state-of-the-art transformer-based models with rule-based methods is a promising research avenue which is worth exploring further especially for geo-referencing wildlife-related Tweets.

Our work is similar to the work presented by Scherrer et al. [2021] which adapted BERT for multivariate regression and explored the effect of various pre-trained and domain-trained BERT models over the prediction task. We build on this research by using a newer and more efficient language model in combination with the techniques described above. We also perform a more extensive analysis.

7.1.4 Extracting Geometric Representations Of Trajectories

After establishing verification and geo-referencing approaches suitable for wildlife-related social media data, we focused on establishing methodologies for trajectory extraction and clustering to support tracking of spatio-temporal phenomena. We build on methods developed by Corcoran and Jones [2018] that used zig-zag homology for identifying moving objects over time. Our trajectory extraction method takes as an input the locations of objects obtained using the methods described in Corcoran and Jones [2018]. Then, we created trajectories by calculating centroids of object's regions at each time slice before connecting the centroids. We also performed clustering to support the easier identification of movement patterns. For these purposes, we used QuickBundle algorithm which was created initially for data from 3D imagery of white tissues. The algorithm has not been applied to trajectory data before, though trajectory data resemble the structure of white tissues. Further, the algorithm is based on a minimum average direct-flip (MDF) distance function which takes into account the sequential nature of streamlines. This makes the algorithm suitable for trajectory data where the sequence in which data points are processed is important as they represent locations of objects at given times. We used cloud imagery data for performing the analysis. However, the methodology can be applied to different types of datasets.

7.2 Contributions

Throughout the earlier chapters of this thesis, work has been conducted towards answering the questions posed in Section 1.2 in the Introduction.

In particular, the questions are now answered more formally.

- **RQ 1: Can social media data serve as a useful supplement to citizen science data portals in representing the spatial and temporal distribution of bio-diversity data?** — Social media data, such as the image sharing platform Flickr, can serve as a useful source of species observation data for certain taxonomic groups, and/or as a repository for dedicated projects. Spatial and temporal analysis suggest that the Flickr dataset best reflects the NBN dataset when considering a purely spatial distribution with no time constraints. The best represented species on Flickr in comparison to NBN are diurnal garden birds, as around 70% of the Flickr posts for them are valid observations relative to the NBN. Further, a fully automated image verification approach for identifying genuine wildlife observations on Flickr facilitates the verification of large and diverse species collections.
- **RQ 2: What are the most efficient text classification approaches for verifying that social media postings are genuine wildlife observations?** — The transformer-based contextualised word representation model, BERT, fine-tuned to the classification task by using a sequential classification layer, performed better than linear classifiers such as Logistic Regression and fastText classification pipeline for identifying genuine wildlife observations on Twitter. Specifically, the BERT model fine-tuned to classification, performed well even for very short Tweets and Tweets with more specialised language (including mentions of Latin species names). This shows the potential of transformer-based models to be used for identifying valuable wildlife-related data among informal social network sources automatically and on a larger scale, independent of the species ob-

served at hand, even when a small training corpus is provided. Further, analysis on the language used in Tweets showed trends in the usage of hashtags related to wildlife observations which are not related to official campaigns. This finding can be used to improve classification models by creating feature selection techniques which assign higher importance to such indicative features. Additionally, such hashtags could be used by informal social network campaigns to encourage people to indicate when they are posting about wildlife.

- **RQ 3: Can deep learning transformer regression models provide an effective means of geo-referencing social media posts?** — The use of a state-of-the-art transformer-based model, RoBERTa trained on the domain and then adapted for multivariate regression proved beneficial for geo-referencing Twitter data when compared to a statistical regression model, BERT model fine-tuned for regression, and the same RoBERTa model trained on a generic dataset. Additionally, enriching the Twitter training set with Flickr labelled data leads to further improvements in the performance of the geo-referencing model. This shows that using large state-of-the-art transformer-based models that have been adapted to the domain and enriching the training corpus with diverse social media datasets are suitable techniques for building geo-referencing models. Further, a hybrid approach based on location name extraction and using RoBERTa regression model for location disambiguation helps further improve the precision of geo-referencing models for social media posts.

- **RQ 4: Do zig zag persistent homology methods have good potential for extracting trajectories of spatio-temporal objects?** — A methodology for extracting, clustering and normalising object trajectories using objects' locations produced by zig-zag homology methods proved efficient and valuable for identifying similar trajectories and patterns of movement for cloud data.

7.3 Future Work

In this section, we discuss ways in which the research in this thesis can be extended in future.

Analysis of a Wider Range of Social Media Datasets The analysis presented in Chapter 3 can be extended by performing comparison between a wider range of social media platforms (such as Twitter and Instagram) and the NBN citizen science data portal. This will provide a better understanding of the the type of social media sites that are suitable to supplement wildlife-related research. While we have performed evaluation for Flickr using NBN dataset as a gold standard, it should be noted that NBN observations are collected by non-professionals. Therefore, it is possible that Flickr observations marked as false positives might be correct due to the absence of NBN observations on that location. In future, analysis can be extended to investigate this problem further. In particular, the use of multiple social media sites can help identify spatial and temporal features of species which are represented better on social networks than official citizen science portals.

Enhance Image and Text Verification Models The image verification approach proposed in Chapter 3 can be improved further by using a combination of inclusive and exclusive tags (i.e. tags used to consider a photo irrelevant) and through the development of more sophisticated computer vision methods for automated identification of individual species.

The text-based verification method, presented in Chapter 4, can also be improved by using larger and newer transformer-based models such as the one we used for geo-referencing Tweets, i.e., RoBERTa. Further, we plan on using the findings on features that are indicative for wildlife observations in order to build feature selection techniques which assign higher importance to such indicative features. This might help improve the accuracy of text classification models especially for scenarios with limited amount of training data.

Enhance Geo-referencing models for Twitter data The use of transformer-based models and transfer learning techniques proved very beneficial for building geo-referencing models. Therefore, we want to expand on these methods by using diverse social media sources, not only for enhancing the training set but also for building language models which are used for performing multivariate regression. Additionally, we want to further improve the precision of the hybrid approach based on location name extraction and regression by training the NER models used for extracting location names to the domain. Finally, considering the wide popularity of location-specific language modelling approaches, we want to expand our experiments with a hybrid approach combining both regression and language modelling methods.

Identification of Movement Patterns Throughout this thesis we focused on the problem of verifying and preparing social media datasets to facilitate its use in research related to wildlife observation and object tracking. Next steps will involve building on the trajectory extraction method by extending it for wildlife observational data and providing visualisation tools for displaying objects movement on a map. We want to focus on using spatio-temporal analysis for normalising trajectory data and identifying direction and patterns of movement for wildlife.

