User Engagement Triggers in Social Media Discourse on Biodiversity Conservation

NINA DETHLEFS, University of Hull, United Kingdom
HERIBERTO CUAYÁHUITL, University of Lincoln, United Kingdom

Studies in digital conservation have increasingly used social media in recent years as a source of data to understand the interactions between humans and nature, model and monitor biodiversity, and analyse online discourse about the conservation of species. Current approaches to digital conservation are for the most part purely frequentist, i.e. focused on easily trackable and quantifiable features, or purely qualitative, which allows a deeper level of interpretation, but is less scalable. Our approach aims to evaluate the applicability of recent advances in deep learning in combination with semi-automatic analysis. We present a multimodal neural learning framework that experiments with different combinations of linguistic and visual features and metadata of tweets to predict user engagement from a function of *likes* and *retweets*. Experimental results show that text is the single most effective modality for prediction when a large amount of training data is available. For smaller datasets, drawing information from multiple modalities can boost performance. Notably, we find a negative effect of large pre-trained language models when dealing with substantially unbalanced datasets. A qualitative analysis into the triggers of user engagement with tweets reveals that it emerges from a combination of online discourse topic and sentiment, and is often amplified by user activity, e.g. when content originates from an influencer account. We find clear evidence of existing sub-communities around specific topics, including *animal photography and sightings*, *illegal wildlife trade and trophy hunting*, *deforestation and destruction of nature* and *climate change and action* in a broader sense.

CCS Concepts: • **Computing methodologies** → **Artificial Intelligence**; Machine learning; • **Applied Computing** → Document management and text processing; • **Social and professional topics** → User characteristics.

Additional Key Words and Phrases: social media analysis, user engagement, multimodal learning, biodiversity conservation, neural networks, large language models

ACM Reference Format:
Nina Dethlefs and Heriberto Cuayáhuitl. 2023. User Engagement Triggers in Social Media Discourse on Biodiversity Conservation. In . ACM, New York, NY, USA, 32 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Social media has served as a rich source of data for studies in conservation science in recent years. Research includes the analysis of images from social media platforms such as Flickr, Instagram and others as well as textual content, e.g. from Twitter, to identify places in nature that humans travel to, species they observe and issues in conservation that are raised in online discussion. Toivonen et al. [82] provide a recent and insightful overview of the use of social media data in conservation science. They categorise existing research into three broad categories: (1) studies on people in nature that aim to understand the interactions between humans and nature, including places that humans visit, value and why; (2) studies in biodiversity monitoring which often focus on data collection, such as sightings and geo-tagging of particular species, and (3) online discussion, which is a broad term encompassing any form of online conversation or

Manuscript submitted to ACM

discourse about conservation, animals or nature, without the in-situ element of the former two. This article aims to contribute to research in the latter of these categories, the analysis of online discussion of conservation-related topics on Twitter.

Specifically, we draw a comparison between data-driven approaches that analyse social media content based on automatically observable features, such as keywords, geo-tags or the presence of images, and qualitative methodologies, such as deep linguistic analysis, or social network models. We see our contribution in bridging the gap between these contrasting methodologies, in generating a deeper understanding of domain discourse dynamics than is possible using data-driven approaches alone, yet developing models that are transferable across datasets without extensive annotation or modelling, and therefore lend themselves to real-time social media analysis. The latter is an important requirement for monitoring in digital conservation, or any social dynamics online. We hope to make a cross-disciplinary contribution to studies in social computing, digital conservation and computational linguistics.

Recent social science research [94] has revealed a 25% threshold to social tipping points, i.e. points of social transformation where change occurs rapidly and suddenly and individual members of society adopt views and/or behaviours that were previously dominated by margin groups. Such social tipping points can relate to technology and energy systems, political, financial or economic trends, or to the general discourse on climate change. This article aims to investigate, from an AI and data-driven perspective, how social media, specifically Twitter, is used by members of the online community to influence the discourse on conservation through textual and visual content. Twitter (known as X since July 2023) was chosen as a data source due to its wide user base, combination of text and image-based content, its adoption in previous studies for comparability [7, 10, 12, 61, 63, 71, 81, 82], and API access for research purposes. We will continue to refer to the platform as Twitter in this article, as it was known during our data collection and research. It is clear from previous studies has conservation-relevant discourse is increasingly taking place on Twitter, in the form of positively-natured activism, as well as malicious wildlife trade, making it a relevant platform for data collection and analysis. Understanding how users react to Twitter-based content can help direct and support conservation action and campaigns. We are especially interested in the multimodal features that characterise such online discourse, including linguistic and visual features of posted content, as well as metadata associated with the user and tweet. In essence, we aim to discover important drivers of user engagement (in the form of *likes* and *retweets*) in online conservation discourse. We attempt to generalise from these features and develop a deep learning framework that can accurately predict user engagement for a given tweet from its multimodal profile. The following research questions form the basis of this article:

(1) What are the defining and recurring topics in social media discourse around the conservation of species?
(2) Who are the sub-communities that participate in such discourse and what are their identifiable characteristics?
(3) What are identifiable (linguistic, visual or meta) characteristics of tweets that function as triggers of online user engagement?
(4) To what extent can recent advances in deep learning for text and image analysis form an effective basis for user engagement prediction?

We present a multimodal deep learning framework that aims to predict user engagement from a combination of text, image, and metadata features. We utilise the most recent architectures in natural language and image processing, and also compare the use of large pre-trained resources, such as language models or image weights. Our experiments demonstrate that text alone is the most effective modality for prediction and outperforms other modalities and combinations by as much as 25% in terms of balanced accuracy. This only holds when sufficient training data is available though. With limited data, combining multiple modalities can help boost performance, where a model that jointly learns from all three modalities achieves second-best performance at 66%, which is 4% better than the next model. Large pre-trained models for language or image processing were found to be less effective, with the language models particularly struggling to learn from unbalanced data.

While overall, we are able to demonstrate some success with recent neural network models for natural language processing, particularly transformer networks, our experiments confirm the findings in other studies that the most relevant insights can be drawn from hybrid methods, i.e. that combine purely AI-driven methods with an element of qualitative analysis. A set of manual annotations on a subset of our dataset were able to uncover deeper patterns of user engagement, that were not apparent from frequency-based methods alone. Our qualitative findings are in line with earlier research that has shown the importance of sentiments for user engagement classification, but not valence, i.e. the strength of the sentiment. Overall, our experiments reveal that user engagement emerges from a combination of user activity on Twitter and the sentiment and topic of the discourse. We find clear evidence of sub-communities of users that engage with specific content, e.g. *wildlife crime*, *deforestation* or *animal sightings*, often driven by influencer behaviour. Quantity-only metrics, such as the number of hashtags, URLs or emojis used in a tweet, were not found to carry much predictive weight, and neither were purely image-based features.

This article is structured as follows. Section 2 discusses related work on digital conservation and analysing the popularity of social media contributions. Section 3 presents details on data collection and labelling and shows basic statistics of the dataset that will be used for analysis. We present our methodology in Section 4, and discuss experiments and results in Section 5. This will involve a quantitative evaluation of our deep learning models, as well as a qualitative analysis of the driving features in digital conservation discourse. We offer a discussion of the findings and drawbacks of our research in Section 7 and finally present conclusions and future work in Section 8.

2 RELATED WORK

In this section we aim to provide a methodological comparison of work in digital conservation studies (Section 2.1) as well as highlight existing findings on what drives popularity of social media contributions in general and across different domains (Section 2.2). We highlight pathways towards significant progress in the automatic and real-time analysis of social media content for conservation science by drawing more heavily on recent advances in deep learning and natural language processing to aid rapid progress.

2.1 Social media analysis for digital conservation

Social media data has been a rich source of insights in studies in digital conservation. Predominant approaches mostly opt for frequency-based analysis of textual or visual content, purely qualitative analysis, or sometimes a hybrid approach that combines these two.

2.1.1 Frequency-based analysis of social media data for digital conservation. A common application of social media analysis in conservation science has been the quantification of visits to particular places, often in nature, as well as the value that humans attribute to them, e.g. as a source of well-being and mental health, or as a venue for leisure activities. In this line of research, it has been common to infer the value of a place in terms of the frequency of visits and social media posts made either about it or from it (based on geolocation). Wood et al. [95] and Gliozzo et al. [26] are relevant studies that attempt to establish a link between human well-being in nature (inferred from post frequency) and wildlife conservation. In a more recent study, Väisäanen et al. [86] also demonstrate the use of various image analysis methods to extract semantic patterns and content from geotagged photographs, with the aim of understanding human activities and interactions with nature.

Apart from analysing post frequencies in relation to geographical location, multiple studies have also explored counting images posted from certain locations, often using an element of geo-tagging [26, 30, 88], or alternative means of determining the location of an image [34]. Some authors have also used keyword search as a data collection tool [7, 57], and identified trends on location popularity by tracking the frequency of keywords on social media. As an example of frequency-based analysis, van Zanten et al. [88] count the number of images posted on photo sharing platforms Panoramio, Flickr and Instagram across Europe and find that, across countries, the frequency of posts seem guided by the accessibility of a landscape, population density, the degree of mountainous terrain and proximity to water, amongst others. Frequency-based methods have been considered as an alternative to traditional high-precision visitor surveys, e.g. by van Zanten et al. [88], as well as Tenkanen et al. [81], who also discuss potential sources of deviation between automatic estimates and surveys. In contrast to these optimistic studies, Levin et al. [45] warn that crowdsourced data can be an unreliable measure for areas that are not generally used for human leisure activities.

Social media analysis has also been used increasingly in recent years as a tool to identify and investigate wildlife crime. This can include the trade of animals (or animal parts) and plants as food, pets, medicines, clothing or trophies. As an example, Eid et al. [21] analyse Facebook posts to identify illegal hunting activities in Jordan. Di Minin et al. [57] use keyword search across a range of platforms in multiple languages to uncover trading-related content. The latter of these studies again demonstrates the significance of frequency-based methods. The authors predefined search terms around specific animal names or animal-based "products", e.g. scales, horns, furs, to identify and track occurrences on social media and flag potentially problematic content. Xu et al. [97] followed a similar path of research and discovered a wider set of keywords (or codewords) in multiple languages that are commonly used in wildlife trade activities online. Fink et al. [25] demonstrate how wildlife trade, in their case of Indonesian songbirds, can be tracked online using web scraping, and can potentially offer opportunities to influence the trade towards more sustainable practices. Apart from using language, recent studies have demonstrated the potential of image analysis methods to identify and monitor illegal wildlife trade. For example, Kulkarni and Di Minin [41] apply deep image analysis to identify exotic animals on sale. Interestingly, the authors show that a key feature for the learning models is to recognise when animals are placed outside their natural environment. In a related study, Cardoso et al. [13] show that state-of-the-art convolutional neural networks can also identify traded pangolins (or their parts) with reasonable accuracy.

Frequency-based methods do not normally analyse the actual contents of posts, such as images or text, and rely only on the occurrence of data points for analysis. This has the advantage of generating basic insights fast, but can compromise the quality of data at the same time, both in terms of noisy data (i.e. including data that is not actually thematically related), as well as missing data (e.g. from lexically or semantically similar keywords that were omitted from the search). For example, somebody posting on Twitter may be commenting on a news headline and be in the area by coincidence without necessarily reflecting an appreciation of the specific geographical spot they are tweeting from. Similarly, keyword search will often find posts that are unrelated to an intended topic - "hedgehog" refers frequently to video game character "Sonic", "jaguar" is often a discussion about cars, and public interest in "reindeer" tends to peek around Christmas. Similarly, references to wildlife crime will mostly lead to general discussion threads of people condemning such activities, and can outweigh the much smaller number of posts that actually offer the sale of illegally poached animal products.

