]

Highlights

Identifying Wildlife Observations on Twitter

Thomas Edwards,Christopher B. Jones,Padraig Corcoran

- Machine learning methods enable identification of genuine wildlife data on Twitter
- Fine-tuning the neural model BERT is highly effective for identifying wildlife Tweets
- A bag of words classifier is a strong baseline for identifying wildlife Tweets
- Wildlife observations on Twitter are associated with distinctive hashtags

Identifying Wildlife Observations on Twitter[⋆,⋆⋆]

Thomas Edwards[a,c,∗,1], Christopher B. Jones[a] and Padraig Corcoran[a]

[a] *School Of Computer Science and Informatics, Cardiff University, Queens Building, 5 The Parade, Roath, Cardiff CF24 3AA, Cardiff, UK*

ARTICLE INFO

Keywords:
wildlife observations
supervised machine learning
natural language processing
social media
crowdsourcing
citizen science

ABSTRACT

Despite the potential of social media for environmental monitoring, concerns remain about the quality and reliability of the information automatically extracted. Notably there are many observations of wildlife on Twitter, but their automated detection is a challenge due to the frequent use of wildlife related words in messages that have no connection with wildlife observation. We investigate whether and what type of supervised machine learning methods can be used to create a fully automated text classification model to identify genuine wildlife observations on Twitter, irrespective of species type or whether Tweets are geo-tagged. We perform experiments with various techniques for building feature vectors that serve as input to the classifiers, and consider how they affect classification performance. We compare three classification approaches and perform an analysis of the types of features that are indicative for genuine wildlife observations on Twitter. In particular, we compare some classical machine learning algorithms, widely used in ecology studies, with state-of-the-art neural network models. Results showed that the neural network-based model Bidirectional Encoder Representations from Transformers (BERT) outperformed the classical methods. Notably this was the case for a relatively small training corpus, consisting of less than 3000 instances. This reflects that fact that the BERT classifier uses a transfer learning approach that benefits from prior learning on a very much larger collection of generic text. BERT performed particularly well even for Tweets that employed specialised language relating to wildlife observations. The analysis of possible indicative features for wildlife Tweets revealed interesting trends in the usage of hashtags that are unrelated to official citizen science campaigns. The findings from this study facilitate more accurate identification of wildlife-related data on social media which can in turn be used for enriching citizen science data collections.

1. Introduction

Studies of wildlife species distribution patterns are increasingly important in the face of rapid ecosystem changes which in turn have implications for disease emergence and spread, as well as food security (Amano et al., 2016; Barve, 2014), climate change, and invasive species biology. Traditional approaches for observing species occurrence involve the participation of professionals. Despite the high reliability of data gathered by professionals, these approaches are time and resource consuming and thus lack broad coverage (Amano et al., 2016). Citizen science campaigns of organised groups of volunteers in partnership with professionals have proven to be very effective for observing wildlife behaviour and have considerable value in ecological conservation studies (Cohn, 2008). The National Biodiversity Network (NBN) Atlas (`https://nbn.org.uk`) portal, which holds the most extensive collection of biodiversity information within the UK, has been populated with datasets collected by citizen scientists. NBN datasets have proved beneficial in studying wildife distribution in a number of studies (Leivesley et al., 2021; Blight et al., 2009). However, the use of social networks for observing the environment in contexts that are unrelated to any particular citizen science programs presents a potentially valuable additional source of species observation datasets.

Social media provide a large number of users with the capability to share images and associated metadata with other users. The significant amount of environmental data on social networks, such as Twitter and Flickr, and their potential for monitoring ecological changes and species movement patterns, has become recognised in recent years (Daume, 2016; Di Minin et al., 2015; Ghermandi and Sinclair, 2019; Edwards et al., 2021). Notably these data sources provide the possibility of timely and (near) real-time monitoring and analysis of species distribution (Ghermandi and Sinclair, 2019; Daume et al., 2014; ElQadi et al., 2017). Further, the large volume of available data comes with lower time and labour resource overheads relative to citizen science campaigns (Ghermandi and Sinclair, 2019; Antoniou et al., 2016; Soliman et al., 2017). Equally, the value of internet sources for gathering wildlife-related data has

[∗] Corresponding author
✉ `t.j.edwards@cs.cf.ac.uk` (T. Edwards)
ORCID(s):

emerged in the context of several planned citizen science initiatives. For instance, the citizen science platform iNaturalist (https://www.inaturalist.org/) has been successfully organising campaigns for observing wildlife. iNaturalist is a web-based and mobile-supported social network which allows individuals to upload photo observations and identify organisms (Aristeidou et al., 2021). 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 (Lowry and Fienen, 2013) crowdsourcing was used as a supplemental method for collecting hydrologic data. An overview of the impact of internet social networks on traditional biodiversity data collection methods in (Di Minin et al., 2015) is optimistic and concludes that social media can potentially play an important role in conservation science. The authors of (Daume, 2016; ElQadi et al., 2017; Barve, 2014) evaluate social media websites such as Flickr and Twitter in comparison to traditional wildlife data portals in order to highlight the potential use of social media for augmenting traditional citizen science data collection methods.