7.4 Summary

The work in this thesis was carried out with the aim of providing methods which can support the use of social media platforms as an unofficial data source of wildlife observations. Initially, we performed a large scale and extensive study, including the conduct of statistical, spatial, and temporal analysis on the suitability of image-sharing platforms to supplement citizen science data portals. We also presented a fully automated image verification approach. This showed that image recognition tools combined with a class-level coarse match between species names and tags returned by the tool can support the verification of large and diverse datasets without affecting the precision. In the next chapters of the thesis, we focused on incorporating the less explored state-of-the-art transformer-based models and transfer learning techniques within social media mining approaches for verifying and geo-referencing text-based social media posts related to wildlife observations. We particularly focused on evaluating the suitability of techniques for low resource settings as labelled data related to wildlife-based social media posts is usually limited. We performed extensive comparison between different classification algorithms and feature extraction and integration methods which showed that contextualised word representations, adapted for the classification task are suitable for identifying wildlife observations on social media, even when more wildlife-specific terminology is used such as species Latin names. Additionally, domain-trained contextualised word representations, fine-tuned for multivariate regression can be very beneficial for geo-referencing social media posts, especially when combined with location name extraction approaches.

This thesis shows that social media can be a rich source of wildlife observational data but it also requires the creation of verification and preparation techniques to support its usage in ecology-related studies. We showed that combining transformer-based models, transfer learning techniques, and/or rule-based approaches can facilitate the verification and geo-referencing of social media datasets even in the presence of more specialised language and a limited amount of labelled data. Further, investigation into trajectory extraction and normalisation methods based on spatio-temporal analysis opens up interesting research avenues for combining social media mining techniques and spatio-temporal analysis for building applications which use social media as a source of species observational data.

Throughout the thesis we used Flickr and Twitter as exemplary representatives of some of the most widely used social media sites which also allow data access for research. Additionally, the two social media sites differ in the nature of the data shared (i.e. images versus short texts) which provides a platform for experimenting with verification methodologies suitable for images and text-based posts. The research presented in this thesis can be applied to other social media platforms. It can also be developed further towards applications integrating social media datasets for identifying and tracking wildlife observations.

Bibliography

Frederick R Adler, Austin M Green, and Çağan H Şekercioğlu. Citizen science in Ecology: A place for humans in nature. *Annals of the New York Academy of Sciences*, 1469(1):52–64, 2020.

Ali Ahani and Mehrbakhsh Nilashi. Coronavirus outbreak and its impacts on global economy: The role of social network sites. *Journal of Soft Computing and Decision Support Systems*, 7(2):19–22, 2020.

Mohammed Ali Al-Garadi, Yuan-Chi Yang, Haitao Cai, Yucheng Ruan, Karen O'Connor, Gonzalez-Hernandez Graciela, Jeanmarie Perrone, and Abeed Sarker. Text classification models for the automatic detection of nonmedical prescription medication use from social media. *BMC medical informatics and decision making*, 21(1):1–13, 2021.

Tatsuya Amano, James DL Lamming, and William J Sutherland. Spatial gaps in global biodiversity information and the role of citizen science. *Bioscience*, 66(5): 393–400, 2016.

Mohd Yousuf Ansari, Amir Ahmad, Shehroz S Khan, Gopal Bhushan, et al. Spatiotemporal clustering: A review. *Artificial Intelligence Review*, 53(4):2381–2423, 2020.

Vyron Antoniou, Cidália Costa Fonte, Linda See, Jacinto Estima, Jamal Jokar Arsanjani, Flavio Lupia, Marco Minghini, Giles Foody, and Steffen Fritz. Investigating the feasibility of geo-tagged photographs as sources of land cover input data. *ISPRS International Journal of Geo-Information*, 5(5):64, 2016.

Lucy M Aplin, Richard E Major, Adrian Davis, and John M Martin. A citizen science approach reveals long-term social network structure in an urban parrot, Cacatua galerita. *Journal of Animal Ecology*, 90(1):222–232, 2021.

Maria Aristeidou, Christothea Herodotou, Heidi L Ballard, Alison N Young, Annie E Miller, Lila Higgins, and Rebecca F Johnson. Exploring the participation of young citizen scientists in scientific research: The case of iNaturalist. *Plos one*, 16 (1):e0245682, 2021.

Sanjeev Arora, Yingyu Liang, and Tengyu Ma. A Simple but Tough-to-Beat Baseline for Sentence Embeddings. In *5th International Conference on Learning Representations, ICLR 2017*, page 16, 2017.

Gowtham Atluri, Anuj Karpatne, and Vipin Kumar. Spatio-Temporal Data Mining: A Survey of Problems and Methods. *ACM Comput. Surv.*, 51(4), aug 2018. ISSN 0360-0300. doi: 10.1145/3161602. URL `https://doi.org/10.1145/3161602`.

Tom A August, Oliver L Pescott, Alexis Joly, and Pierre Bonnet. AI Naturalists Might Hold the Key to Unlocking Biodiversity Data in Social Media Imagery. *Patterns*, 1 (7):100116, 2020.

Mariette Awad and Rahul Khanna. *Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers*, chapter Support Vector Regression, pages 67–80. Apress, Berkeley, CA, 2015.

Lars Backstrom, Eric Sun, and Cameron Marlow. Find me if you can: Improving geographical prediction with social and spatial proximity. In *Proceedings of the 19th international conference on World wide web*, pages 61–70, 2010.

Katherine Bailey and Sunny Chopra. Few-shot text classification with pre-trained word embeddings and a human in the loop. *arXiv preprint arXiv:1804.02063*, 2018.

Vijay Barve. Discovering and developing primary biodiversity data from social networking sites: A novel approach. *Ecological Informatics*, 24:194–199, 2014.

Jonathan Bassi, Sukanya Manna, and Yu Sun. Construction of a geo-location service utilizing microblogging platforms. In *2016 IEEE Tenth International Conference on Semantic Computing (ICSC)*, pages 162–165. IEEE, 2016.

Derya Birant and Alp Kut. ST-DBSCAN: An algorithm for clustering spatial–temporal data. *Data & knowledge engineering*, 60(1):208–221, 2007.

Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.", 2009.

Alberto Blazquez-Herranz, Juan-Ignacio Caballero-Garzon, Albert Zilverberg, Christian Wolff, Alejandro Rodríguez-Gonzalez, and Ernestina Menasalvas. Clustering Moving Object Trajectories: Integration in CROSS-CPP Analytic Toolbox. *Applied Sciences*, 11(8):3693, 2021.

Andrew J Blight, A Louise Allcock, Christine A Maggs, and Mark P Johnson. Intertidal molluscan and algal species richness around the UK coast. *Marine ecology progress series*, 396:235–243, 2009.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017.

Rick Bonney, Caren B Cooper, Janis Dickinson, Steve Kelling, Tina Phillips, Kenneth V Rosenberg, and Jennifer Shirk. Citizen science: A developing tool for expanding science knowledge and scientific literacy. *BioScience*, 59(11):977–984, 2009.

Bernhard E Boser, Isabelle M Guyon, and Vladimir N Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152, 1992.