2.1.2 Automatic processing with open access knowledge resources . To increase the reliability of insights that can be drawn from social media data, some studies have combined pure frequency analysis with other data sources, e.g. information about the presence of night lights to identify urban areas [44], or open-source GIS data to better model the geographical context of posts [45]. Again other studies have integrated the use of sentiment analysis to gain a deeper understanding of positive or negative values associated with places and potential triggers of such sentiments. In a study on the Great Barrier Reef, Becken et al. [7] retrieve geo-tagged Twitter posts containing any of a set of predefined keywords relating to reef monitoring or reef-related activities, such as the state of coral and water, sightings of marine life or leisure activities. Using a lexicon-based sentiment analysis approach, the authors find that little information is given that could be used for reef monitoring. Tweets seemed overall positive, which the authors speculate is at least partially due to a bias towards touristic visitors tweeting about their experiences. The overall positive stance that humans take towards certain forms of wildlife and conservation activities online is confirmed in a recent study by van Houten et al. [33] in the area of scientific publishing. The authors show based on the automatic reading of scientific journal abstracts on the reintroduction of species that sentiments have become increasingly positive over the last 40 years, potentially indicating the growing success of conservation programs.

Hybrid methods that combine an automatic element of processing, e.g. frequency analysis, with a knowledge-based component, i.e. mostly a hand-curated resource, generally aim to filter or in some way structure the information they may obtain from a purely statistical approach. This can be successful but is dependent on the quality of the resource. As an example, sentiment analysis lexicons, which have been used in a number of conservation studies, have often been curated carefully over long periods of time and are therefore of high quality. They still struggle though with ambiguity (e.g. negative words that are used in a positive context "wicked", "insane"), with negation ("not good"), with unknown words, paraphrases, or humorous, ironic or sarcastic contents, which are notoriously difficult to spot computationally. Recent advances in statistical language models [56, 65], especially those that can model the context of linguistic content [18, 67] can circumvent some of these problems. Also, Kulkarni and Di Minin [40] demonstrate success using recent natural language processing techniques to identify news articles and social media content that are relevant to specific topics of interest in biodiversity monitoring, which can be valuable for data collection.

2.1.3 Qualitative analysis and interpretability . Purely qualitative approaches lie somewhat at the opposite end of the automation spectrum in comparison to the approaches discussed so far. While data collection is still done through an API, analysis is human-guided and manual. Qualitative analysis often delivers meaningful findings but faces constraints on the amount of data that can be incorporated. Representative examples that apply this form of analysis to research on digital conservation include e.g. Willemen et al. [92], who analyse online photographs of IUCN Red List endangered species to infer their popularity. Hausmann et al. [30] analyse labelled images of animals to compare social media observation surveys against traditional surveys. The authors discuss particularly the risks of bias of an online vs an in-situ community of wildlife observers. Barry [6] analyses people's reaction to images of grazing cows and other livestock. A case in point for this type of research is a study by Macdonald et al. [53] that seeks to understand behaviour triggers via social media analysis on the sudden and world-wide attention to the killing of a lion in a National Park by a trophy hunter. The authors point to idiosyncratic features in the narrative as likely sources, such as the lion's English nickname "Cecil", the identifiability of the killer as a Western trophy hunter or the circumstances of the lion being lured into his death. Understanding such triggers can be of vital importance for conservation efforts as they shine a light on exactly what causes a willingness in humans to condemn wildlife crime or take action against it. Identifying such broad factors and generalising them into a systematic framework for the understanding of wildlife crime and public reaction can play an important role in designing campaigns deliberately and gaining public support for conservation projects.

Specifically in the field of social media analysis, a complementary strand of research focuses on social network analysis, as commonly carried out in social science research [91]. Social network analysis typically aims to represent user networks based on metadata and engagement, e.g. likes, retweets, followers, conversational threads, etc. which allows researchers to reconstruct various network dynamics. For example, in the domain of climate change and biodiversity conservation, previous research has shown a strong relationship between social media users' individual features, such as their political orientation [46, 87], socioeconomic status [19], willingness to form risky beliefs [38], and certain views on climate change and biodiversity, i.e. specifically whether they support action or not. In the context of interaction patterns, it has also been shown that a majority of online forums are "internally coherent", i.e. supporters and sceptics of climate change action both have a preference to interact within their own group [93], a phenomenon which previous studies have referred to as *echo chambers*. Anderson and Huntington [1] show that sentiment in open/mixed opinion forums is often more negative [23] than in echo chambers, while more recent research by Tyagi et al. [85] reveals a trend for activists to increasingly interact outside their group. Related to the topic of content and sentiment analysis, there is an active strand of research that looks at framing in climate change discourse, i.e. the concepts that social groups draw on when discussing or posting about climate change and conservation, see e.g. a study by Hopke and Hestres [32] on framing of global warming in terms of crisis discourse.

In summary, qualitative research can generate invaluable insights, e.g. in the case of social network analysis on the relationships between individual attributes and observable behaviour on social media. However, qualitative data analysis is expensive, time-consuming and resource-intensive to conduct, and therefore applicable in practice only to limited and specific research questions and datasets. This is particularly relevant for studies in framing, which rely on high-quality manual expert annotation, and do not transfer well across domains or datasets due to the specialised nature of online discourse forums.

2.2 Popularity prediction of social media contributions

This section will review existing research on predicting the popularity, or user engagement, with different social media content, and the underlying factors that seem to drive such engagement. Overall, it appears that different domains of online discourse operate with different patterns of engagement and participation, though some general rules are also discovered. As an example from the domain of digital conservation, Papworth et al. [63] present findings on what drives social media uptake of new pieces of conservation research, when they appear. While they find that the scientific journal has the highest influence, there is a small trend for mammals and "charismatic" animals to be featured over other species. This observation is confirmed in research by Roberge [71] who discovered a bias towards larger animals, often mammals. Findings by Di Minin et al. [59] are more balanced. The authors discover that preferences for particular species differ between different groups of users, with some travelling to see the "Big Five" while others are more interested in a broader range of species. This is confirmed by Hausmann et al. (2017) [30] who found that factors including the socio-economic condition of countries, geographical characteristics, accessibility and proximity to civilisation were more important indicators of touristic visits (as estimated from Instagram) than the presence of charismatic species such as the "Big Five". Fink et al. [24] use sentiment analysis to show a correlation between social media and news coverage of conservation events, and observe particularly strong sentiments in tragic outlier events, such as the extinction of a species. Other studies have analysed user engagement, or post popularity, in a broader range of domains beyond digital conservation.

In recent related work, Mahdikhani [54] uses linguistic features to predict the popularity of a tweet, defined in terms of its retweets. The author presents a model for a dataset of 1.25M tweets during the Covid-19 pandemic and compares the use of topic modelling (LDA [8]), TF-IDF word vectors [78], word embeddings [56], bag-of-word models [29], and combinations of the above as input features to a voting ensemble of classifiers. The topic modelling phase is used particularly to induce content features of tweets, such as the emotion in a tweet (fear, anger, job, sadness) paired with its valence score for intensity. It is shown that highly emotional tweets attract higher popularity than more information-based tweets. Mahdikhani's research is in line with earlier work that has shown that content-based features tend to have higher predictive power than e.g. the number of followers of a tweeting account on its own [42]. In an early study on Twitter, Naveed et al. [61] find that hashtags, URLs and usernames increase the likelihood of a tweet being retweeted. The authors also confirm the importance of emotional features, where negative emotions tend to increase retweets, as well as positive tweets with high arousal or dominance (i.e. indicating exciting or intense news). Emojis were found to intensify these trends, the use of rude words de-intensifies them. Question marks were found to be a further indicator of retweets. The engagement with tweets has also been found to be contextually dependend. In the political domain, Rivadeneira et al. [70] find that sentiment can help predict engagement with tweets, but the polarity (positive / negative) depends on a user's political views.

The importance of URLs and emojis were also shown in a recent study by Chung et al. [16]. The authors investigate the engagement with tweets particularly targeted at women by analysing tweets from the Women Can Code theme. The study reveals that information-based tweets receive more engagement than action or community tweets. This finding seems contrary to work by Mahdikhani [54] above, who found a negative tendency for engagement with information-based tweets, though did not consider the additional dimensions of action or community, and focused on emotion intensity instead. Chung et al. further observe a positive effect from emojis and URLs, and a negative effect for engagement with hashtags and videos. Using images or photos had no demonstrated effect in this study.

Other studies in this area have focused on other particular domains and use cases, particularly the effectiveness of marketing and branding on social media. For example, Guidry et al. [28] present a focused study in the area of social marketing, specifically investigating which type of tweets from non-profit organisations tend to attract the most stakeholder engagement. The purpose was to create a model that can help non-profit organisations engage with the public effectively. In this context, the authors found that public information tweets were likely to receive more engagement than marketing tweets, and that call-to-action tweets were more popular than fundraising or event promotion.

Also focused on marketing, Zadeh and Sharda [99] present a mathematical framework to model the popularity of tweets in the context of branding. The authors model popularity through variables *likes*, *retweets* and *replies*. They observe state-of-the-art performance against comparable frameworks and prefer a model that predicts good user engagement soon after a post was made. While the model achieved good prediction accuracy, the authors do not provide an analysis of content features beyond temporal relationships. Zadeh and Sharda's study can be seen as part of a cluster of approaches that focus on surface features such as temporal connections between tweets and reactions, and the number of followers, tweets, account age, etc. to model cause-and-effect relationships without the use of additional content features. Other approaches in this cluster include work by Zaman et al. [100], who use a Bayesian approach to model retweets and show that it is possible to an extent to model the effects of a tweet from the number of followers and time of tweet. Lymperopoulos [52] also models tweet engagement from temporal and follower features based on a novel model inspired by an RC-Circuit.

While the above studies focus on analysing linguistic features in combination with metadata from tweets, Joseph et al. [37] offer an analysis of visual content on social media. The authors investigate the specific sub-topic of using images in tweets and the effects that this can create. They find that historical features (likes, statuses, historical likes) and transactional variables (creation time of post, age of profile, tweet length) have higher predictive power than any specific image properties. These features seem to indicate a certain set of acquired skills by experienced tweeters, who may learn to post in such a way as to maximise their effects over time. While we cannot verify this from the features in Joseph et al.'s study, it seems to be supported by work by Tavazoee et al. [80], who studied the evolution of tweets in the 2016 US presidential election.

2.3 Summary and research gaps

Purely data-driven and qualitative approaches both have distinct advantages. While the former offer automated analyses that can easily scale to large amounts of data with the potential for live or real-time analysis, e.g. monitoring, or identification of recent events, the level of insight is often shallow. For example, patterns or trends may be discovered without that their drivers and context are fully understood. At the same time, qualitative approaches, including traditional linguistic or social network analyses, can generate deep understanding and create new knowledge, but are very expensive to carry out, and do not transfer well across research domains, datasets, and do not lend themselves to real-time analysis. An important research gap therefore lies in developing hybrid approaches that combine the benefits of both paradigms of analysis. Also, many existing studies rely on a single modality, e.g. focusing exclusively on text, on images, or on social network features, opening up a further research gap on approaches that attempt to model multiple modalities congruently [9, 11, 62, 76, 83, 90]. Finally, much of the existing literature on user engagement with social media content relies on getting to know its users, e.g. modelling user networks, recognising interests, sometimes users' socio-economic features, and identifying influencers [27, 35, 72], etc. Other studies have attempted to model engagement based on surface features alone, such as the use of hashtags, emojis, etc. A research gap exists in methodologies that create a deep understanding of the social dynamics of a domain, in a way that is scalable and transferable to new domains and topics with relative computational ease.