Despite the potential 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). 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 indicating a specific observation of the presence of a 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. Previous research on verifying social media data for wildlife studies is limited to the use of manual or semi-automatic approaches and are limited to observing a small amount of well-known species (Daume, 2016; ElQadi et al., 2017; Barve, 2014). Notably some automated techniques for validating genuine wildlife observations on social networks are based on image verification rather than text verification techniques (ElQadi et al., 2017; Estima et al., 2014; Antoniou et al., 2016; Di Minin et al., 2018). While image verification techniques are undoubtedly very useful, there are many 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 Tweets and Tweets that are actual wildlife observations. Although existing studies highlight the potential value of social network data for supplementing traditional biodiversity data collections, little progress has been made to date on developing reliable automated methods for exploiting the textual content of social media postings for tasks such as the study of species distribution.

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 [1]. 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 transformer 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 of wildlife occurrences. This can help the creation of targeted campaigns that influence social media trends in order to produce higher quality data. Our main contributions are:

[1] The implementation for the classification methods and the dataset are freely available at: `https://github.com/te9055/Social-Media-Wildlife-Distribution`

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 paper is structured as follows: In Section 1.1 we introduce our main text classification concepts and methods. Section 1.2 presents related work on applications of machine learning to classifying social media that relate to the environment, highlighting differences from our work. Our methods and materials are described in detail in Section 2. In Section 3 we present the results from our classification models and relevant analyses. Section 4 discusses findings and implications of the study and Section 5 concludes the paper.

1.1. Supervised Text Classification

Text 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, the large number of misspellings and the jargon language used, as well as the presence of polysemous words (Bouazizi and Ohtsuki, 2019). In the following, we describe our classical machine learning approaches as well as the state-of-the-art neural network models used for text classification (see Section 1.1.1) which have become commonly adopted for categorising social media data. Further, we explain methods for representing the features often used as input for classification models (see Section 1.1.2).

1.1.1. Machine Learning Approaches

Classical Machine Learning Models: Machine learning algorithms such as SVM and Logistic Regression, coupled with feature vectors that represent the frequency of occurrence of individual words, have traditionally been used for performing text classification. Despite their simplicity, they can provide a strong baseline 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). A drawback of such approaches is that they are limited in their capacity to deal with out-of-vocabulary (OOV) 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). The fastText pipeline classifier addresses this problem with an approach based on word embeddings (see also section 1.1.2) 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. We implement this form of classifier and compare it with more advanced approaches that use deep learning.

Neural Network Machine Learning Models: 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 networks (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 backpropagation 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., 2018). These limitations 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). In our work we apply the BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2018) 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.

1.1.2. Feature Representation Methods

We distinguish between three main types of feature representation 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.

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 re-trained, with a fine tuning process, 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). The pre-training can be expected to have exposed the model to a much wider vocabulary and range of language uses than in the application-specific training dataset. However, the representations for the words (based on word embeddings) from pre-training might be similar to those of related words that do appear in the application training set, which allows the model to generalise better when it is applied to unseen test 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.

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 statistical machine learning approaches (Peng and Dean, 2007; Mikolov et al., 2013b), described in Section 1.1.1 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). The vector with an element for words in the entire vocabulary is referred to as a bag of words (BOW). 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 Embeddings: 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 models (Mikolov et al., 2013a). The Skip-gram model 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. In our methods described in Section 2, we perform experiments with three popular word embedding models.

Language Models: 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, 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., 2018). The BERT (Peters et al., 2018; Devlin et al., 2018) transformer model that we use here has been pre-trained on large amounts of generic data (Books Corpus and English Wikipedia). This pre-trained model can be fine-tuned to a specific task (being classification of Tweets in our case) by adding a single additional output layer (a classification layer) to the neural network architecture. In our work we have experimented with using the pre-trained model directly and with fine-tuning the model, as explained further in Section 2.

1.2. Related Work: Text classification for wildlife data

Here we review related work that employs machine learning methods for detecting wildlife and related environmental data from social media as well as related work that uses social media for detecting events in the context of emergency response.

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 text associated with Flickr photographs and the visual features of 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 (Jeawak et al., 2017) use SVM classifiers that takes 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. Jeawak et al. (2020) propose a collective classification model to predict similar environmental phenomena, again combining Flickr tags with environmental data, to define a neighbourhood structure. An iterative approach predicts what is present at an individual location based on neighbouring data that includes elements of the training data (analogously to interpolation methods). An associated study (Jeawak et al., 2018) using only Flickr data focused on bird species distribution and 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 an SVM classifier, 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.