Mondher Bouazizi and Tomoaki Ohtsuki. Multi-class sentiment analysis on twitter: Classification performance and challenges. *Big Data Mining and Analytics*, 2(3): 181–194, 2019. doi: 10.26599/BDMA.2019.9020002.

Danah Boyd and Nicole Ellison. Social network sites: Definition, history, and scholarship. *IEEE Engineering Management Review*, 3(38):16–31, 2010.

Peter Bridgewater, Sandra Knapp, Christian Prip, and M MacDavette. GBIF Review 2009. From Prototype to full operation: Managing expectations. *Copenhagen*, 39, 2010.

Eleanor D Brown and Byron K Williams. The potential for citizen science to produce reliable and useful information in ecology. *Conservation Biology*, 33(3):561–569, 2019.

Peter Bubenik. Statistical topological data analysis using persistence landscapes. *The Journal of Machine Learning Research*, 16(1):77–102, 2015.

Peter Bubenik and Paweł Dłotko. A persistence landscapes toolbox for topological statistics. *Journal of Symbolic Computation*, 78:91–114, 2017.

Sergio G. Burdisso, Marcelo Errecalde, and Manuel Montes y Gomez. A text classification framework for simple and effective early depression detection over social media streams. *Expert Systems with Applications*, 133:182–197, 2019.

Pete Burnap, Rachel Gibson, Luke Sloan, Rosalynd Southern, and Matthew Williams. 140 characters to victory?: Using Twitter to predict the UK 2015 General Election. *Electoral Studies*, 41:230–233, 2016. ISSN 0261-3794. doi: https://doi.org/10.1016/j.electstud.2015.11.017. URL `https://www.sciencedirect.com/science/article/pii/S0261379415002243`.

David Camacho, Ángel Panizo-LLedot, Gema Bello-Orgaz, Antonio Gonzalez-Pardo, and Erik Cambria. The four dimensions of social network analysis: An overview of research methods, applications, and software tools. *Information Fusion*, 63: 88–120, 2020.

Gunnar Carlsson and Vin De Silva. Zigzag persistence. *Foundations of computational mathematics*, 10(4):367–405, 2010.

Carlos Castillo. *Big crisis data: Social media in disasters and time-critical situations*. Cambridge University Press, 2016.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. Universal Sentence Encoder. *CoRR*, abs/1803.11175, 2018. URL `http://arxiv.org/abs/1803.11175`.

Francine Chen, Dhiraj Joshi, Yasuhide Miura, and Tomoko Ohkuma. Social media-based profiling of business locations. In *Proceedings of the 3rd ACM Multimedia Workshop on Geotagging and Its Applications in Multimedia*, pages 1–6, 2014.

Jindong Chen, Yizhou Hu, Jingping Liu, Yanghua Xiao, and Haiyun Jiang. Deep short text classification with knowledge powered attention. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6252–6259, 2019.

Zhiyuan Cheng, James Caverlee, and Kyumin Lee. 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*, pages 759–768, 2010.

Jeffrey P Cohn. Citizen science: Can volunteers do real research? *BioScience*, 58(3): 192–197, 2008.

Çağrı Çöltekin and Taraka Rama. Tübingen-oslo at SemEval-2018 task 2: SVMs perform better than RNNs in emoji prediction. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 34–38, 2018.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. Supervised Learning of Universal Sentence Representations from Natural Language Inference Data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1070. URL `https://www.aclweb.org/anthology/D17-1070`.

Padraig Corcoran and Christopher B Jones. Modelling topological features of swarm behaviour in space and time with persistence landscapes. *IEEE Access*, 5:18534–18544, 2017.

Padraig Corcoran and Christopher B Jones. Robust tracking of objects with dynamic topology. In *Proceedings of the 26th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 428–431. ACM, 2018.

Elizabeth M Daly, Freddy Lecue, and Veli Bicer. Westland row why so slow? Fusing social media and linked data sources for understanding real-time traffic conditions. In *Proceedings of the 2013 international conference on Intelligent user interfaces*, pages 203–212, 2013.

Per-Erik Danielsson. Euclidean distance mapping. *Computer Graphics and image processing*, 14(3):227–248, 1980.

Stefan Daume. Mining Twitter to monitor invasive alien species: An analytical framework and sample information topologies. *Ecological Informatics*, 31:70–82, 2016.

Stefan Daume, Matthias Albert, and Klaus von Gadow. Forest monitoring and social media – Complementary data sources for ecosystem surveillance? *Forest Ecology and Management*, 316:9–20, 2014.

MohammadReza Davari, Leila Kosseim, and Tien Bui. TIMBERT: Toponym Identifier For The Medical Domain Based on BERT. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 662–668, 2020.

Adrian Davis, Richard E Major, Charlotte E Taylor, and John M Martin. Novel tracking and reporting methods for studying large birds in urban landscapes. *Wildlife Biology*, 2017(4), 2017.

David E Davis and Ray L Winstead. Estimating the numbers of wildlife populations. *Wildlife Society*, 1980.

Chris De Rouck, Olivier Van Laere, Steven Schockaert, and Bart Dhoedt. Georeferencing Wikipedia pages using language models from Flickr. In *10th International Semantic Web Conference (ISWC 2011)*, 2011.

Grant DeLozier, Jason Baldridge, and Loretta London. Gazetteer-independent toponym resolution using geographic word profiles. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015.

Xuelian Deng, Yuqing Li, Jian Weng, and Jilian Zhang. Feature selection for text classification: A review. *Multimedia Tools & Applications*, 78(3), 2019.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *CoRR*, abs/1810.04805, 2018. URL `http://arxiv.org/abs/1810.04805`.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL `https://aclanthology.org/N19-1423`.

Enrico Di Minin, Henrikki Tenkanen, and Tuuli Toivonen. Prospects and challenges for social media data in conservation science. *Frontiers in Environmental Science*, 3: 63, 2015.

Laura Di Rocco, Michela Bertolotto, Barbara Catania, Giovanna Guerrini, and Tiziano Cosso. Extracting fine-grained implicit georeferencing information from microblogs exploiting crowdsourced gazetteers and social interactions. In *AGILE international conference on geographic information science*, 2016.

Laura Di Rocco, Federico Dassereto, Michela Bertolotto, Davide Buscaldi, Barbara Catania, and Giovanna Guerrini. Sherloc: A knowledge-driven algorithm for geolocating microblog messages at sub-city level. *International Journal of Geographical Information Science*, 35(1):84–115, 2021.

Cathal Doyle, Rodreck David, Yevgeniya Li, Markus Luczak-Roesch, Dayle Anderson, and Cameron M Pierson. Using the web for science in the classroom: Online citizen science participation in teaching and learning. In *Proceedings of the 10th ACM Conference on Web Science*, pages 71–80, 2019.

Thomas Edwards, Christopher Jones, and Padraig Corcoran. Extracting Geometric Representations of Trajectories Using Topological Data Analysis. geocomputation 2019. *The University of Auckland*, 2019.