3 DATA COLLECTION AND PREPARATION

3.1 Data collection and keywords

We used the Twitter API to collect a set of 1,003,059 tweets over a time span of six months, between November 2020 and May 2021. Keywords for the Twitter search were drawn from two sources:

- We obtained the names of all animals that were listed as vulnerable or endangered on the IUCN Red List[1] as a download in November 2020. This list was filtered to include only palearctic mammals, to reduce the number of search terms from about 19,000 down to 5,561. As the list uses the scientific names of animals, e.g. "ursus maritimus" instead of polar bear, we used a script to automatically convert scientific names to common names using Wikipedia. The intuition was that the latter would be much more commonly found on Twitter than the former Latin names. We found in our data collection that some animals were never tweeted about, leaving us with a remaining set of 4,305 that had at least one tweet over the time frame of our search (see list of all hashtags used on Github[2]) . The resulting distribution of keywords used as Twitter search terms is shown in Figure 1. Our

[1] https://www.iucnredlist.org/
[2] URL anonymised as per author instructions.

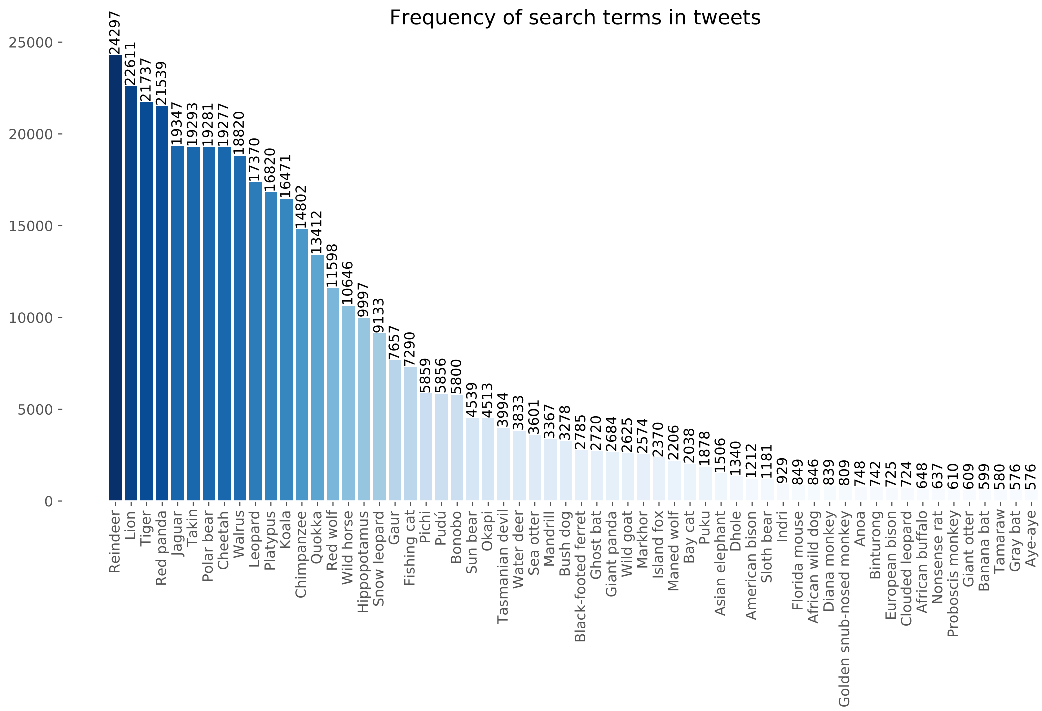

Fig. 1. Frequency distribution of animal names used as keywords for data collection.

initial search led to 3,571,350 tweets. After removing retweets and duplicates, this part of our search led to a dataset of 477,228 tweets.

- As a second step, we collected tweets from a set of other conservation-related search terms. This was based on our initial analysis showing that animal-keyword tweets are receiving predominantly positive attention (in line with earlier research [33, 92]), and we were aiming to have a more balanced dataset between positive and negative topics. We therefore chose the 100 most frequent conservation-related hashtags from our initial data collection as a set of further search terms. The distribution of these hashtags is shown in Figure 2. Adding tweets with specific hashtags, we obtained another 5,081,008 tweets, of which 525,831 remained after duplicates were removed.

Combining the two steps above, we are left with a dataset of 1,003,059 tweets for analysis.[3] While it is impossible to fully rule out collecting bot-generated tweets, we used Indiana University's Botometer[4] tool (which identifies user accounts that are likely bots) [98] in both data collection steps above to minimise the chance of collecting non-human-generated content. This led to a collection of tweets from 640,541 unique users, with an average of 1.78 tweets per user (± 8.27), a minimum of 1 and a maximum of 1,994. Table 1 shows statistics of users and tweet behaviour.

Table 2 shows the features available for each tweet, including metadata provided by the API, as well as additional labels described below.

[3]In line with the Twitter API's Terms & Conditions we are not able to share the collected dataset of tweets with the community. However, we make a list of tweet IDs available to support replicability of our results, see [anonymised Github URL].
[4]https://botometer.osome.iu.edu/

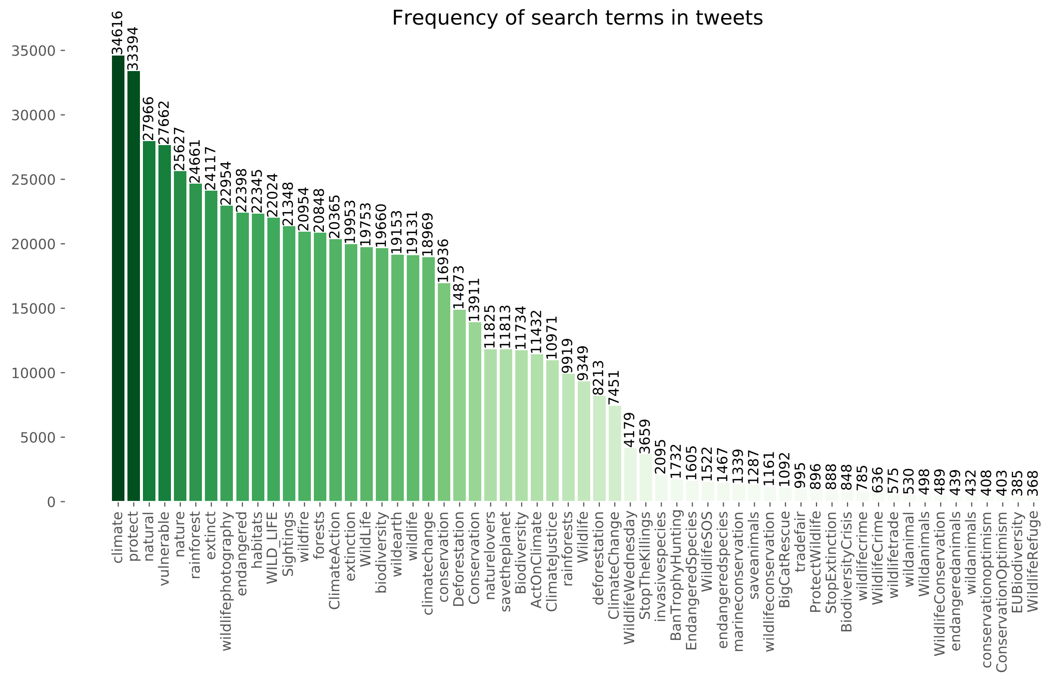

Fig. 2. Frequency distribution of conservation-related hashtags used for data collection.

Feature	Min	Max	Mean	Std	25%	50%	75%
User following	0.0	109,496,900	22,866.06	519,275.4	100	442	1,768
User followers	0.0	1,619,266	1,683.122	7,811.361	155	147	1,276
Tweets per user	0.0	6,329,018	32,859.21	120,846.5	1,419	6,648	24,673

Table 1. Basic statistics on user tweets, followers and following.

3.2 Data labelling and annotation

Besides metadata provided by the Twitter API, we added a set of additional annotation features on the image content, sentiment and valence and user reactions to tweets, described below.

Image processing. From the 1,003,059 tweets we collected, 186,461 had image content attached to them, and 160,196 were downloadable for analysis (other files were deleted, empty or corrupted). As an initial step, we used OpenCV to automatically detect human faces in the images and blur them with Gaussian noise to protect the individuals' privacy during analysis. Next, our aim was to annotate the data further in terms of whether or not an image included an animal. To this end, we manually sorted 150 images of animals and 150 without animals into separate folders (binary classification) - a small dataset seemed sufficient given the relative ease of this task. We trained a Convolutional Neural Network (CNN) (see specifications in Section 4.2) over 10 epochs using pre-trained ImageNet embeddings with another 10 epochs of fine-tuning, and a train-test split of 80%-20% with 5-fold cross-validation. This yielded an overall accuracy of 93% on the held-out test data. To compare with a standard benchmark, we also experimented with a YOLO image detector [69]. YOLO achieved a positive classification rate of 91% for images containing animals and a misclassification

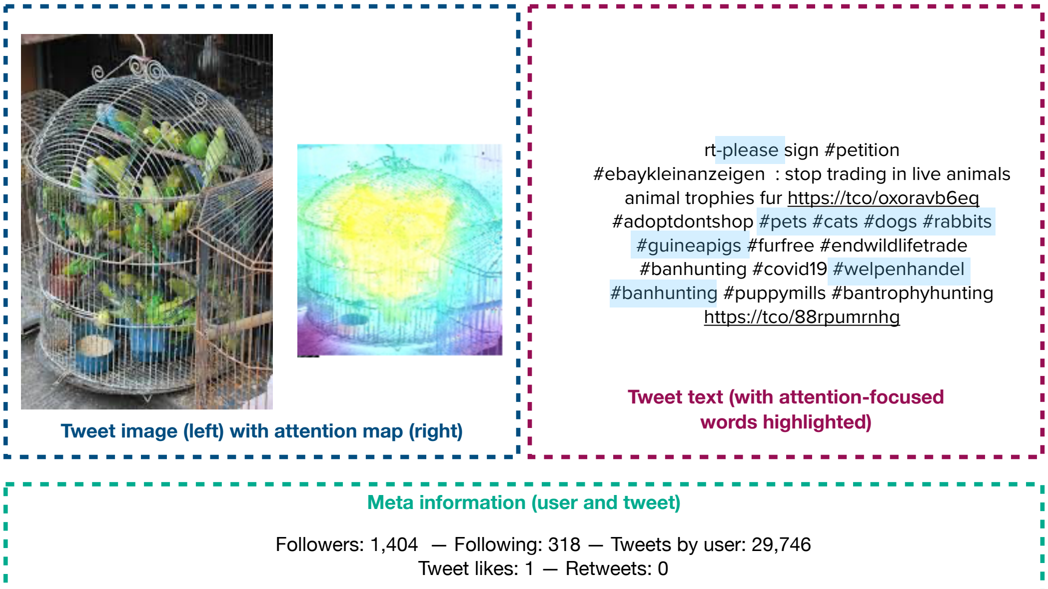

Fig. 3. Example tweet including image information, text and metadata. Attention maps were generated from the 12-head transformer network described in Section 4.1 for text, and the CNN in Section 4.2 for the image attention map.

Metadata from tweets	tweet id, text, user name (anonymised), user description, user location, user following, user followers, tweets by user, date when user account was created, date of tweet, number of retweets, number of likes, hashtags, links to any media (images or video), search term used to identify tweet.
Image label	Image contains an animal: binary
Sentiment label	Binary (discrete) sentiment estimated with DistilBERT
Valence	DistilBERT sentiment valence (float -1 to 1), continuous strength of sentiment.
Reaction label	High or low reaction to a tweet based user engagement with a tweet, i.e. number of "retweets" + "likes".

Table 2. Tweet attributes collected or labelled for analysis.

rate of 16%, yielding an overall accuracy of 84% in comparison to our bespoke model. It is noteworthy that YOLO only contains a subset of 11 animals of the ones we considered.