The work proposed by Xu et al. (2019) used Twitter to detect and classify suspicious wildlife trafficking and sale using an unsupervised machine learning topic model combined with keyword filtering and manual annotation. The study was limited to studying two wildlife animals and related products: elephant ivory and pangolin. The authors used the clustering method bi-term topic model (BTM) to categorize similar text into related topic clusters. BTM is an unsupervised machine learning topic model that uses natural language processing (NLP) to categorize short forms of text in a given number of groups (topics) by analyzing the correlations between words and topics. Our work differs from this study significantly in our focus on identifying Tweets that are observations of wildlife as opposed to ones concerned with illegal sales, and that we adopt a fully automated supervised machine learning approach.

Monkman et al. (2018) present 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. Natural language processing software was used in a ruled based system to classify sentences based on their inclusion of terms from a manually constructed lexicon.

Additionally, recent work on text classification for social media has focused on disaster management and hate speech recognition (Reynard and Shirgaokar, 2019; Huang et al., 2019; Gambäck and Sikdar, 2017; Li et al., 2018; Poria et al., 2016) mainly using CNN or SVM classifiers. A more generalised classification model for filtering crisis Tweets was proposed in Li et al. (2018), based on the use of pre-trained and specialised corpus-trained word embeddings for representing the Tweets' vocabulary. Two approaches were presented for building Tweets embedding vectors, the first being 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 was evaluated with Naive Bayes, Random Forest, K-nearest Neighbours and SVM classifiers.

An extensive analysis was conducted on how different word embedding models affect classification performance. In our work, which is for a different task, we also evaluate the use of sentence embedding methods but we differ in experimenting with and demonstrating the benefits of transformer based neural network methods.

The approaches described above perform classification with either classical machine learning methods or using traditional neural models such as CNN that assume the availability of large amounts of training data. Further, there is limited comparison between different classification models and how their performance is affected by the use of different feature representation methods. We present an extensive comparison between three different classification approaches and various feature representation methods, and their suitability for small and task-specific social network collections. Further, we propose a fully automatic approach for predicting wildlife observations, regardless of the species that need to be studied.

2. Materials and Methods

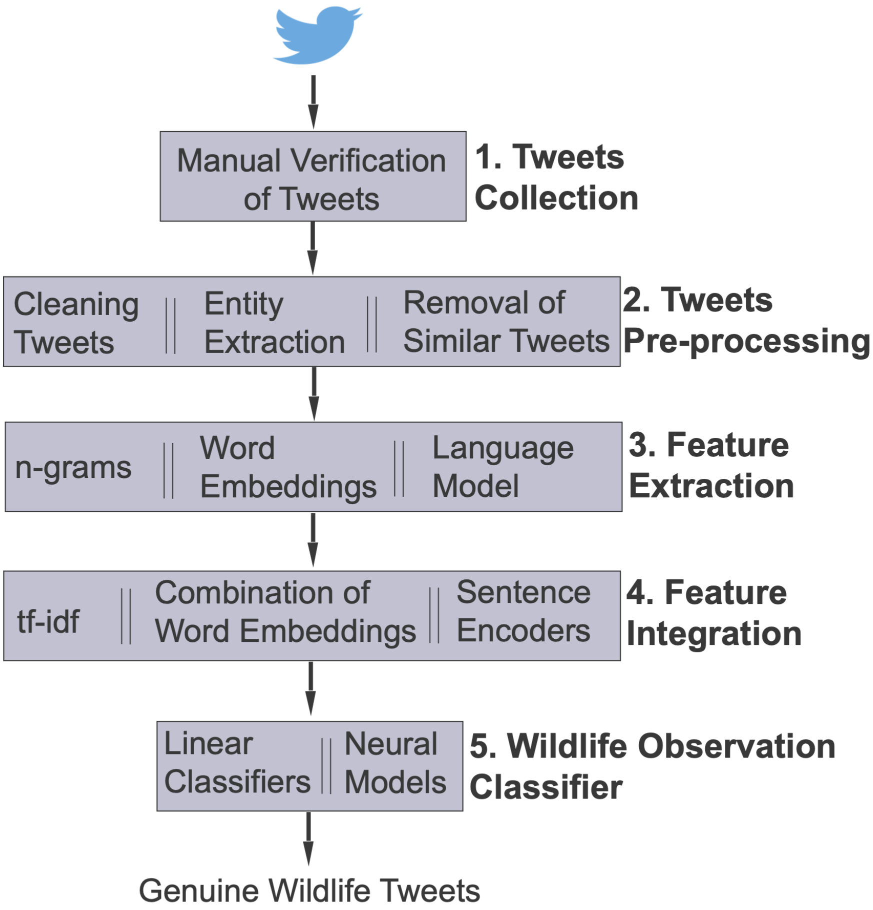

Figure 1: Overview of the methodology followed to build a classifier including main steps ('Tweets collection', 'Tweets pre-processing', 'Feature Extraction', 'Feature Integration', 'Wildlife Observation Classifier') as well as the different methods we experimented with during each of these steps.