Thomas Edwards, Christopher B. Jones, Sarah E. Perkins, and Padraig Corcoran. Passive citizen science: The role of social media in wildlife observations. *Plos one*, 16(8):e0255416, 2021.

Thomas Edwards, Christopher B. Jones, and Padraig Corcoran. Identifying wildlife observations on Twitter. *Ecological Informatics*, 67:101500, 2022a. ISSN 1574-9541. doi: https://doi.org/10.1016/j.ecoinf.2021.101500. URL `https://www.sciencedirect.com/science/article/pii/S1574954121002910`.

Thomas Edwards, Christopher B. Jones, and Padraig Corcoran. A hybrid approach for geo-referencing tweets: Language model regression and gazetteer disambiguation. *International Journal of Geographical Information Science (IJGIS)*, under review, 2022b.

Jacob Eisenstein, Brendan O'Connor, Noah A. Smith, and Eric P. Xing. A Latent Variable Model for Geographic Lexical Variation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1277–1287, Cambridge, MA, October 2010a. Association for Computational Linguistics. URL `https://aclanthology.org/D10-1124`.

Jacob Eisenstein, Brendan O'Connor, Noah A Smith, and Eric P Xing. A latent variable model for geographic lexical variation. In *Proceedings of the 2010 conference on empirical methods in natural language processing*, pages 1277–1287. Association for Computational Linguistics, 2010b.

P Ekta, Priya Bundela, and Richa Dewan. Tweet analysis for real-time event detection and earthquake reporting system development. *International Research Journal of Engineering and Technology (IRJET)*, 4(5), 2017.

Moataz Medhat ElQadi, Alan Dorin, Adrian Dyer, Martin Burd, Zoë Bukovac, and Mani Shrestha. Mapping species distributions with social media geo-tagged images: Case studies of bees and flowering plants in australia. *Ecological informatics*, 39: 23–31, 2017.

Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, volume 96, pages 226–231, 1996.

Kawin Ethayarajh. Unsupervised random walk sentence embeddings: A strong but simple baseline. In *Proceedings of The Third Workshop on Representation Learning for NLP*, pages 91–100, 2018.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. LIBLINEAR: A library for large linear classification. *Journal of machine learning research*, 9(Aug):1871–1874, 2008.

Damien R Farine, Ariana Strandburg-Peshkin, Tanya Berger-Wolf, Brian Ziebart, Ivan Brugere, Jia Li, and Margaret C Crofoot. Both nearest neighbours and long-term affiliates predict individual locations during collective movement in wild baboons. *Scientific reports*, 6(1):1–10, 2016.

Zhenni Feng and Yanmin Zhu. A Survey on Trajectory Data Mining: Techniques and Applications. *IEEE Access*, 4:2056–2067, 2016. doi: 10.1109/ACCESS.2016.2553681.

Tommaso Fornaciari and Dirk Hovy. Identifying Linguistic Areas for Geolocation. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 231–236, Hong Kong, China, November 2019a. Association for Computational Linguistics. doi: 10.18653/v1/D19-5530. URL `https://aclanthology.org/D19-5530`.

Tommaso Fornaciari and Dirk Hovy. Geolocation with attention-based multitask learning models. In *EMNLP Workshop W-NUT: The 5th Workshop on Noisy User-generated Text*. Association for Computational Linguistics, 2019b.

Steffen Fritz, Ian McCallum, Christian Schill, Christoph Perger, Linda See, Dmitry Schepaschenko, Marijn Van der Velde, Florian Kraxner, and Michael Obersteiner. Geo-Wiki: An online platform for improving global land cover. *Environmental Modelling & Software*, 31:110–123, 2012.

Björn Gambäck and Utpal Kumar Sikdar. Using Convolutional Neural Networks to Classify Hate-Speech. In *Proceedings of the First Workshop on Abusive Language Online*, pages 85–90, Vancouver, BC, Canada, August 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-3013. URL `https://www.aclweb.org/anthology/W17-3013`.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. AllenNLP: A deep semantic natural language processing platform. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pages 1–6, Melbourne, Australia, July

2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-2501. URL `https://aclanthology.org/W18-2501`.

Eleftherios Garyfallidis, Matthew Brett, and Ian Nimmo-Smith. Fast dimensionality reduction for brain tractography clustering. In *16th Annual Meeting of the Organization for Human Brain Mapping*, 2010.

Eleftherios Garyfallidis, Matthew Brett, Marta Morgado Correia, Guy B Williams, and Ian Nimmo-Smith. Quickbundles, a method for tractography simplification. *Frontiers in neuroscience*, 6:175, 2012.

Judith Gelernter, Gautam Ganesh, Hamsini Krishnakumar, and Wei Zhang. Automatic gazetteer enrichment with user-geocoded data. In *Proceedings of the Second ACM SIGSPATIAL International Workshop on Crowdsourced and Volunteered Geographic Information*, pages 87–94, 2013.

Andrea Ghermandi and Michael Sinclair. Passive crowdsourcing of social media in environmental research: A systematic map. *Global environmental change*, 55:36–47, 2019.

Yoav Goldberg. A primer on neural network models for natural language processing. *Journal of Artificial Intelligence Research*, 57:345–420, 2016.

Milan Gritta, Mohammad Taher Pilehvar, and Nigel Collier. A pragmatic guide to geoparsing evaluation. *Language resources and evaluation*, pages 1–30, 2019.

Yiming Gu, Zhen Sean Qian, and Feng Chen. From Twitter to detector: Real-time traffic incident detection using social media data. *Transportation research part C: emerging technologies*, 67:321–342, 2016.

Christi J. Guerrini, Mary A. Majumder, Meaganne J. Lewellyn, and Amy L. McGuire. Citizen science, public policy. *Science*, 361(6398):134–136, 2018. ISSN 0036-8075. doi: 10.1126/science.aar8379. URL `https://science.sciencemag.org/content/361/6398/134`.

Pritam Gundecha and Huan Liu. Mining social media: A brief introduction. *New directions in informatics, optimization, logistics, and production*, pages 1–17, 2012.

Yuting Guo, Xiangjue Dong, Mohammed Ali Al-Garadi, Abeed Sarker, Cecile Paris, and Diego Mollá Aliod. Benchmarking of Transformer-Based Pre-Trained Models on Social Media Text Classification Datasets. In *Proceedings of the The 18th Annual Workshop of the Australasian Language Technology Association*, pages 86–91, Virtual Workshop, December 2020. Australasian Language Technology Association. URL `https://aclanthology.org/2020.alta-1.10`.

Matthias Häberle, Martin Werner, and Xiao Xiang Zhu. Geo-spatial text-mining from Twitter – A feature space analysis with a view toward building classification in urban regions. *European journal of remote sensing*, 52(sup2):2–11, 2019.

Adam Hart, Richard Stafford, Anne Goodenough, and Simon Morgan. The role of citizen science and volunteer data collection in zoological research, 2012.