Using our bespoke trained classifier, we annotated the remainder of the dataset with the relevant binary label set. A qualitative analysis of a sample of classified images was conducted. Images that were mis-classified were mostly those that contained animals in the background, or in a secondary illustration, drawing, or other non-clear cut representation. Animals in the foreground or in close-up were generally recognised.

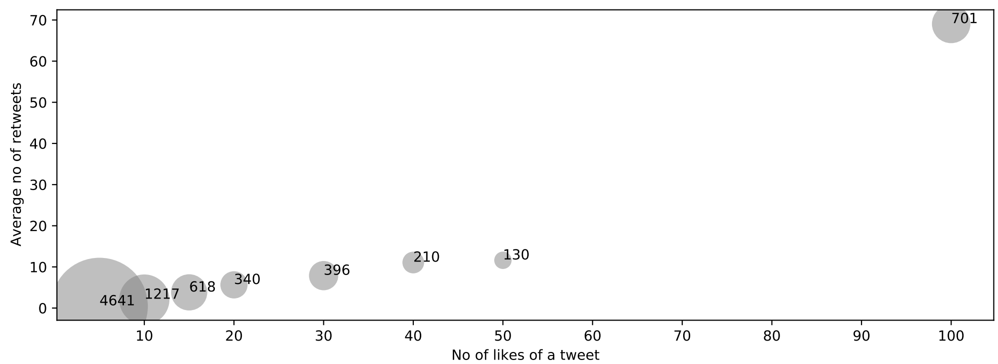

Fig. 4. Plot showing binned numbers of *likes* against average number of *retweets* in the group. We define 35 as a threshold between *low* and *high* engagement with tweets.

Sentiment and valence analysis. We added sentiment labels to each tweet based on its text, using HuggingFace's pre-trained DistilBERT module. DistilBERT[5] [75] is a smaller, more efficient variant of BERT [18] that assigns a valence score to a text segment between -1 (negative) and +1 (positive), which can be discretised into a binary sentiment label.

On manual inspection of the resulting multimodal and sentiment-labelled dataset, we observed that a majority of animal photos receive positive sentiment. This includes close-up images of animals, frequently in the wild, but also in zoos or images of pets, or in gardens (e.g. birds). Some memes that are made to look like realistic animals (or use images of some) are also in this positive category, as are realistic drawings of animals or costumes. Negative images can in some cases be very similar (e.g. close-ups of animals), but contain more images of cages, some dead or injured animals, images of traps, e.g. snarls, plastic, deforestation, and trophy hunters. Images are also of objects, e.g. weapons, bones, tusks, skulls. A sizeable number of negative images of animals are with humans (e.g. a hunter posing) or animals on stretchers, though some of them have a positive dimension (e.g. animals being helped). Images of dishes and body parts are also generally negative.

Engagement labels. Finally, we categorise tweets according to how much user engagement, they receive, measured in terms of their number of *likes* and *retweets*. The average number of *likes* in our dataset is 27.43 with a standard deviation of 168.64, a median of 3, a maximum of 6,630 and minimum value of 0. The average number of *retweets* is lower at 7.79 with a standard deviation of 49.68, a median of 1 and maximum value of 2,708. We add the number of *retweets* + *likes* and consider a *high reaction* score > 35, and *low* otherwise. This is based on an analysis illustrated in Figure 4. The figure shows the binned number of *likes* that a tweet receives mapped against the average number of *retweets* received by the bin. For example, we can see that 130 tweets receive between 40 and 50 *likes*, and the average number of *retweets* of this group is 10. Based on this analysis, we postulated 35 as a threshold to distinguish a tweet that receives *high* engagement from other that receive *low* engagement. In the remainder of the article, we are mostly interested in identifying the factors that lead to a tweet receiving a high reaction score. We acknowledge that our measure of engagement is does not incorporate valuable information on specific users, their tweet behaviour and wider social networks, which have been shown to be relevant prediction features in previous work [4, 14, 19, 38, 39, 46, 87].

[5]https://huggingface.co/docs/transformers/model_doc/distilbert

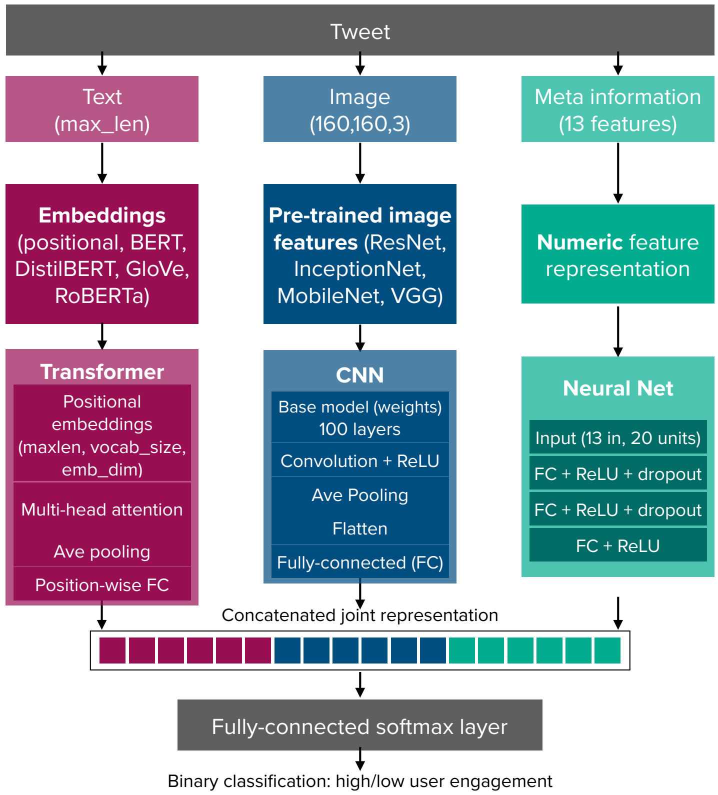

Fig. 5. Learning architecture of our neural network for multimodal user engagement classification. Learning models for text analysis (transformer network on the left), image analysis (CNN in the centre), and metadata (feedforward neural network on the right), are compiled, and then their final layer representations are concatenated.

4 METHODOLOGY

Our methodology is based on a deep learning model that combines information from three disparate sources drawn from the tweets: textual data analysis (Section 4.1), image data analysis (Section 4.2) and metadata (Section 4.3). Given that all three sources of data seem to contain valuable information that may determine user engagement with specific tweets, we want to explore to what extent treating all three data streams jointly can lead to better performance than the individual models alone.

4.1 Text analysis

We pre-process our text data by removing special characters ([\/:*'?¿";!< >-,]) and converting all text to lowercase. We use a sequence length of 35 words which we found covers the majority of tweets in our dataset as tweets are limited to 280 characters.

Our text analysis module is based on a transformer network [89] architecture with positional embeddings that represent the input to our learning model as $\mathbf{X}_{\text{TEXT}} \in \mathbb{R}^{n \times d}$. If we assume that $\mathbf{X}_{\text{TEXT}}$ is based on a d-dimensional embedding representation for n tokens (words) in a sequence, we can compute positional embeddings for $\mathbf{X}_{\text{TEXT}}$ as

$\mathbf{X_{TEXT}} + \mathbf{P}$, where $\mathbf{P}$ is a positional embedding matrix $\mathbf{P} \in \mathbb{R}^{n \times d}$. The elements of the i^{th} row and the $(2j)^{th}$ or the $(2j+1)^{th}$ column is then given as:

$$p_{i,2j} = \sin\left(\frac{i}{10000^{2j/d}}\right), p_{i,2j+1} = \cos\left(\frac{i}{10000^{2j/d}}\right). \tag{1}$$

The resulting embedding matrix $\mathbf{P}$ represents positions in a sequence as rows and different positional encoding dimensions as columns. Therefore, $\mathbf{X_{TEXT}} + \mathbf{P}$ can be represented as a matrix of n rows (one per token) of d columns (dimensions). This representation serves as an input to our transformer model for text analysis and is illustrated in Figure 5 (pink stream on the left). The transformer is implemented as a stack of layers, including a multi-head self-attention layer wit global average pooling, followed by a position-wise feedforward neural network [22].

The intuition behind self-attention is that each input token can pay attention to any other token during processing, which is computationally efficient due to parallelisation, and allows a wider linguistic context to be taken into account for prediction making [48, 64]. The inputs $\mathbf{X_{TEXT}} + \mathbf{P}$ described above are mapped to matrices $\mathbf{q}$ (query), $\mathbf{k}$ (keys) and $\mathbf{v}$ (values) with learnable weight matrices $\mathbf{W}_i^{(q)} \in \mathbb{R}^{p_q \times d_q}$, $\mathbf{W}_i^{(k)} \in \mathbb{R}^{p_k \times d_k}$ and $\mathbf{W}_i^{(v)} \in \mathbb{R}^{p_v \times d_v}$. This helps to find how the inputs interact together, and to determine the attention between input tokens (self-attention):

$$\mathbf{h}_i = softmax(\frac{\mathbf{QK}^\top}{\sqrt{d_k}})\mathbf{V}, \tag{2}$$

where $\mathbf{Q} = \mathbf{qW}_i^{(q)}$, $\mathbf{K} = \mathbf{kW}_i^{(k)}$, and $\mathbf{V} = \mathbf{vW}_i^{(v)}$. Multi-head attention applies multiple self-attention computations in parallel and later concatenates them, so that each "head" $\mathbf{h}_i$ can pay attention to different tokens during prediction making, thus gathering more relevant information towards the overall prediction. The position-wise feedforward neural network that follows the self-attention layer is a multi-layer perceptron with two hidden layers that work from the sequence of positions in our input text.

A transformer network as described in this section represents our baseline implementation for text analysis. Our experiments in Section 5 compare this model against large pre-trained language models, BERT [89] using bidirectional self-attention, DistilBERT [75] using knowledge distillation, and RoBERTa [49] using a robustly optimised BERT, to assess the benefits of pre-trained vs domain-specific embeddings.

4.2 Image analysis

We apply basic data augmentation to all images in our dataset, including horizontally flipping the images and applying a 20% degree random rotation to enhance the diversity of individual images. Data augmentation is applied to images in the training set only, and led to an additional 19,809 image-label pairs. This augmented dataset serves as input $\mathbf{X_{IMG}}$ to our learning model, where each image $x_{img,i} \in \mathbf{X_{IMG}}$ is a 160x160 3-dimensional RGB pixel matrix. For image analysis, we use a Convolutional Neural Network (CNN) [43] with pre-trained ImageNet weights and fine-tuning. The specific pre-trained model is varied across experiments (see Section 5), though MobileNet V2 [74] is light-weight and gives us consistent results. We inherit the layers of the pre-trained *base model* and feed pre-processed augmented data $\mathbf{X_{IMG}}$ into this base model. We then stack a two-dimensional convolutional layer with ReLU activation on top of the base model, and an average global pooling layer, followed by a fully-connected prediction layer. We use an Adam optimiser and a sparse categorical cross-entropy loss function with a learning rate of 0.0001. The resulting model is then trained with an initial set of epochs during which layers on the pre-trained base model remain frozen. After the initial training phase we unfreeze these layers and fine-tune the model. The image analysis phase is shown in the blue (middle) stream in Figure 5 with the relevant layers used in the image analysis CNN.