Our study aims to develop a text classifier model, for identifying wildlife observations on social media sites. We have built classification models using the Python programming language and standard Python libraries suitable for classification and use of neural network models, such as the Sklearn and Hugging Face transformers libraries.

Our methodology consists of five main steps, *Tweets collection*, *Pre-processing*, *Feature Extraction*, *Feature Integration*, and finally training a *Wildlife Observation Classifier*. See Figure 1 for overall flow of the methodology and Figure 2 for an example of a Tweet being processed using the classification methodology. During the collection and pre-

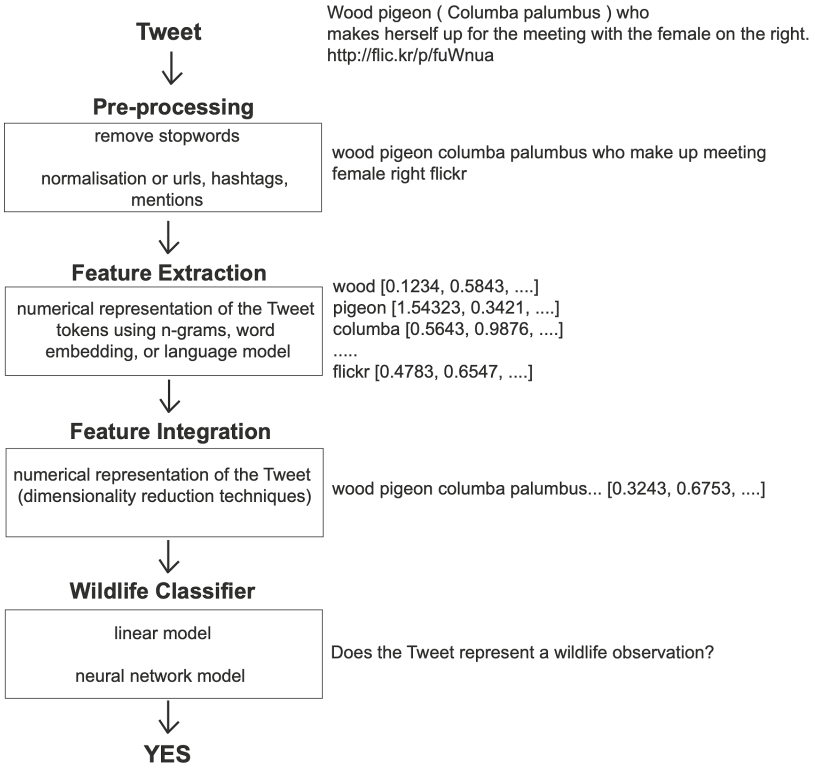
Tweet
Wood pigeon (Columba palumbus) who
makes herself up for the meeting with the female on the right.
http://flic.kr/p/fuWnua
Pre-processing
remove stopwords
normalisation or urls, hashtags, mentions
wood pigeon columba palumbus who make up meeting female right flickr
Feature Extraction
numerical representation 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 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).

processing steps (see Sections 2.1 and 2.2) we gather Tweets related to wildlife, from which stop words are removed, tokens normalised, and duplicates are removed. During *Feature Extraction* (see Section 2.3) we build word feature vectors for the corpus using techniques based on the feature representation approaches described in Section 1.1.2. 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 2.4). Finally, we experiment with three classifiers for building a *Wildlife Observation Classifier* (Section 2.5). These are based on the main supervised machine learning approaches explained in Section 1.1.

2.1. 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. Thus, 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 2.6, Table 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.

2.2. Tweets Pre-processing

Cleaning Tweets Stanford NLP Core is used for pre-processing the dataset, in particular for part of speech tagging. Stop words were removed using the NLTK stop word list. Following tokenization 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 and perform part-of-speech tagging in order to identify noun phrases. 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. Represent Tweets as bag-of-words (BOW)
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 one of the Tweets, then the two Tweets are considered similar and the Tweet with the flagged up threshold is removed.

2.3. Feature Extraction

We performed experiments with three main types of feature extraction techniques, as described in Table 1. They are reflective of the main approaches for building feature representations, identified earlier in the paper, 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. We have performed experiments with three pre-trained word embedding models, Word2Vec (Mikolov et al., 2013a), fastText (Bojanowski et al., 2017) and GloVe (Pennington et al., 2014). Word2Vec embeddings (Mikolov et al., 2013a) are generated using a skip-gram approach for learning term embeddings from raw text. 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. We have also included GloVe pre-trained embeddings as they have 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., 2018) introduced in Section 1.1.1. As explained in Section 1.1.2, 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. As explained in Section 1.1.2 there are two steps in the BERT architecture: pre-training and fine-tuning, both of which we use here in different experiments. We use the base pre-trained BERT model to create a

Table 1
Feature Extraction Step — A summary of main methods deployed during this step.