Jochen Hartmann, Juliana Huppertz, Christina Schamp, and Mark Heitmann. Comparing automated text classification methods. *International Journal of Research in Marketing*, 36(1):20–38, 2019. ISSN 0167-8116. doi: https://doi.org/10.1016/j.ijresmar.2018.09.009. URL `https://www.sciencedirect.com/science/article/pii/S0167811618300545`.

J Mason Heberling, Joseph T Miller, Daniel Noesgaard, Scott B Weingart, and Dmitry Schigel. Data integration enables global biodiversity synthesis. *Proceedings of the National Academy of Sciences*, 118(6), 2021.

Matthew Honnibal and Ines Montani. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. *To appear*, 7(1):411–420, 2017.

Jeremy Howard and Sebastian Ruder. Universal Language Model Fine-tuning for Text Classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, 2018.

Wen Hua, Dat T Huynh, Saeid Hosseini, Jiaheng Lu, and Xiaofang Zhou. Information Extraction From Microblogs: A Survey. *Int. J. Softw. Informatics*, 6(4):495–522, 2012.

Xiao Huang, Zhenlong Li, Cuizhen Wang, and Huan Ning. Identifying disaster related social media for rapid response: A visual-textual fused CNN architecture. *International Journal of Digital Earth*, 0(0):1–23, 2019. doi: 10.1080/17538947.2019.1633425. URL `https://doi.org/10.1080/17538947.2019.1633425`.

Mans Hulden, Miikka Silfverberg, and Jerid Francom. Kernel density estimation for text-based geolocation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015.

Yohei Ikawa, Maja Vukovic, Jakob Rogstadius, and Akiko Murakami. Location-based insights from the social web. In *Proceedings of the 22nd international conference on World Wide Web*, pages 1013–1016, 2013.

Muhammad Imran, Ferda Ofli, Doina Caragea, and Antonio Torralba. Using ai and social media multimodal content for disaster response and management: Opportunities, challenges, and future directions, 2020.

Diana Inkpen. Text mining in social media for security threats. In *Recent Advances in Computational Intelligence in Defense and Security*, pages 491–517. Springer, 2016.

Zahedeh Izakian, M. Saadi Mesgari, and Robert Weibel. A feature extraction based trajectory segmentation approach based on multiple movement parameters. *Engineering Applications of Artificial Intelligence*, 88:103394, 2020. ISSN 0952-1976. doi: https://doi.org/10.1016/j.engappai.2019.103394. URL `https://www.sciencedirect.com/science/article/pii/S0952197619303124`.

Anuj Jaiswal, Wei Peng, and Tong Sun. Predicting time-sensitive user locations from social media. In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 870–877, 2013.

Ivan Jarić, Ricardo A Correia, Barry W Brook, Jessie C Buettel, Franck Courchamp, Enrico Di Minin, Josh A Firth, Kevin J Gaston, Paul Jepson, Gregor Kalinkat, et al. iEcology: harnessing large online resources to generate ecological insights. *Trends in Ecology & Evolution*, 35(7):630–639, 2020.

Tommi Jauhiainen, Marco Lui, Marcos Zampieri, Timothy Baldwin, and Krister Lindén. Automatic language identification in texts: A survey. *Journal of Artificial Intelligence Research*, 65:675–782, 2019.

Shelan Jeawak, Christopher Jones, and Steven Schockaert. Mapping wildlife species distribution with social media:Augmenting text classification with species names. *10th International Conference of Geographic Information Science (GIScience 2018)*, 2018.

Shelan S Jeawak, Christopher B Jones, and Steven Schockaert. Using Flickr for characterizing the environment: An exploratory analysis. In *13th International Conference on Spatial Information Theory (COSIT 2017)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2017.

Shelan S. Jeawak, Christopher B. Jones, and Steven Schockaert. Embedding Geographic Locations for Modelling the Natural Environment Using Flickr Tags and Structured Data. In Leif Azzopardi, Benno Stein, Norbert Fuhr, Philipp Mayr, Claudia Hauff, and Djoerd Hiemstra, editors, *Advances in Information Retrieval*, pages 51–66, Cham, 2019. Springer International Publishing.

Shelan S Jeawak, Christopher B Jones, and Steven Schockaert. Predicting the environment from social media: A collective classification approach. *Computers, Environment and Urban Systems*, 82:101487, 2020.

Thorsten Joachims. Text categorization with Support Vector Machines: Learning with many relevant features. In *European conference on machine learning*, pages 137–142. Springer, 1998.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of Tricks for Efficient Text Classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431. Association for Computational Linguistics, April 2017.

Morteza Karimzadeh, Scott Pezanowski, Alan M MacEachren, and Jan O Wallgrün. GeoTxt: A scalable geoparsing system for unstructured text geolocation. *Transactions in GIS*, 23(1):118–136, 2019.

Joshua D Kent and Henry T Capello Jr. Spatial patterns and demographic indicators of effective social media content during the Horsethief Canyon fire of 2012. *Cartography and Geographic Information Science*, 40(2):78–89, 2013.

Sheila Kinsella, Vanessa Murdock, and Neil O'Hare. " i'm eating a sandwich in Glasgow" modeling locations with Tweets. In *Proceedings of the 3rd international workshop on Search and mining user-generated contents*, pages 61–68, 2011.

Giorgos Kordopatis-Zilos, Symeon Papadopoulos, and Ioannis Kompatsiaris. Geotagging Text Content With Language Models and Feature Mining. *Proceedings of the IEEE*, 105(10):1971–1986, 2017. doi: 10.1109/JPROC.2017.2688799.

Sayali Kulkarni, Shailee Jain, Mohammad Javad Hosseini, Jason Baldridge, Eugene Ie, and Li Zhang. Spatial Language Representation with Multi-level Geocoding. *ArXiv*, abs/2008.09236, 2020.

Abhinav Kumar and Jyoti Prakash Singh. Location reference identification from Tweets during emergencies: A deep learning approach. *International journal of disaster risk reduction*, 33:365–375, 2019.

Olivier Van Laere, Steven Schockaert, Vlad Tanasescu, Bart Dhoedt, and Christopher B. Jones. Georeferencing Wikipedia Documents Using Data from Social Media Sources. *ACM Trans. Inf. Syst.*, 32(3), July 2014a. ISSN 1046-8188. doi: 10.1145/2629685. URL https://doi.org/10.1145/2629685.

Olivier Van Laere, Steven Schockaert, Vlad Tanasescu, Bart Dhoedt, and Christopher B Jones. Georeferencing Wikipedia documents using data from social media sources. *ACM Transactions on Information Systems (TOIS)*, 32(3):12, 2014b.

Juan Carlos Laso Bayas, Linda See, Myroslava Lesiv, Martina Dürauer, Ivelina Georgieva, Dmitry Schepaschenko, Mathias Karner, Olga Danylo, Hedwig Bartl, Anto Subash, et al. Experiences from Recent Geo-Wiki Citizen Science Campaigns in the Creation and Sharing of New Reference Data Sets on Land Cover and Land Use. In *EGU General Assembly Conference Abstracts*, pages EGU21–10871, 2021.