4.3 Metadata

We also include metadata from tweets into our multimodal analysis, consisting of features: user location (discrete), presence of animals in image (binary), DistilBERT sentiment (binary), DistilBERT valence (float), presence of emojis (discrete), search term (discrete), number of user followers (int), number of accounts that user follows (int), and tweets made by user (int). We refer to the input from the metadata stream as $\mathbf{X_{META}}$. Given the tabular form of the data, this information is modelled by a standard feedforward neural network that computes an increasingly abstract hidden representation of $\mathbf{X_{META}}$ captured in the hidden state $\mathbf{g}$, which is computed through updates to a non-linear activation function $f(\mathbf{x}_{META,t}, \mathbf{g}_{t-1})$ at timestep t. We apply two fully-connected layers, with 10 and 5 hidden units each, with a dropout rate of 0.2 on each layer and a ReLU activation function. The architecture is illustrated in Figure 5 alongside the text and image models. We use Adam optimisation and a categorical cross-entropy loss function to minimise the loss expected and generated outputs. This classifier is used to predict user engagement from metadata only in Table 3 below, and is also used in the joint setting with other modalities.

4.4 Multimodal

As illustrated in Figure 5, the final multimodal learning model concatenates the last layers of each of the image $\mathbf{OUT_{IMG}}$, text $\mathbf{OUT_{TEXT}}$ and metadata $\mathbf{OUT_{META}}$ models into a single layer representation. We stack a prediction layer on top and train the model over ten epochs with a batch size of 32, Adam optimisation and sparse categorical cross-entropy loss. Different dual combinations of modalities reported in Section 6 omit one of the layers, but otherwise follow the same principle.

5 EXPERIMENTS AND RESULTS

We describe our experimental setup in this section followed by a quantitative evaluation of our learning models. We then present an evaluation that explores qualitative aspects of tweets and different modalities in more detail.

5.1 Experimental setup

We compare five different experimental setups: two models using text only, one using images only, one using only metadata, and one multimodal setup that combines different modalities. All learning models use the same train-test split which is 80% to 20% for the joint subset of data, i.e. all tweets that have both text and images associated to them. Our test dataset for all experiments contains 1,650 tweets. This leaves 1,001,409 tweets containing text for training, and a much smaller training dataset of 6,603 tweets that contains text and images for joint training. The metadata (only) results are computed from the full 1,001,409 data points.

Text-only baselines.

- **Transformer networks** as described in Section 4.1 with 2, 8 or 12 heads. Embedding representations in these models are learnt from the domain data without pre-trained language models. Our implementation uses one hidden layer (512 units) with layer normalisation (ϵ=1e-6) [3] and dropout (0.1). We use a batch size of 32 for these experiments and Adam optimisation. We use an embedding dimension of 128, maximum sequence length of 35 words and a vocabulary size of 16,33,395.

- **BiGRU with Glove** uses GloVe [65] embeddings with its pre-trained Twitter word embeddings *glove.twitter.27B.100d*. The learning model is a Gated Recurrent Unit (GRU) [17] with two bidirectional layers (512, 256 units) and 0.2 of dropout. We use an embedding vector length of 100, and a maximum sequence length of 35. Other parameters are shared with the transformer networks.
- **BERT** [18] is applied for comparison as a large pre-trained language model. We use *bert-base-uncased* embeddings with a sequence length of 35. Our model stacks an additional fully-connected layer (512 units and ReLU activation) on top of the BERT embeddings, as well as a softmax prediction layer. Both layers use a 0.01 L2 kernel regulariser.
- We also compare with a **DistilBERT** [75] model using *distilbert-base-uncased* embeddings, which is considered a lighter-weight model based on BERT with 40% less parameters. Training parameters are shared with the BERT model above. Both use Adam optimisation.
- Finally we use a **RoBERTa** [49] model with *roberta-base* embeddings, which is also based on BERT but uses dynamic masking and hence more training data. Model and training parameters were the same as for the other BERT and DistilBERT above.

We also compare our text-only classifiers in two conditions: using the full textual data set available (*full data*), and using only the subset of data samples that also contain an image (*joint subset*). The latter is necessarily the only data that is available for joint learning.

Image-only baselines. To establish prediction performance for an image analysis only task, we use a CNN learning model as introduced in Section 4.2. Our experiments focus on varying the pre-trained image weights, while keeping the remainder of the model setup and parameters constant. We compare **MobileNetV2** [74], **VGG19** [77], **InceptionNetV3** [79] and **ResNet50V2** [31]. As a baseline comparison to image classifiers with pre-trained weights, we also report results with a **Standard CNN** (3 convolutional layers with 16 filters and 3 kernels, max pooling and ReLU activation) that learns domain weights from scratch. All models are trained for 10 epochs initially and then fine-tuned for another 10 epochs before generating predictions.

Metadata-only baselines. We also experiment with a set of baselines to predict user engagement with tweets from metadata of those tweets alone. Our **Neural Net** baseline (multi-layer perceptron) uses two fully-connected layers with a dropout rate of 0.2 and a sigmoid activation function. We use Adam optimisation and a binary cross-entropy loss function. The neural network model is compared with a set of non-neural standard machine learning baselines, which can generally be expected to show decent performance given the tabular nature of the data: a **Naive Bayes** classifier, a **decision tree**, **random forest**, and an **XGBoost** classifier.

6 RESULTS

Table 3 shows quantitative results for all models according to test accuracy, balanced accuracy, precision, recall, F1 score and the model parameters. The test set is unbalanced with a majority baseline of 86%, which makes test accuracy a slightly less interesting metric to consider in the table below.

6.1 Learning results

Inspecting initially the *balanced accuracy* score of our results, we can see that the **images only** category fails to learn a balanced prediction model for both output classes *high* and *low* user engagement. Overall, our **MobileNet** model appears to be the most successful of the pre-trained image weight models, and is also the most efficient to train given

Model	Test Accuracy	Balanced Accuracy	Precision	Recall	F1	Model parameters
Images only						
Standard CNN	**0.87**	0.50	0.76	**0.87**	0.81	3,301,028
CNN+MobileNetV2+FT	**0.87**	0.50	**0.83**	**0.87**	**0.81**	2,260,546
CNN+VGG19+FT	0.86	0.50	0.78	0.86	**0.81**	10,586,178
CNN+InceptionV3+FT	0.84	0.50	0.78	0.83	0.80	21,806,882
CNN+ResNet50V2+FT	0.84	**0.51**	0.78	0.84	**0.81**	23,568,898
Text only (joint subset)						
Transformer (2 heads)	0.85	0.58	0.79	0.77	0.78	1,455,774
Transformer (8 heads)	0.86	0.57	0.78	0.77	0.77	1,480,926
Transformer (12 heads)	0.86	0.55	0.80	0.84	**0.81**	1,497,694
BiGRU with GloVe	**0.87**	0.50	0.73	0.86	0.79	165,073,334
BERT	0.86	0.51	**0.83**	0.86	**0.81**	109,876,994
DistilBERT	**0.87**	**0.60**	0.82	0.79	0.80	66,757,634
RoBERTa	**0.87**	0.50	0.76	**0.87**	**0.81**	125,040,386
Text only (full data)						
Transformer (2 heads)	0.95	0.84	0.95	0.95	0.95	52,281,182
Transformer (8 heads)	0.95	0.87	**0.96**	0.96	**0.96**	52,306,334
Transformer (12 heads)	0.95	**0.91**	0.95	0.95	0.95	52,323,102
BiGRU with GloVe	0.94	0.50	0.73	0.86	0.79	165,073,334
BERT	**0.97**	0.50	0.93	**0.97**	0.95	109,876,994
DistilBERT	**0.97**	0.50	0.93	**0.97**	0.95	66,757,634
RoBERTa	**0.97**	0.50	0.93	**0.97**	0.95	125,040,386
Metadata only						
Naive Bayes	0.84	0.54	0.79	0.84	0.80	-
Decision Tree	0.83	**0.66**	0.83	0.83	0.83	-
Random Forest	0.86	0.50	0.73	0.86	0.79	-
XGBoost	**0.87**	0.57	**0.88**	**0.87**	0.83	-
Neural Net (MLP)	0.85	0.50	0.73	0.86	**0.86**	-
Multimodal models						
Joint encoder text-images	0.86	0.59	**0.81**	**0.86**	**0.82**	18,554,338
Joint encoder text-meta	0.86	**0.62**	0.80	0.77	0.78	1,552,004
Joint encoder images-meta	0.86	0.50	0.73	**0.86**	0.79	17,017,636
Joint encoder all	**0.87**	0.57	**0.81**	0.84	**0.82**	18,561,988

Table 3. Results for predicting reactions for tweets in terms of images, text or metadata only and models that use a joint representation.

the smallest number of parameters. While *recall* numbers are better, the results overall seem to confirm earlier research that has shown that predicting user engagement from visual features is difficult [12, 15, 20, 37, 51], lending motivation to the exploration of multimodal features.

For **text only** models, we see a similar pattern for the models trained from the smaller *joint subset* of tweets. The **DistilBERT, RoBERTa** and **BiGRU with GloVe** models achieve the joint best *test accuracy*, followed by the transformer networks, but overall performance is low. This improves when the larger *full data* set of text samples is taken into account, which train from our full set of 1,001,409 text-based tweets. We can see that the pre-trained language models **BERT**, **DistilBERT** and **RoBERTa** all achieve high *test accuracy* and *precision, recall* and *F1* scores. However, their performance drops sharply when looking at *balanced accuracy*. This drop is not observed for the transformers

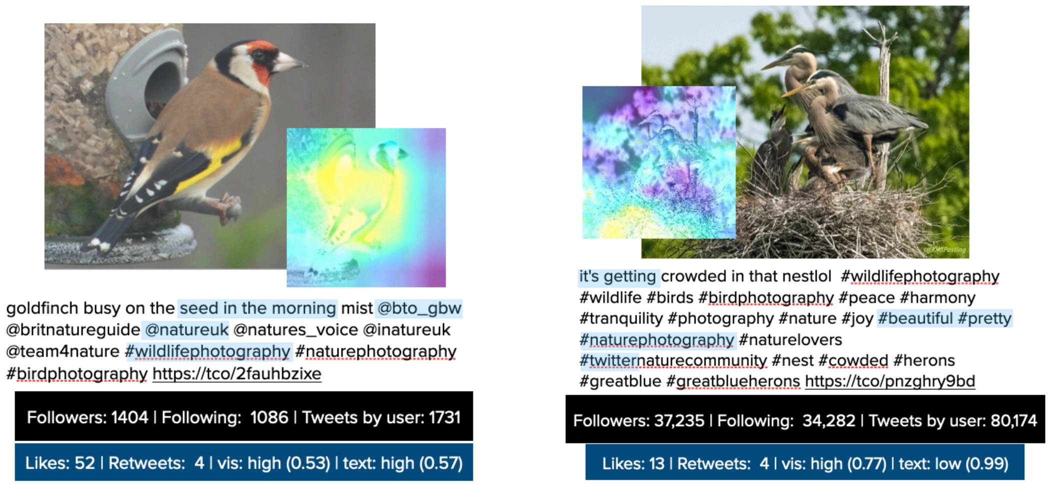

Fig. 6. Example tweets from the domain of bird photography. Confidence level of classifier is shown as a probability.

that are trained from in-domain data without pre-trained embeddings. A larger amount of training data clearly has a substantial effect on the performance of user engagement prediction models, and linguistic features carry substantial predictive value towards this task.

Looking finally at the **metadata** models, we see learning success for the **XGBoost** model and the **Decision Tree** classifier with a maximum *balanced accuracy* of 0.66. These results are noticeably better than images alone, but are far from the 90+% text-based models trained from *full data*.

Based on these results, we combined a **12-head transformer network**, **CNN+MobileNetV2+FT** and the **Neural Network** for metadata into a single model to compute the multimodal joint results at the bottom of Table 3. The transformer was chosen because it has overall the best performance of the text models. While the BERT variants score slightly higher in some metrics, this is not consistent, and the vanilla transformer has only 2.4% of the parameters of the smallest BERT model (DistilBERT). For predictions from metadata, the neural network did not show high performance across metrics, but allows the extraction of learnt weights into a joint multimodal model, and was therefore chosen based on architectural considerations. We can see that the best *recall* and *F1* results in this section are achieved by a model that combines all three modalities (**joint encoder all**). However, the best *balanced accuracy* results are from a model that combines text and metadata only (**joint encoder text-meta**). None of the multimodal model combinations perform nearly as highly as the text only model trained from *full data*. This is presumably due to a small amount of image data in comparison with the amount of text and metadata.