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 unidirectional 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., 2018)	We use the base pre-trained BERT language model, which has been trained on Books Corpus and English Wikipedia

sentence encoding (see next section) that will be used as input to a Logistic Regression classifier, while also using the fine-trained version, with BERT's built-in classifier, in subsequent experiments.

2.4. Feature Integration

In this step, we generate Tweet classification feature vectors using three main approaches. 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[2]. This can be regarded as a bag of words (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 2.

2.5. Wildlife Observation Classifier

We use three types of classifier where each classifier represents one of the main text classification methods outlined in Section 1.1.1. 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 in the previous section. Thus these features include sentence representations based on average pre-trained Word2Vec,

[2] 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

Table 2
Feature Integration Step — A summary of main methods deployed during this step.

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 3
Wildlife Observation Classifier — A summary of classification approaches used to build a classification model.

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) including 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., 2018)	We use pre-trained BERT word representation model and add a final sequence classification layer

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 3

2.6. Dataset

We selected a subset of the Tweets collection 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 the first author. 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 0.978 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 3.3 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. 1769384) 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.[3]. An overview of the manually labelled subset of the Tweets collection which was used for training the classifier is presented in Table 5. Analysis of the

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

Table 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.

#Tweets	1769384
#Tokens	31780390
Avg Length	18

Table 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).

	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 6
Tweets distribution per species — limited to the 20 best represented species on Twitter ('#Tweets' refers to number of Tweets per species).

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

distribution of Tweets per species (see Table 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, Mallard).

3. Results

3.1. Evaluation experiments

As mentioned in Section 2, 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., 2018) 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

Table 7
Results per classification approach ('p' refers to precision, 'r' refers to recall).

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%)
Logistic Regression	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%)

(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 ensures that each class is (approximately) equally represented across each test fold.

3.2. Classification Results

The baseline classifier based on frequency scores of n-grams as features provided remarkably good precision 93.33% and recall 95.07% (see Table 7). The feature extraction method based on noun phrase and named entity terms rather than 2-gram representation did 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 96.1% and an f1-score of 94.7%.

3.3. Indicative Features Analysis

We performed an analysis on the features indicative for wildlife using the manually annotated Tweets. The results in Figures 3-5 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.

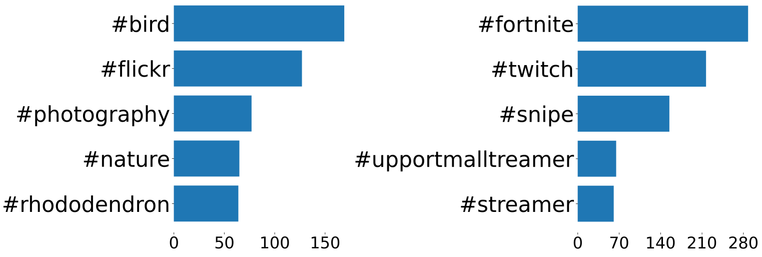

Figure 3: The ten most frequent hashtags per class label, Tweets with genuine wildlife observations (left), Tweets with false wildlife Tweets (right).

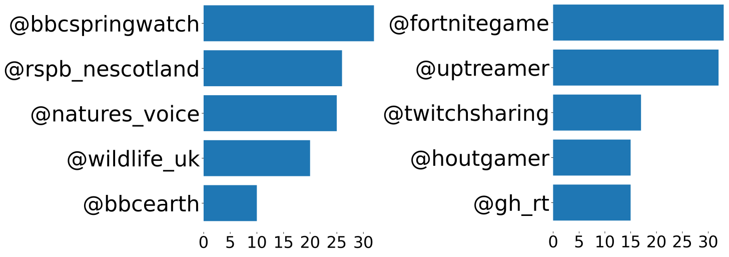

Figure 4: The ten most frequent mentions per class label, Tweets with genuine wildlife observations (left), Tweets with false wildlife Tweets (right)

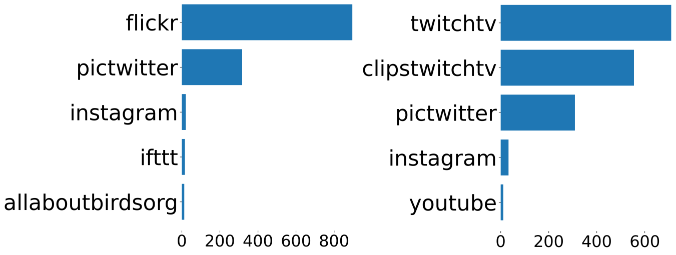

Figure 5: The ten most frequent URL links class label, Tweets with genuine wildlife observations (left), Tweets with false wildlife Tweets(right)

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 3 to 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 a wildlife observation. There is a single false wildlife Tweet with a wildlife indicative URL (i.e., *'instagram'*).

Table 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

	'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

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 fine-tuned BERT model.