Jessica A Leivesley, Robyn A Stewart, Victoria Paterson, and Dominic J McCafferty. Potential importance of urban areas for water voles: Arvicola amphibius. *European Journal of Wildlife Research*, 67(1):1–4, 2021.

Daniel Leung and Shawn Newsam. Exploring geotagged images for land-use classification. In *Proceedings of the ACM multimedia 2012 workshop on Geotagging and its applications in multimedia*, pages 3–8, 2012.

Vladimir I Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710, 1966.

Hongmin Li, D Caragea, X Li, and Cornelia Caragea. Comparison of Word Embeddings and Sentence Encodings as Generalized Representations for Crisis Tweet Classification Tasks. *Proceedings of the ISCRAM Asian Pacific 2018 Conference, New Zealand*, page 13, 2018.

Jing Li, Xueming Qian, Ke Lan, Peilun Qi, and Arunabh Sharma. Improved image GPS location estimation by mining salient features. *Signal Processing: Image Communication*, 38:141–150, 2015.

Huan Liu, Fred Morstatter, Jiliang Tang, and Reza Zafarani. The good, the bad, and the ugly: Uncovering novel research opportunities in social media mining. *International Journal of Data Science and Analytics*, 1(3):137–143, 2016.

Junhua Liu, Trisha Singhal, Lucienne T.M. Blessing, Kristin L. Wood, and Kwan Hui Lim. CrisisBERT: A Robust Transformer for Crisis Classification and Contextual Crisis Embedding. In *Proceedings of the 32nd ACM Conference on Hypertext and Social Media*, pages 133–141, New York, NY, USA, 2021. Association for Computing Machinery.

Kan Liu and Lu Chen. Medical Social Media Text Classification Integrating Consumer Health Terminology. *IEEE Access*, 7:78185–78193, 2019. doi: 10.1109/ACCESS.2019.2921938.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019a. URL `http://arxiv.org/abs/1907.11692`.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019b.

Armando López-Cuevas, José Ramírez-Márquez, Gildardo Sanchez-Ante, and Kash Barker. A community perspective on resilience analytics: A visual analysis of community mood. *Risk Analysis*, 37(8):1566–1579, 2017.

Ezequiel Lopez-Lopez, Jorge Carrillo-de Albornoz, and Laura Plaza. Combining Transformer-Based Models with Traditional Machine Learning Approaches for Sexism Identification in Social Networks at EXIST 2021. *PLOS One*, 2021.

Christopher S Lowry and Michael N Fienen. CrowdHydrology: Crowdsourcing hydrologic data and engaging citizen scientists. *GroundWater*, 51(1):151–156, 2013.

Sergio Luna and Michael J. Pennock. Social media applications and emergency management: A literature review and research agenda. *International Journal of Disaster Risk Reduction*, 28:565–577, 2018. ISSN 2212-4209. doi: https://doi.org/10.1016/

j.ijdrr.2018.01.006. URL `https://www.sciencedirect.com/science/article/pii/S221242091830030X`.

Arjun Magge, Ari Klein, Antonio Miranda-Escalada, Mohammed Ali Al-Garadi, Ilseyar Alimova, Zulfat Miftahutdinov, Eulalia Farre, Salvador Lima-López, Ivan Flores, Karen O'Connor, et al. Overview of the sixth social media mining for health applications (# smm4h) shared tasks at NAACL 2021. In *Proceedings of the Sixth Social Media Mining for Health (# SMM4H) Workshop and Shared Task*, pages 21–32, 2021.

Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. The Stanford coreNLP natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pages 55–60, 2014.

Saravanapriya Manoharan and Radha Senthilkumar. An intelligent fuzzy rule-based personalized news recommendation using social media mining. *Computational intelligence and neuroscience*, 2020, 2020.

Yago Martin, Susan L Cutter, Zhenlong Li, Christopher T Emrich, and Jerry T Mitchell. Using geotagged Tweets to track population movements to and from Puerto Rico after Hurricane Maria. *Population and Environment*, 42(1):4–27, 2020.

Matej Martinc and Senja Pollak. Combining n-grams and deep convolutional features for language variety classification. *Natural Language Engineering*, 25(5):607–632, 2019.

Jean Damascène Mazimpaka and Sabine Timpf. Trajectory data mining: A review of methods and applications. *Journal of Spatial Information Science*, 2016(13):61–99, 2016.

Andrew McCallum, Kamal Nigam, et al. A comparison of event models for naive bayes text classification. In *AAAI-98 workshop on learning for text categorization*, volume 752, pages 41–48. Citeseer, 1998.

Bryan McCann, James Bradbury, Caiming Xiong, and Richard Socher. Learned in Translation: Contextualized Word Vectors. *Advances in Neural Information Processing Systems*, 30, 2017.

Jonathan Mellon. Internet search data and issue salience: The properties of Google Trends as a measure of issue salience. *Journal of Elections, Public Opinion & Parties*, 24(1):45–72, 2014.

Fernando Melo and Bruno Martins. Geocoding textual documents through the usage of hierarchical classifiers. In *Proceedings of the 9th Workshop on Geographic Information Retrieval*, pages 1–9, 2015.

Stephen Merity, Nitish Shirish Keskar, and Richard Socher. Regularizing and Optimizing LSTM Language Models. In *International Conference on Learning Representations*, pages 1–13, 2018.

Danny Merkx and Stefan L Frank. Comparing Transformers and RNNs on predicting human sentence processing data. *arXiv e-prints*, pages arXiv–2005, 2020.

Stuart E Middleton, Giorgos Kordopatis-Zilos, Symeon Papadopoulos, and Yiannis Kompatsiaris. Location extraction from social media: Geoparsing, location disambiguation, and geotagging. *ACM Transactions on Information Systems (TOIS)*, 36(4): 1–27, 2018.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013a.

Tomáš Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies*, pages 746–751, 2013b.

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. SemEval-2018 task 1: Affect in Tweets. In *Proceedings of the 12th international workshop on semantic evaluation*, pages 1–17, 2018.

M Mohsin. Tiktok statistics that you need to know in 2020 [infographic]. *Oberlo.[online] Available at: https://www. oberlo. com/blog/tiktok-statistics [Accessed: 24 March 2020]*, 10.

Graham G Monkman, Michel J Kaiser, and Kieran Hyder. Text and data mining of social media to map wildlife recreation activity. *Biological conservation*, 228:89–99, 2018.

Aibek Musaev, De Wang, Saajan Shridhar, Chien-An Lai, and Calton Pu. Toward a real-time service for landslide detection: Augmented explicit semantic analysis and clustering composition approaches. In *2015 IEEE International Conference on Web Services*, pages 511–518. IEEE, 2015.

Justin R Ortiz, Hong Zhou, David K Shay, Kathleen M Neuzil, Ashley L Fowlkes, and Christopher H Goss. Monitoring influenza activity in the United States: A comparison of traditional surveillance systems with Google Flu Trends. *PloS one*, 6(4):e18687, 2011.