As a further layer of analysis, Table 4 shows confidence levels for the best performing models per category, alongside statistical significance (based on a Wilcoxon Signed Rank test) and effect size r. We can see earlier results confirmed with the text-only models showing the highest confidence in their predictions and the lowest Brier score (which measures the accuracy of probabilistic predictions) overall, followed by the set of joint models.

Figures 6 and 7 illustrate example tweets alongside their text, image and metadata, and can provide a more intuitive illustration of the prediction models. Specifically, the two tweets in Figure 6 are both examples of bird photography. The

Fig. 7. Example tweets with negative topics, deforestation and illegal wildlife trade. Confidence level of classifier is shown as a probability.

tweet on the left was made by an account with a smaller amount of followers and accounts they follow, with a smaller amount of tweets. In general this does not seem to be a high profile Twitter account. The specific tweet however, which features a close-up of a bird, in a relatively informal position, received 52 likes and 4 retweets, which according to our categorisation in Section 3.2 is classed as *high* user engagement. In contrast, the tweet on the right was made by a much more high profile account and looks professionally taken (as also indicated by the photographer's signature in the lower right-hand corner). Yet the actual tweet received *low* engagement. We can see from the visual attention maps that the bird in the left-hand image is recognised, but not the group of birds on the right-hand side. Nonetheless our image classifier predicts *high* engagement for both, though with different confidence levels (0.53 for the single bird and 0.96 for the group). The text classifier on the other hand predicts correctly based on a set of keywords and hashtags.

In contrast to the positive-natured bird tweets, tweets in Figure 7 deal with more negative topics: deforestation (left) and wildlife trade (right). Again we can see that the left-hand post was made with a moderately active user account. The image classifier recognises the correct region of interest in the image (patch of missing trees), however does not make a correct engagement prediction. The text classifier is correct again based on a small set of keywords. The tweet on the right-hand side was made from a more active account. The text-based prediction and image-based predictions are both incorrect this time, with the image attention map roughly focusing on the correct region of the image, but with low confidence.

Overall while attention maps from text and image classifiers can help us understand the predictions that are generated for individual tweet instances, they are not helpful for the discovery of broader and more general patterns of user engagement with conservation-related social media content.

6.2 Qualitative analysis

This section presents an annotation scheme for a subset of our data in an attempt to uncover more of the general features and patterns that are at work in our dataset. We know from earlier research (see Section 2) that people prefer

Model	Probabilistic Confidence of Predictions (overall)	(correct)	(incorrect)	Brier score	Effect size r
Joint modality models					
Joint encoder text-images**	0.936 ± 0.086	0.941 ± 0.078	0.903 ± 0.119	0.121	0.81
Joint encoder text-meta**	0.851 ± 0.089	0.857 ± 0.083	0.812 ± 0.114	0.116	0.93
Joint encoder images-meta**	0.821 ± 0.029	0.822 ± 0.030	0.814 ± 0.024	0.123	0.99
Joint encoder all**	0.871 ± 0.129	0.883 ± 0.120	**0.799** ± 0.158	0.121	NA
Text only					
Transformer (12 heads), full data**	**0.970** ± 0.081	**0.981** ± 0.060	**0.792** ± 0.158	**0.039**	0.58
Distilbert (full data)**	0.963 ± 0.0705	0.966 ± 0.064	0.865 ± 0.142	**0.031**	0.50
Distilbert (joint data subset)**	0.927 ± 0.0441	0.929 ± 0.043	0.908 ± 0.045	0.115	0.94
Images only					
CNN+MobileNetV2+FT**	0.896 ± 0.072	0.896 ± 0.066	0.893 ± 0.095	0.155	0.99
Metadata only					
Decision Tree**	**1.0** ± 0.0	**1.0** ± 0.0	1.0 ± 0.0	0.147	0.49

Table 4. Analysis of the confidence (and standard deviation ±) of different models in their output predictions. Statistical significance at $p < 0.0001$ is shown as ** and is computed against the fully joint model *Joint encoder all*.

Attribute	**Description**	Max	Min	Std	Mean	Median
Emojis	Emojis in tweet	10	0	1.6	0.55	0
Urls	URLs in tweet	3	0	0.54	1.34	1
Hashtags	Hashtags in tweet	27	0	4.13	3.2	2
Usernames	Usernames in tweet	12	0	3.33	1.05	0
Place names	Geographical place names in tweet	4	0	0.87	0.55	0
Names	Names in tweet	5	0	0.53	0/19	0

Table 5. Context annotation categories and attributes. Numbers are given for an annotated sub-set of 1,650 tweets. Each category refers to the total raw count occurring in a tweet and is represented by an integer.

Image content	Categories of image content (all binary)	true	false
Weapons, guns	Visible in image.	34	1616
Animal and people	Both visible in image.	86	1564
People in image	Visible in image.	270	1380
Graphics	Graph, plot or other illustration.	94	1556
Text imposed	Text imposed over visual content.	552	1098
Public figure	Recognisable person in the image.	25	1625
Beautiful nature	Image of (often pristine) nature.	128	1522
Destruction of nature	Deforestation, drilling, fires, etc.	148	1502
Animal in distress	Visible in image.	88	1562
Animal in image	Visible in image.	851	799
Animal in focus	Visible in image.	618	1032

Table 6. Image content annotation categories and attributes. Numbers are given for an annotated sub-set of 1,650 tweets.

to engage with media content on "charismatic" animals over other less prominent species [30, 58, 63, 71], and that negative emotions, high valence [54] the use of emojis [16], hashtags, usernames [61], and URLs [16] lead to positive engagement of users with social media content in some contexts. This is also the case for information-based tweets,

Content / Speech Act	**Discrete category tweet of speech act**	Frequency
Call for Action (CFA)	A call for action, e.g. to react, to participate, to change behaviour, etc.	136
CFA + Event	A call for action to attend an event	2
CFA + Advert	A call for action to support a commercial organisation or make a purchase.	38
Community	A tweet that is relevant to a specific community e.g. bird watchers and is mostly relevant to them.	713
Community + Advert	An advert that appeals to a sub-community	39
Community + CFA	A call for action directed at a specific sub-community, e.g. to act react, sign a petition, etc.	64
Information	Factual tweet that conveys information	232
Information + Advert	An advert that is based on a factual situation	35
Information + CFA	A call for action motivated by factual information	121
Information + Community	Information relevant only to a sub-community	113
Information + Community + CFA	A sub-community is called to act based on information mostly relevant only to them.	8
Advert	A tweet advertising a product, company, etc.	146
Fundraising	A fundraising tweet, e.g. a call to donate	2

Table 7. Tweet semantic / speech act content annotation categories and attributes. Numbers are given for an annotated sub-set of 1,650 tweets.

according to some studies [16]. In contrast to some of those findings, other studies have found a negative influence of hashtags [16] or information-based tweets on user engagement [54], as well as marking tweets [28, 99] or those that call for action, e.g. prompting users to respond, react, donate, share, or similar [16]. To the best of our knowledge, there are no studies so far that have found a demonstrable influence of image-based content on user engagement, either positive or negative [16, 37].

6.2.1 Content annotations. To capture known features of user engagement, we manually annotated a small portion of our dataset, i.e. 20% (or 1,650 tweets) of the multimodal *joint data* corresponding to our test set above. Table 5 shows annotation categories for textual content in individual tweets alongside basic statistics. These features are mostly objective and can be extracted from tweets semi-automatically, e.g. via the @ or # symbols. Table 6 focuses on image-based content, where categories are binary. The table shows class distributions for each attribute, e.g. if animals were visible in the image, or people, destruction of nature, or other relevant categories. The specific categories were chosen based on empirical inspection and prominent visual categories in our dataset. While image-based features are also objective in their nature, they are less easy to extract reliably via automatic processing, and were therefore hand-annotated by the authors. Finally, Table 7 lists speech acts that individual tweets can represent. These categories are based on earlier research presented above, and combinations of base categories. The table shows the frequency of different speech acts in our dataset. Speech acts were based on the presumed intent behind a tweet and its linguistic features. For example, *information-based* tweets were not fact-checked, so information may or may not be fact-based and genuine, but tweets were annotated as such based on their linguistic presentation.

6.2.2 Content analysis. Figure 8 shows a correlation matrix for metadata associated with tweets. These features were extracted automatically and were introduced in Section 3.2 above. We can make the following observations:

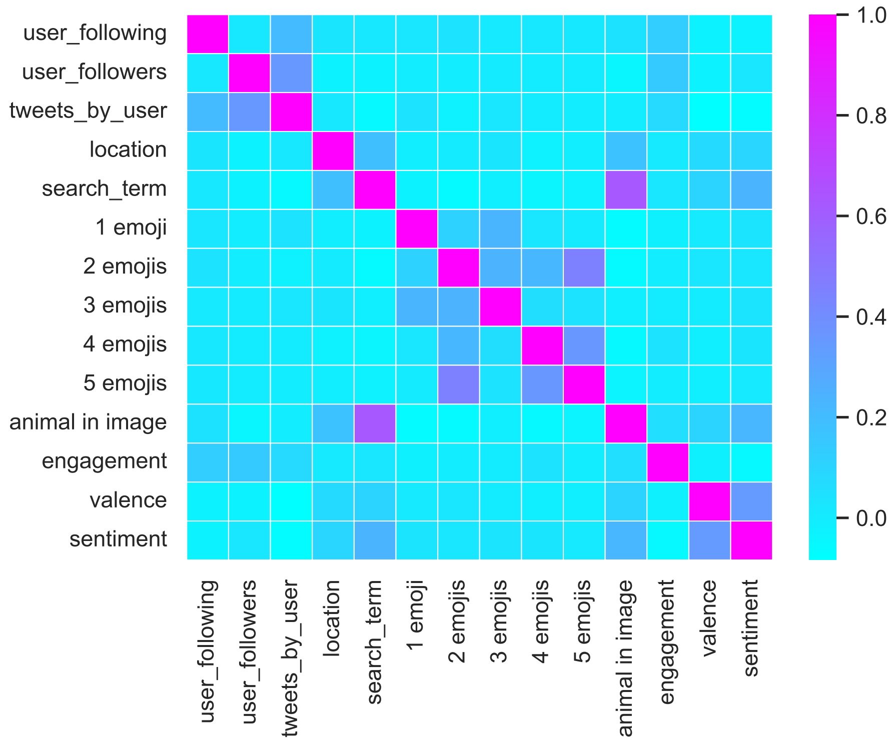

Fig. 8. Correlation matrix for metadata, including tweet-related information, sentiment, reaction and valence. Features were introduced in Section 3.2.

(1) Emojis tend not to occur in isolation, but in clusters. In other words, if a tweet uses emojis, it is likely to use more than one.
(2) Images of animals tend to occur with specific search terms (e.g. animal names), and tend to carry positive sentiment, albeit slightly.
(3) Users with a more active social media network, i.e. who follow other users and have a fair amount of followers themselves, tend to receive more engagement with their posts/tweets.

Figure 9 shows correlations for the content-based features annotated and shown in Tables 5-7 above. We observe the following:

(1) Humans and animals that are visible in images (*Animal and people*) often correlate with weapons and guns also being present (strong correlation at 0.55), and often represent a distress situation for the animal (*Animal in distress*).
(2) The number of hashtags in a tweet correlates moderately with images of animals (*Animal in focus*) at 0.20, and animals in focus correlate weakly with the amount of engagement that the tweet receives (0.15).
(3) Given results from previous research, it is worth noting that we did not observe any correlation effects from the speech act used in a tweet.