3.4. 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 3.1). A confusion matrix of the performance of the classification models is given in Table 8.

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 and also a mix between English and other languages. A false positive example of the latter for fastText, where BERT correctly classifies the Tweet as 'Wildlife' observation, is: *'le petit oison des bernaches du canada branta canadensis canada goose pic.twitter.com/'*.

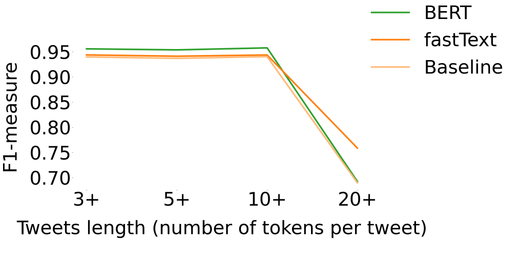

Figure 6: **Comparison between the performance of baseline, fastText, and BERT classifiers for different length of Tweets.**

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 6).

4. Discussion

Classification results and error analysis presented in Sections 3.2 and 3.4 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 pre-trained embeddings of GloVe which were trained on 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 it 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 ecology-related data), compared to corpus-trained word embeddings which are trained on a smaller but more task-specific dataset.

It may also be noted that 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) and in correctly classifying 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, the BERT deep learning 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 1.2 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 1.2

Analysis of 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, but they are not indicative of Tweets with actual 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 alone of some of these hashtags cannot be 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 the language models with contextual word embeddings employed here to identify wildlife observations, as opposed to simply selecting Tweets according to the presence of particular terms, allows us to identify the semantics of terminology used to make wildlife observations, rather than being dependent on a fixed vocabulary. Thus terms with similar meaning will have similar representations which can help accurate classification despite the diverse spelling or diversity of terminology. This allows our methods to be applied to a wider range of species, geographical regions and even different languages than would otherwise be the case. It is also possible to envisage that, in future, classification models could be improved by creating feature selection techniques which assign higher importance to the sort of indicative features identified in this work.

The statistical analysis, presented in Table 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 (Edwards et al., 2021; August et al., 2020) which suggests that there are common trends among different social networks on the type of species they represent. A more detailed analysis of the value of Twitter for collecting species-specific data is outside the scope of this research. 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.

5. Conclusion

In this work 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. We performed experiments with three classification approaches: a classical (linear) Logistic Regression, the fastText pipeline and the fine-tuned BERT transformer 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. The best performance of .96 for each of precision, recall and F1 score was obtained using the fine-tuned base BERT model in which word embeddings are adapted to their context. In particular, the fine-tuned BERT model proved 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 in 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.

References

Amano, T., Lamming, J.D., Sutherland, W.J., 2016. Spatial gaps in global biodiversity information and the role of citizen science. Bioscience 66, 393–400.

Antoniou, V., Fonte, C.C., See, L., Estima, J., Arsanjani, J.J., Lupia, F., Minghini, M., Foody, G., Fritz, S., 2016. Investigating the feasibility of geo-tagged photographs as sources of land cover input data. ISPRS International Journal of Geo-Information 5, 64.

Aristeidou, M., Herodotou, C., Ballard, H.L., Young, A.N., Miller, A.E., Higgins, L., Johnson, R.F., 2021. Exploring the participation of young citizen scientists in scientific research: The case of inaturalist. Plos one 16, e0245682.

Arora, S., Liang, Y., Ma, T., 2017. A simple but tough-to-beat baseline for sentence embeddings, in: 5th International Conference on Learning Representations, ICLR 2017, p. 16.

August, T.A., Pescott, O.L., Joly, A., Bonnet, P., 2020. AI naturalists might hold the key to unlocking biodiversity data in social media imagery. Patterns 1, 100116.

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

Blight, A.J., Allcock, A.L., Maggs, C.A., Johnson, M.P., 2009. Intertidal molluscan and algal species richness around the uk coast. Marine ecology progress series 396, 235–243.

Bojanowski, P., Grave, E., Joulin, A., Mikolov, T., 2017. Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics 5, 135–146.

Bouazizi, M., Ohtsuki, T., 2019. Multi-class sentiment analysis on twitter: Classification performance and challenges. Big Data Mining and Analytics 2, 181–194. doi:10.26599/BDMA.2019.9020002.

Cer, D., Yang, Y., Kong, S., Hua, N., Limtiaco, N., John, R.S., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., Sung, Y., Strope, B., Kurzweil, R., 2018. Universal sentence encoder. CoRR abs/1803.11175. URL: http://arxiv.org/abs/1803.11175, arXiv:1803.11175.

Chen, J., Hu, Y., Liu, J., Xiao, Y., Jiang, H., 2019. Deep short text classification with knowledge powered attention, in: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 6252–6259.

Cohn, J.P., 2008. Citizen science: Can volunteers do real research? BioScience 58, 192–197.

Çöltekin, Ç., Rama, T., 2018. 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, pp. 34–38.