Feyza Altunbey Ozbay and Bilal Alatas. Fake news detection within online social media using supervised artificial intelligence algorithms. *Physica A: Statistical Mechanics and its Applications*, 540:123174, 2020. doi: https://doi.org/10.1016/j.physa.2019.123174.

Marco Palomino, Tim Taylor, Ayse Göker, John Isaacs, and Sara Warber. The online dissemination of nature–health concepts: Lessons from sentiment analysis of social media relating to 'nature-deficit disorder'. *International Journal of Environmental Research and Public Health*, 13(1):142, 2016.

Jonathan D Paul, Wouter Buytaert, Simon Allen, Juan A Ballesteros-Cánovas, Jagat Bhusal, Katarzyna Cieslik, Julian Clark, Sumit Dugar, David M Hannah, Markus

Stoffel, et al. Citizen science for hydrological risk reduction and resilience building. *Wiley Interdisciplinary Reviews: Water*, 5(1):e1262, 2018.

Jorge David Gonzalez Paule, Yeran Sun, and Yashar Moshfeghi. On fine-grained geo-localisation of Tweets and real-time traffic incident detection. *Information Processing & Management*, 56(3):1119–1132, 2019.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830, 2011.

Thorsten Brants Ashok C. Popat Peng and Xu Franz J. Och Jeffrey Dean. Large Language Models in Machine Translation. *EMNLP-CoNLL 2007*, page 858, 2007.

Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global Vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proceedings of NAACL-HLT*, pages 2227–2237, 2018.

Matthew Pittman and Brandon Reich. Social media and loneliness: Why an Instagram picture may be worth more than a thousand Twitter words. *Computers in Human Behavior*, 62:155–167, 2016. ISSN 0747-5632. doi: https://doi.org/10.1016/j.chb.2016.03.084. URL `https://www.sciencedirect.com/science/article/pii/S0747563216302552`.

Soujanya Poria, Erik Cambria, Devamanyu Hazarika, and Prateek Vij. A deeper look into sarcastic Tweets using deep convolutional neural networks. *arXiv preprint arXiv:1610.08815*, 2016.

Anjuman Prabhat and Vikas Khullar. Sentiment classification on big data using NaÃ¯ve Bayes and logistic regression. In *2017 International Conference on Computer Communication and Informatics (ICCCI)*, pages 1–5, 2017. doi: 10.1109/ICCCI.2017.8117734.

Tobias Preis, Helen Susannah Moat, and H Eugene Stanley. Quantifying trading behavior in financial markets using Google Trends. *Scientific reports*, 3(1):1–6, 2013.

Reid Priedhorsky, Aron Culotta, and Sara Y Del Valle. Inferring the origin locations of Tweets with quantitative confidence. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, pages 1523–1536, 2014.

Hemant Purohit and Steve Peterson. *Social Media Mining for Disaster Management and Community Resilience*, pages 93–107. Springer International Publishing, Cham, 2020. ISBN 978-3-030-48099-8. doi: 10.1007/978-3-030-48099-8_5. URL `https://doi.org/10.1007/978-3-030-48099-8_5`.

Afshin Rahimi, Duy Vu, Trevor Cohn, and Timothy Baldwin. Exploiting Text and Network Context for Geolocation of Social Media Users. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1362–1367, 2015.

Afshin Rahimi, Timothy Baldwin, and Trevor Cohn. Continuous representation of location for geolocation and lexical dialectology using mixture density networks. *arXiv preprint arXiv:1708.04358*, 2017a.

Afshin Rahimi, Trevor Cohn, and Timothy Baldwin. A neural model for user geolocation and lexical dialectology. *arXiv preprint arXiv:1704.04008*, 2017b.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, 2019.

Darcy Reynard and Manish Shirgaokar. Harnessing the power of machine learning: Can Twitter data be useful in guiding resource allocation decisions during a natural disaster? *Transportation Research Part D: Transport and Environment*, 77:449–463, 2019.

Stephen Roller, Michael Speriosu, Sarat Rallapalli, Benjamin Wing, and Jason Baldridge. Supervised text-based geolocation using language models on an adaptive grid. In *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning*, pages 1500–1510, 2012.

Natalia C Roos and Guilherme O Longo. Critical information for fisheries monitoring may be available in social media. *Aquatic Conservation: Marine and Freshwater Ecosystems*, 31(9):2420–2428, 2021.

Abeed Sarker, Maksim Belousov, Jasper Friedrichs, Kai Hakala, Svetlana Kiritchenko, Farrokh Mehryary, Sifei Han, Tung Tran, Anthony Rios, Ramakanth Kavuluru, Berry de Bruijn, Filip Ginter, Debanjan Mahata, Saif M Mohammad, Goran Nenadic, and Graciela Gonzalez-Hernandez. Data and systems for medication-related text classification and concept normalization from Twitter: Insights from the Social Media Mining for Health (SMM4H)-2017 shared task. *Journal of the American Medical Informatics Association*, 25(10):1274–1283, 10 2018. URL `https://doi.org/10.1093/jamia/ocy114`.

Yves Scherrer, Nikola Ljubešić, et al. 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), 2020.

Yves Scherrer, Nikola Ljubešić, et al. 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, 2021.

Stefan Schweter and Alan Akbik. FLERT: Document-Level Features for Named Entity Recognition, 2020.

Pavel Serdyukov, Vanessa Murdock, and Roelof Van Zwol. Placing Flickr photos on a map. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 484–491. ACM, 2009.

Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter*, 19(1):22–36, 2017.

Jonathan Silvertown. A new dawn for citizen science. *Trends in ecology & evolution*, 24(9):467–471, 2009.

Asamaporn Sitthi, Masahiko Nagai, Matthew Dailey, and Sarawut Ninsawat. Exploring land use and land cover of geotagged social-sensing images using naive Bayes classifier. *Sustainability*, 8(9):921, 2016.

Marta Skreta, Alexandra Luccioni, and David Rolnick. Spatiotemporal Features Improve Fine-Grained Butterfly Image Classification. In *Conference on Neural Information Processing Systems*, 2020.

Aiman Soliman, Kiumars Soltani, Junjun Yin, Anand Padmanabhan, and Shaowen Wang. Social sensing of urban land use based on analysis of Twitter users' mobility patterns. *PlOS ONE*, 12(7):e0181657, 2017.

Kristin Stock. Mining location from social media: A systematic review. *Computers, Environment and Urban Systems*, 71:209 – 240, 2018a.

Kristin Stock. Mining location from social media: A systematic review. *Computers, Environment and Urban Systems*, 71:209–240, 2018b.

Oliver C Stringham, Stephanie Moncayo, Katherine GW Hill, Adam Toomes, Lewis Mitchell, Joshua V Ross, and Phillip Cassey. Text classification to streamline online wildlife trade analyses. *PloS one*, 16(7):e0254007, 2021.