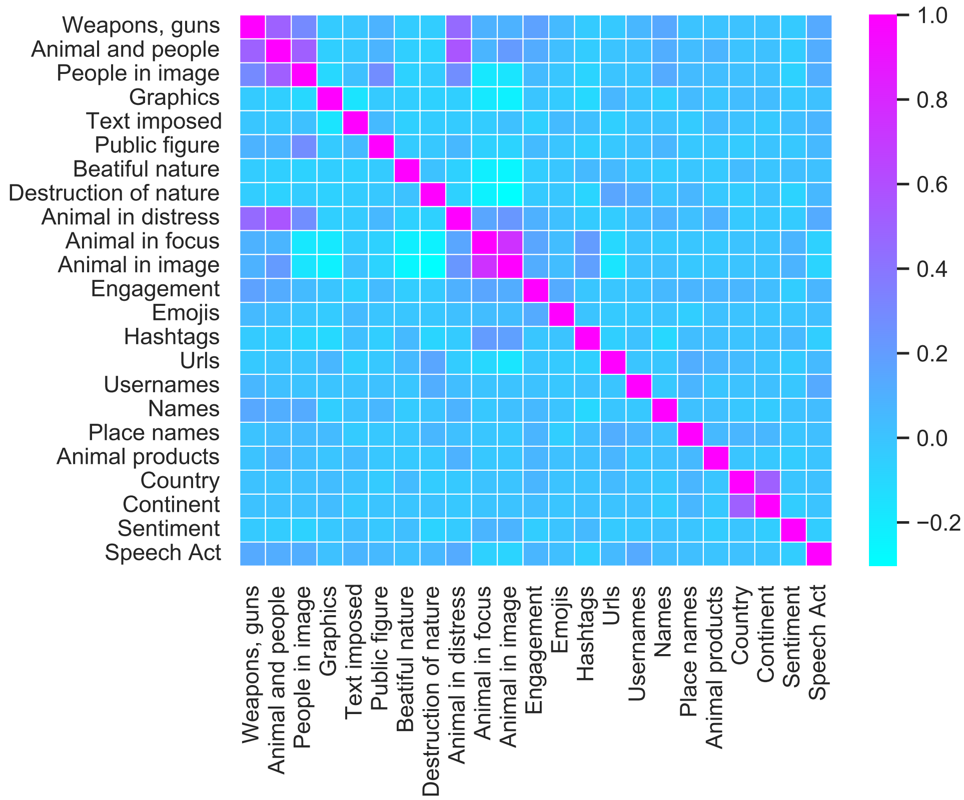

Fig. 9. Correlation matrix for textual and visual content features, annotated on 1,650 sample tweets from the test set, paired with sentiment and engagement labels.

While the correlation analyses give us some insight into the drivers of user engagement, it is still difficult to formulate concrete patterns or heuristics to predict user engagement for individual tweets. As a next analysis step, we therefore applied clustering to our annotated data to see what further insights could be gained. Figure 10 shows the results of a KMeans++ cluster analysis of the set of all merged content and metadata features. We experimented with the number of clusters K empirically and using the Elbow method, and found six clusters to yield a good representation.

As a next step to clustering, we wanted to find out which features are prominent in each of the clusters to create an understanding of the groupings and interactions of content and metadata features in our data. We approached this by measuring the distance between individual features across two clusters at a time. The distance function is based on the mean value of a category inside a cluster, e.g. the mean value of images that show “destruction of nature" (binary), or the users following a tweeting account (int), the numbers of hashtags in a tweet (int), or in fact the engagement (binary) that tweets in a cluster receive. As a second step, we aimed to identify those features that were highly indicative of a data point's membership of a particular cluster. Specifically, for each feature a, we compute the mean value of the feature in a cluster c and subtract it from the value of another cluster:

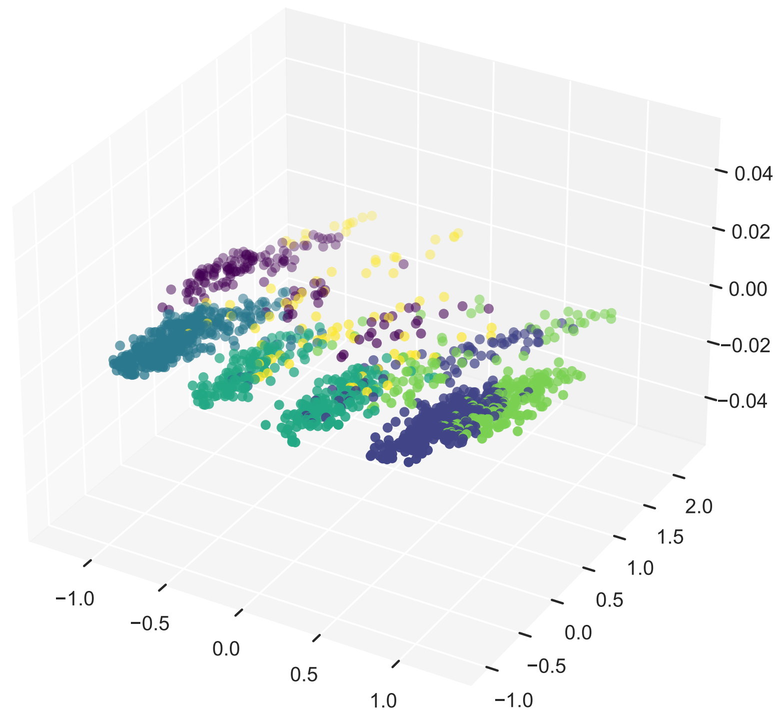

Fig. 10. KMeans++ cluster analysis using the joint set of metadata and content features, where $K = 6$. We see five clear clusters appear, with data points in the last cluster being more dispersed.

$$distance = m_a^{c1} - m_a^{c2}, \tag{3}$$

where a is a feature under consideration from our context and metadata features, and $c1$ and $c2$ are two clusters under comparison. The purpose of this comparison is to determine those features that have high relevance for specific clusters. For example, if all members of a cluster are entirely positive in sentiment, their mean value will be 1.0. In contrast, a fully negative cluster will have a mean of 0.0. We compared distance functions based on mean, standard deviation and variance, and found that mean distance is a good measure to discern individual clusters. Based on this analysis, we were able to identify the following clusters:

- **Cluster 1:** This cluster mostly features animal photography, with *animals in image* (0.88) and *in focus* (0.75), high *engagement* (1.0) and high *valence* in some cases (0.30). *Sentiment* is reasonably high in this cluster (0.82). Some images can have *text imposed* on them (0.65), which e.g. refers to additional information, or represents the signature of the photographer. This cluster includes users with an active network of followers (0.65) and accounts that they follow (0.66). We call this cluster *Animal photography by influencers*.

- **Cluster 2:** This cluster also focuses on animal photography, it always contains *animals* (1.0), usually *in focus* (0.80), with high *sentiment* (0.85), but low *engagement* (0.0). The cluster has a moderate amount of followers (0.40) and accounts they follow (0.44). We call this cluster *Animal photography by non-influencers*.
- **Cluster 3:** This cluster shows *animals and people* together (1.0), where animals are often *in distress* (0.65) and *in focus* (0.53), at times *weapons and guns* are also visible (0.34). There is comparatively high *sentiment* (0.71) in this cluster, though *valence* is lower (0.18) and a reasonable amount of *engagement* (0.30), with a good amount of followers (0.51) and users followed (0.46). We call this the *Animal cruelty and illegal wildlife trade cluster*.
- **Cluster 4:** This cluster features images of nature, either pristine *"beautiful" nature* (0.42) or the *destruction of nature* (0.16) with high *sentiment* (0.93) and *valence* (0.40), but no *engagement* (0.0). Accounts followed are high in this cluster (0.64) and so are followers (0.76). We call this our *nature cluster*.
- **Cluster 5:** This cluster features mixed content with some *people in images* (0.25), some *graphics* (0.14) and *text* (0.20), some *destruction of nature* (0.26) with overall lower *engagement* (0.13), *sentiment* (0.32) and *valence* (0.08). The cluster has a generous amount of followers (0.53) and accounts they are following (0.50). We call this cluster *Negative mixed cluster*.
- **Cluster 6:** This cluster focuses on information, e.g. *infographics* or facts, represented as *text imposed* on an image (1.0). *People* can be present (0.25), but no *animals* in this cluster (0.0), with high *sentiment* (0.86) and low *engagement* (0.05). This cluster has a moderate level of followers (0.40) and accounts they follow (0.38). We this call this our *information cluster*.

These clusters reveal a number of concrete patterns and feature interactions that can help determine the user engagement that a tweet will receive. The level of activity, measured by followers and accounts followed, is a clear indicator of engagement. Other topical and semantic properties also play a role, e.g. tweets about wildlife trade, poaching, animal photography or the preservation of nature receive a fair amount of engagement, but mostly so in combination with active user accounts. The clusters were able give insight into a number of different sub-communities that are active on Twitter and that share, and engage with, specific thematic content.

It seems clear from our two different types of analyses that both unsupervised clustering and classification complement each other in trying to construct an understanding of this domain. Classification is able to address large amounts of input data as features are mostly automatically obtained from raw Twitter data, and could therefore inform a "real time" use case, if need be. At the same time, a clustering analysis was able to generate deeper insights into the interactions between user engagement and the semantic features of tweets, but was based on time-consuming manual annotation, so comes at a much higher cost for a small amount of data.

7 DISCUSSION OF LIMITATIONS

The use of social media data for the analysis of social, behavioural, geographical and other phenomena can have distinct advantages, including for digital conservation, as illustrated by earlier research presented in Section 2, as well as our own research in this article. Social media can overcome problems of sample size, temporal and spatial constraints and allows easy and fast data access. Lopez et al. [50] also argue for the value of social media as an observation tool of actual behaviour – as social media posts are largely unsolicited, they can give insights into perspectives and preferences that may not have been discovered in a more structured form of data collection that may prompt certain types of responses. In that sense social media is also an ideal tool for explorative research. Consequently, we were able in this article to consider data from a much larger, and potentially more diverse and geographically dispersed, set of social media users, than we may have reached with more traditional forms of crowdsourcing or questionnaire-based data collection. At the same time, Lopez et al. warn of the risk of mis-interpretation of social media posts, especially as the context of a tweet is not always provided when scraping via an API: posts are presented as individual artefacts when in reality they occurred as part of a thread or conversation in the context of which they should be interpreted.

One important issue when using social media as a tool for behaviour analysis is a significant population bias [73]. It is known that younger users are overrepresented on social media with different demographics favouring different platforms (see e.g. Mislove et al. [60]; or Mellon and Prosser [55]). This has been confirmed in the context of digital conservation studies, e.g. in a comparison of platform preferences of different types of users discussed above [30]. In our research, we can see a clear bias in the geographical reach of our study. A majority of our tweets originate from English-speaking countries, many in the Northern hemisphere. While only a minority of tweets are linked to a geolocation (about 1%), 54.51% of them originated from the US or the UK (28% US, 26% UK), followed by 7.74% from India, 4.5% from Canada, followed by decreasing percentages from Australia (3.24%), South Africa (2.52%), New Zealand (2.16%), Germany (1.98%), Kenya (1.80%), Belgium (1.53%), Ireland (1.35%), France (1.08%), Pakistan (1.08%), Finland (0.99%), Nigeria (0.90%), and other individual countries (14.68%). This can lead to bias in the view points represented in individual studies and often results in an Anglo-centric focus [96].