Conneau, A., Kiela, D., Schwenk, H., Barrault, L., Bordes, A., 2017. Supervised learning of universal sentence representations from natural language inference data, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Copenhagen, Denmark. pp. 670–680. URL: https://www.aclweb.org/anthology/D17-1070, doi:10.18653/v1/D17-1070.

Daume, S., 2016. Mining twitter to monitor invasive alien species—an analytical framework and sample information topologies. Ecological Informatics 31, 70–82.

Daume, S., Albert, M., von Gadow, K., 2014. Forest monitoring and social media–complementary data sources for ecosystem surveillance? Forest Ecology and Management 316, 9–20.

Davis, A., Major, R.E., Taylor, C.E., Martin, J.M., 2017. Novel tracking and reporting methods for studying large birds in urban landscapes. Wildlife Biology 2017.

Deng, X., Li, Y., Weng, J., Zhang, J., 2019. Feature selection for text classification: A review. Multimedia Tools & Applications 78.

Devlin, J., Chang, M.W., Lee, K., Toutanova, K., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 .

Di Minin, E., Fink, C., Tenkanen, H., Hiippala, T., 2018. Machine learning for tracking illegal wildlife trade on social media. Nature ecology & evolution 2, 406.

Di Minin, E., Tenkanen, H., Toivonen, T., 2015. Prospects and challenges for social media data in conservation science. Frontiers in Environmental Science 3, 63.

Edwards, T., Jones, C.B., Perkins, S.E., Corcoran, P., 2021. Passive citizen science: The role of social media in wildlife observations. Plos one 16, e0255416.

ElQadi, M.M., Dorin, A., Dyer, A., Burd, M., Bukovac, Z., Shrestha, M., 2017. Mapping species distributions with social media geo-tagged images: case studies of bees and flowering plants in australia. Ecological informatics 39, 23–31.

Estima, J., Fonte, C.C., Painho, M., 2014. Comparative study of land use/cover classification using flickr photos, satellite imagery and corine land cover database, in: 17th Conference on Geographic Information Science, p. 4.

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

Fan, R.E., Chang, K.W., Hsieh, C.J., Wang, X.R., Lin, C.J., 2008. Liblinear: A library for large linear classification. Journal of machine learning research 9, 1871–1874.

Fritz, S., McCallum, I., Schill, C., Perger, C., See, L., Schepaschenko, D., Van der Velde, M., Kraxner, F., Obersteiner, M., 2012. Geo-wiki: An online platform for improving global land cover. Environmental Modelling & Software 31, 110–123.

Gambäck, B., Sikdar, U.K., 2017. Using convolutional neural networks to classify hate-speech, in: Proceedings of the First Workshop on Abusive Language Online, Association for Computational Linguistics, Vancouver, BC, Canada. pp. 85–90. URL: https://www.aclweb.org/anthology/W17-3013, doi:10.18653/v1/W17-3013.

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

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

Howard, J., Ruder, S., 2018. 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), pp. 328–339.

Huang, X., Li, Z., Wang, C., Ning, H., 2019. Identifying disaster related social media for rapid response: a visual-textual fused cnn architecture. International Journal of Digital Earth 0, 1–23. URL: https://doi.org/10.1080/17538947.2019.1633425, doi:10.1080/17538947.2019.1633425, arXiv:https://doi.org/10.1080/17538947.2019.1633425.

Jauhiainen, T., Lui, M., Zampieri, M., Baldwin, T., Lindén, K., 2019. Automatic language identification in texts: A survey. Journal of Artificial Intelligence Research 65, 675–782.

Jeawak, S.S., Jones, C.B., Schockaert, S., 2017. 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.

Jeawak, S.S., Jones, C.B., Schockaert, S., 2018. Mapping wildlife species distribution with social media: Augmenting text classification with species names. 10th International Conference of Geographic Information Science (GIScience 2018) .

Jeawak, S.S., Jones, C.B., Schockaert, S., 2019. Embedding geographic locations for modelling the natural environment using flickr tags and structured data, in: Azzopardi, L., Stein, B., Fuhr, N., Mayr, P., Hauff, C., Hiemstra, D. (Eds.), Advances in Information Retrieval, Springer International Publishing, Cham. pp. 51–66.

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

Joachims, T., 1998. Text categorization with support vector machines: Learning with many relevant features, in: European conference on machine learning, Springer. pp. 137–142.

Joulin, A., Grave, E., Bojanowski, P., Mikolov, T., 2017. 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, Association for Computational Linguistics. pp. 427–431.

Kent, J.D., Capello Jr, H.T., 2013. Spatial patterns and demographic indicators of effective social media content during thehorsethief canyon fire of 2012. Cartography and Geographic Information Science 40, 78–89.

Leivesley, J.A., Stewart, R.A., Paterson, V., McCafferty, D.J., 2021. Potential importance of urban areas for water voles: Arvicola amphibius. European Journal of Wildlife Research 67, 1–4.