Han Su, Shuncheng Liu, Bolong Zheng, Xiaofang Zhou, and Kai Zheng. A survey of trajectory distance measures and performance evaluation. *The VLDB Journal*, 29(1): 3–32, 2020.

LN Swamy and JV Gorabal. Logistic regression-based classification for reviews analysis on E-commerce based applications. In *Frontiers in intelligent computing: Theory and applications*, pages 323–334. Springer, 2020.

Chuanqi Tan, Fuchun Sun, Tao Kong, Wenchang Zhang, Chao Yang, and Chunfang Liu. A survey on deep transfer learning. In *International conference on artificial neural networks*, pages 270–279. Springer, 2018.

Erik F. Tjong Kim Sang and Fien De Meulder. 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*, pages 142–147, 2003. URL `https://www.aclweb.org/anthology/W03-0419`.

Santosh Tokala, Vaibhav Gambhir, and Animesh Mukherjee. Deep Learning for social media health text classification. In *Proceedings of the 2018 EMNLP workshop SMM4H: the 3rd social media mining for health applications workshop & shared task*, pages 61–64, 2018.

Sergio Torrijos, Alejandro Bellogín, and Pablo Sánchez. Discovering related users in location-based social networks. In *Proceedings of the 28th ACM conference on user modeling, adaptation and personalization*, pages 353–357, 2020.

David Tuxworth, Dimosthenis Antypas, Luis Espinosa-Anke, Jose Camacho-Collados, Alun Preece, and David Rogers. Deriving Disinformation Insights from Geolocalized Twitter Callouts, 2021.

Belén Usero, Virginia Hernández, and Cynthia Quintana. Social Media Mining for Business Intelligence Analytics: An Application for Movie Box Office Forecasting. In *Intelligent Computing*, pages 981–999. Springer, 2022.

Steven Van Canneyt, Steven Schockaert, and Bart Dhoedt. Discovering and characterizing places of interest using Flickr and Twitter. *International Journal on Semantic Web and Information Systems (IJSWIS)*, 9(3):77–104, 2013.

Olivier Van Laere, Steven Schockaert, and Bart Dhoedt. Georeferencing Flickr resources based on textual meta-data. *Information Sciences*, 238:52–74, 2013. ISSN 0020-0255. doi: https://doi.org/10.1016/j.ins.2013.02.045. URL `https://www.sciencedirect.com/science/article/pii/S002002551300162X`.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

Eelke Visser, Emil HJ Nijhuis, Jan K Buitelaar, and Marcel P Zwiers. Partition-based mass clustering of tractography streamlines. *NeuroImage*, 54(1):303–312, 2011.

Ulrike Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17 (4):395–416, 2007.

Henrik von Scheel, Zakaria Maamar, and Mona von Rosing. Social Media and Business Process Management. In Mark von Rosing, August-Wilhelm Scheer, and Henrik von Scheel, editors, *The Complete Business Process Handbook*, pages 381–398. Morgan Kaufmann, Boston, 2015. ISBN 978-0-12-799959-3. doi: https://doi.org/10.1016/B978-0-12-799959-3.00018-5.

Strother H Walker and David B Duncan. Estimation of the probability of an event as a function of several independent variables. *Biometrika*, 54(1-2):167–179, 1967.

Sheng Wang, Zhifeng Bao, J Shane Culpepper, and Gao Cong. A survey on trajectory data management, analytics, and learning. *ACM Computing Surveys (CSUR)*, 54(2): 1–36, 2021.

Benjamin Wing and Jason Baldridge. Simple supervised document geolocation with geodesic grids. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pages 955–964, 2011.

Benjamin Wing and Jason Baldridge. Hierarchical discriminative classification for text-based geolocation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 336–348, 2014.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *ArXiv*, abs/1910.03771, 2019.

Mike Worboys and Matt Duckham. Monitoring qualitative spatiotemporal change for geosensor networks. *International Journal of Geographical Information Science*, 20(10):1087–1108, 2006. doi: 10.1080/13658810600852180. URL `https://doi.org/10.1080/13658810600852180`.

Liang Wu and Huan Liu. Tracing Fake-News Footprints: Characterizing Social Media Messages by How They Propagate. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, WSDM '18, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450355810. doi: 10.1145/3159652.3159677. URL `https://doi.org/10.1145/3159652.3159677`.

Yijun Xiao and Kyunghyun Cho. Efficient character-level document classification by combining convolution and recurrent layers. *arXiv preprint arXiv:1602.00367*, 2016.

Zheren Yan, Can Yang, Lei Hu, Jing Zhao, Liangcun Jiang, and Jianya Gong. The integration of linguistic and geospatial features using global context embedding for automated text geocoding. *ISPRS International Journal of Geo-Information*, 10(9): 572, 2021.

Dingqi Yang, Terence Heaney, Alberto Tonon, Leye Wang, and Philippe Cudré-Mauroux. CrimeTelescope: Crime hotspot prediction based on urban and social media data fusion. *World Wide Web*, 21(5):1323–1347, 2018a.

Zhilin Yang, Zihang Dai, Ruslan Salakhutdinov, and William W Cohen. Breaking the Softmax Bottleneck: A High-Rank RNN Language Model. In *International Conference on Learning Representations*, pages 1–18, 2018b.

Chris Yesson, Peter W Brewer, Tim Sutton, Neil Caithness, Jaspreet S Pahwa, Mikhaila Burgess, W Alec Gray, Richard J White, Andrew C Jones, Frank A Bisby, et al. How global is the global biodiversity information facility? *PloS one*, 2(11): e1124, 2007.

Manzhu Yu, Qunying Huang, Han Qin, Chris Scheele, and Chaowei Yang. Deep learning for real-time social media text classification for situation awareness - using Hurricanes Sandy, Harvey, and Irma as case studies. *International Journal of Digital Earth*, 12(11):1230–1247, 2019.

Guan Yuan, Penghui Sun, Jie Zhao, Daxing Li, and Canwei Wang. A review of moving object trajectory clustering algorithms. *Artificial Intelligence Review*, 47(1): 123–144, 2017.

Reza Zafarani, Mohammad Ali Abbasi, and Huan Liu. *Social media mining: An introduction*. Cambridge University Press, 2014.

Wei Zhang and Judith Gelernter. Geocoding location expressions in twitter messages: A preference learning method. *Journal of Spatial Information Science*, 2014(9):37–70, 2014.

Yu Zhang, Wenzhou Wu, Qi Wang, and Fenzhen Su. A geo-event-based geospatial information service: A case study of typhoon hazard. *Sustainability*, 9(4):534, 2017.

Xin Zheng, Jialong Han, and Aixin Sun. A survey of location prediction on Twitter. *IEEE Transactions on Knowledge and Data Engineering*, 30(9):1652–1671, 2018.

Yu Zheng. Location-based social networks: Users. In *Computing with spatial trajectories*, pages 243–276. Springer, 2011.

Xiao Zhong and David Enke. Predicting the daily return direction of the stock market using hybrid machine learning algorithms. *Financial Innovation*, 5(1):1–20, 2019.