Leung, D., Newsam, S., 2012. Exploring geotagged images for land-use classification, in: Proceedings of the ACM multimedia 2012 workshop on Geotagging and its applications in multimedia, pp. 3–8.

Levenshtein, V.I., 1966. Binary codes capable of correcting deletions, insertions, and reversals, in: Soviet physics doklady, pp. 707–710.

Li, H., Caragea, D., Li, X., Caragea, C., 2018. 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 , 13.

Lowry, C.S., Fienen, M.N., 2013. Crowdhydrology: crowdsourcing hydrologic data and engaging citizen scientists. GroundWater 51, 151–156.

Martinc, M., Pollak, S., 2019. Combining n-grams and deep convolutional features for language variety classification. Natural Language Engineering

25, 607–632.

McCallum, A., Nigam, K., et al., 1998. A comparison of event models for naive bayes text classification, in: AAAI-98 workshop on learning for text categorization, Citeseer. pp. 41–48.

McCann, B., Bradbury, J., Xiong, C., Socher, R., 2017. Learned in translation: Contextualized word vectors. Advances in Neural Information Processing Systems 30.

Merity, S., Keskar, N.S., Socher, R., 2018. Regularizing and optimizing lstm language models, in: International Conference on Learning Representations, pp. 1–13.

Merkx, D., Frank, S.L., 2020. Comparing transformers and rnns on predicting human sentence processing data. arXiv e-prints , arXiv–2005.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J., 2013a. Distributed representations of words and phrases and their compositionality, in: Advances in neural information processing systems, pp. 3111–3119.

Mikolov, T., Yih, W.t., Zweig, G., 2013b. 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, pp. 746–751.

Mohammad, S., Bravo-Marquez, F., Salameh, M., Kiritchenko, S., 2018. Semeval-2018 task 1: Affect in tweets, in: Proceedings of the 12th international workshop on semantic evaluation, pp. 1–17.

Monkman, G.G., Kaiser, M.J., Hyder, K., 2018. Text and data mining of social media to map wildlife recreation activity. Biological conservation 228, 89–99.

Palomino, M., Taylor, T., Göker, A., Isaacs, J., Warber, S., 2016. 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, 142.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al., 2011. Scikit-learn: Machine learning in python. Journal of machine learning research 12, 2825–2830.

Peng, T.B.A.C.P., Dean, X.F.J.O.J., 2007. Large language models in machine translation. EMNLP-CoNLL 2007 , 858.

Pennington, J., Socher, R., Manning, C., 2014. Glove: Global vectors for word representation, in: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pp. 1532–1543.

Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L., 2018. Deep contextualized word representations, in: Proceedings of NAACL-HLT, pp. 2227–2237.

Poria, S., Cambria, E., Hazarika, D., Vij, P., 2016. A deeper look into sarcastic tweets using deep convolutional neural networks. arXiv preprint arXiv:1610.08815 .

Reynard, D., Shirgaokar, M., 2019. 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.

Scherrer, Y., Ljubešić, N., et al., 2021. 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. p. 135–140.

Soliman, A., Soltani, K., Yin, J., Padmanabhan, A., Wang, S., 2017. Social sensing of urban land use based on analysis of twitter users' mobility patterns. PloS one 12, e0181657.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need, in: Advances in neural information processing systems, pp. 5998–6008.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Brew, J., 2019. Huggingface's transformers: State-of-the-art natural language processing. ArXiv abs/1910.03771.

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

Xu, Q., Li, J., Cai, M., Mackey, T.K., 2019. Use of machine learning to detect illegal wildlife product promotion and sales on twitter. Frontiers in Big Data 2, 28.

Yang, Z., Dai, Z., Salakhutdinov, R., Cohen, W.W., 2018. Breaking the softmax bottleneck: A high-rank rnn language model, in: International Conference on Learning Representations, pp. 1–18.

Zhong, X., Enke, D., 2019. Predicting the daily return direction of the stock market using hybrid machine learning algorithms. Financial Innovation 5, 1–20.

Appendix A: Glossary

Neural Network: Neural networks are machine learning algorithms inspired by the structure of the human brain. Neural networks are comprised of layers of nodes, 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[4].

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.

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

Recurrent Neural Network (RNN): RNNs are feed-forward NNs, which process text in a sequential manner where sentences are processed word by word. Previous input is represented as the hidden state of the recurrent computation and each new input is processed and combined with the 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 between 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 architectures address the problems associated with earlier neural network models by using an attention mechanism where each word representation in a sentence 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. They can be applied directly or be 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 in various ways including shallow 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 the Books corpus and Wikipedia data. This pre-trained model can be fine-tuned for various tasks by adding a single output layer.

GloVe: An embedding model where a weighted least squares regression dimensionality reduction procedure uses a co-occurrence counts matrix. For this paper, we used the 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.