Issues around representativeness and bias are exacerbated by a lack of replicability of social media research and systematic evaluation, see Arts et al. [2] for a conservation-perspective. Such issues arise largely because T&Cs of social media platforms nearly always forbid the sharing or further distribution of any collected data to protect user privacy and commercial interests. This issue has become more pertinent by the change of Twitter's / X's data access policy. While the platform was one of the last to offer an open API for research purposes until mid-2023, this data source is now largely lost to the community, which will affect research on digital conservation in future. Comparable platforms such as Mastodon still allow data access, yet have a much smaller user population and less established online communities. The general inaccessibility of social media for research prevents common benchmarks to be established as is typical in other fields of machine learning and artificial intelligence, such as computer vision and natural language processing, amongst others. These communities share a set of core datasets for benchmarking, run competitions and increasingly release code, allowing for comparability of approaches and replicability of research. We attempt to support the replicability of our research by providing the list of tweet IDs that were used in our experiments.

8 CONCLUSION AND FUTURE WORK

We presented a multimodal neural learning architecture and experimented with different combinations of text, image and metadata of tweets to predict user engagement with Twitter content. Engagement was based on a function of the number *likes* and *retweets* that a tweet receives. We find that a transformer network trained with a large amount of text from the target domain performs best, outperforming models that consider other modalities, such as images, or tabular metadata. Importantly, we observed a negative effect of large pre-trained language models when working with a domain-specific unbalanced dataset. While models such as BERT and variants, e.g. DistilBERT and RoBERTa, outperform other models on test accuracy and recall, this is not confirmed for metrics that take the unbalanced nature of the data into account, such as balanced accuracy. At the same time we find that in the absence of a generous text dataset, improved prediction performance can be achieved through a combination of multiple modalities, e.g. taking information from tweet properties and user account into consideration. In accordance with previous research, we find a negligible effect of image features on user engagement. A Chi-Squared test reveals a highly significant effect of the presence of an image on user engagement: $X^2(1, N = 1,138,093), p = .00001$, however we were not able to identify clear visual patterns or themes that guide this engagement. Rather it seems that images serve to get a user's initial attention, while other modalities have a higher impact on whether the user ultimately engages with a tweet or not.

On a more qualitative level, we were interested in the specifically defining topics in the social media discourse on conservation. We find clear recurrent threads on *wildlife photography and animal sightings*, the *protection of vulnerable species and illegal wildlife trade* and *trophy hunting*, *rainforests and deforestation*, as well as *climate change and climate action* in a more general sense. All of these topics prompt interest and user engagement in principle, and this was found to be amplified when content originates from active user accounts, i.e. Twitter users with a high number of followers and accounts they follow. We also found evidence of sub-communities around these topics that share and engage with similar thematic content. Sentiment is a clear predictor of engagement, and can be positive or negative, depending on the sub-community and content. With regards to our research question on identifiable linguistic, visual or metadata features that are predictive of user engagement, we discovered six distinct topical clusters that help structure the discourse in our dataset. Based on these clusters, we show that engagement emerges from a combination of topic, user activity and sentiment, rather than a single set of distinctly identifiable features.

Reviewing our original research questions, we make the following findings in this article.

(1) What are the defining and recurring topics in social media discourse around the conservation of species?
  - Domain topics emerge from a combination of text and image content, and include *animals*, *weapons and guns*, *animals and people*, *beautiful nature*, *descruction of nature*, *information and infographics* (Table 6 and Sec. 6.2).
  - Speech-act related topics such as *call for action*, *fundraising*, *community*, *events* and *adverts*, were found to have no observable effect in our data (Table 7 and Sec. 6.2).

(2) Who are the sub-communities that participate in such discourse and what are their identifiable characteristics?
  - Social media discourse on conservation topics is dominated by sub-communities of users that already care about certain topics, such as *wildlife photography*, *protection of nature or vulnerable species*, or *climate change information*, and are likely to engage with new content on the same topic as they have before (Sec. 6.2).

(3) What are identifiable (linguistic, visual or meta) characteristics of tweets that function as triggers of online user engagement?
  - Text is the most effective modality to predict user engagement from tweets, in comparison with visual features or metadata, but this only holds when sufficient amounts of training data is available (Sec. 6.1).
  - With limited amounts of training data, combinations of modalities, such as text with metadata, can boost performance over single-modality models (Sec. 6.1). Metadata related to a user's social network, such as their followers or number of tweets, seems particularly informative (Sec. 6.2).

(4) To what extent can recent advances in deep learning for text and image analysis form an effective basis for user engagement prediction?
  - User engagement emerges from a combination of user activity, online conversation topic and sentiment, where both positive and negative tweets receive engagement for different topics (Sec. 6.2). We were able to model this relationship effectively using state-of-the-art transformer networks (Sec. 6.1).
  - In contrast, large pre-trained language models can have a negative effect on prediction performance when dealing with a substantially unbalanced dataset (Sec. 6.1).

Future work can drill down further into the specific linguistic features that drive user engagement with individual tweets. While our method of analysis did not lend itself to discovering broad trends in linguistic features, e.g. word categories, rhetorical structure or stylistic devices, beyond individual attention maps, some patterns are clearly present given the success of transformer-based engagement prediction. A deeper-level discourse analysis can likely reveal some of them. Similarly, a social network analysis can be applied to explore how a better understanding of individual users, e.g. their location, interests, social network membership and topical interests, can further support engagement modelling for conservation-related content. There is also the possibility to apply data enhancing methods, e.g. paraphrase generation, to address the unbalanced nature of the dataset. Similarly, recent dual learning approaches, such as CLIP [66], VisualBERT [47], MulT [84], Zorro [68] or ALIGN [36], amongst others, may be used to augment our larger text-only dataset, and create a richer set of examples for multimodal analysis [5]. In the same vein, it can be explored if dual learning methods can create richer representations of the modality allocation of social media content than our simpler feature concatenation approach. This may lead to additional insights on the contribution of multimodality in engaging users.

ACKNOWLEDGMENTS

We acknowledge the VIPER high-performance computing facility of the University of Hull and its support team for carrying out the experiments reported in this article.

REFERENCES

[1] Ashley A. Anderson and Heidi E. Huntington. 2017. Social Media, Science, and Attack Discourse: How Twitter Discussions of Climate Change Use Sarcasm and Incivility. *Science Communication* 39, 5 (2017), 598–620. https://doi.org/10.1177/1075547017735113 arXiv:https://doi.org/10.1177/1075547017735113

[2] Koen Arts, Rene van der Wal, and William M. Adams. 2015. Digital technology and the conservation of nature. *Ambio* 44 (2015), 661–673.

[3] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. 2016. Layer Normalization. http://arxiv.org/abs/1607.06450 cite arxiv:1607.06450.

[4] María Teresa Ballestar, Marta Martín-Llaguno, and Jorge Sainz. 2022. An artificial intelligence analysis of climate-change influencers' marketing on Twitter. *Psychology & Marketing* 39, 12 (2022), 2273–2283. https://doi.org/10.1002/mar.21735 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/mar.21735

[5] Tadas Baltrusaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2019. Multimodal Machine Learning: A Survey and Taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.* 41, 2 (2019), 423–443.

[6] Sheila Barry. 2014. Using social media to discover public values, interests, and perceptions about cattle grazing on park lands. *Environmental Management* 53(2) (2014), 454–464.

[7] Susanne Becken, Bela Stantic, Jinyan Chen, Ali Reza Alaeia, and Rod M.Connolly. 2017. Monitoring the environment and human sentiment on the Great Barrier Reef: assessing the potential of collective sensing. *Journal of Environmental Management* 203 (2017), 87–97.

[8] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *J. Mach. Learn. Res.* 3 (mar 2003), 993–1022.

[9] Damian Borth, Tao Chen, Rongrong Ji, and Shih-Fu Chang. 2013. SentiBank: large-scale ontology and classifiers for detecting sentiment and emotions in visual content. In *Proc. of the International Multimedia Conference (MM)*.

[10] Lydia Bryan-Smith, Jake Godsall, Franky George, Kelly Egode, Nina Dethlefs, and Dan Parsons. 2023. Real-time social media sentiment analysis for rapid impact assessment of floods. *Computers & Geosciences* 178 (2023). https://https://www.sciencedirect.com/science/article/pii/S0098300423001097

[11] Lydia Bryan-Smith, Jake Godsall, Franky George, Kelly Egode, Nina Dethlefs, and Dan Parsons. 2023. Real-time social media sentiment analysis for rapid impact assessment of floods. *Computers & Geosciences* 178 (2023), 105405. https://doi.org/10.1016/j.cageo.2023.105405

[12] Spencer Cappallo, Thomas Mensink, and Cees G.M. Snoek. 2015. Latent Factors of Visual Popularity Prediction. In *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval* (Shanghai, China) *(ICMR '15)*. Association for Computing Machinery, New York, NY, USA, 195–202. https://doi.org/10.1145/2671188.2749405

[13] Ana Sofia Cardoso, Sofiya Bryukhova, Francesco Renna, Luís Reino, Chi Xu, Zixiang Xiao, Ricardo Correia, Enrico Di Minin, Joana Ribeiro, and Ana Sofia Vaz. 2023. Detecting wildlife trafficking in images from online platforms: A test case using deep learning with pangolin images. *Biological Conservation* 279 (2023), 109905. https://doi.org/10.1016/j.biocon.2023.109905

[14] Chang-Feng Chen, Wen Shi, Jing Yang, and Hao-Huan Fu. 2021. Social bots' role in climate change discussion on Twitter: Measuring standpoints, topics, and interaction strategies. *Advances in Climate Change Research* 12, 6 (2021), 913–923. https://doi.org/10.1016/j.accre.2021.09.011

[15] Junhong Chen, Dayong Liang, Zhanmo Zhu, Xiaojing Zhou, Zihan Ye, and Xiuyun Mo. 2019. Social Media Popularity Prediction Based on Visual-Textual Features with XGBoost *(MM '19)*. Association for Computing Machinery, New York, NY, USA, 2692–2696. https://doi.org/10.1145/3343031.3356072

[16] Angie Chung, hongjoo Woo, and Kang-Bok Lee. 2020. Understanding the information diffusion of tweets of a non-profit organization that targets female audiences: an examination of Women Who Code's tweets. *Journal of Communication Management* (09 2020). https://doi.org/10.1108/JCOM-05-2020-0036

[17] Junyoung Chung, Çaglar Gülçehre, Kyunghyun Cho, and Yoshua Bengio. 2014. *Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling*. Technical Report Arxiv report 1412.3555. Université de Montréal. Presented at the Deep Learning workshop at NIPS2014.

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

[19] Trevor Diehl, Brigitte Huber, Homero Gil de Zúñiga, and James Liu. 2019. Social Media and Beliefs about Climate Change: A Cross-National Analysis of News Use, Political Ideology, and Trust in Science. *International Journal of Public Opinion Research* 33, 2 (11 2019), 197–213. https://doi.org/10.1093/ijpor/edz040 arXiv:https://academic.oup.com/ijpor/article-pdf/33/2/197/39770180/edz040.pdf

[20] Keyan Ding, Ronggang Wang, and Shiqi Wang. 2019. Social Media Popularity Prediction: A Multiple Feature Fusion Approach with Deep Neural Networks. In *Proceedings of the 27th ACM International Conference on Multimedia* (Nice, France) *(MM '19)*. Association for Computing Machinery, New York, NY, USA, 2682–2686. https://doi.org/10.1145/3343031.3356062

[21] Ehab Eid and Ramzi Handal. 2017. Illegal hunting in Jordan using social media to assess impacts on wildlife. *Oryx: Cambridge University Press* 52 (4) (2017), 730 – 735.

[22] A. Zhang et al. 2021. *Dive into Deep Learning*. https://d2l.ai/

[23] M Falkenberg, A Galeazzi, M Torricelli, N Di Marco, F Larosa, M Sas, A Mekacher, W Pearce, F Zollo, W Quattrociocchi, and A Baronchelli. 2022. Growing polarization around climate change on social media. *Nature Climate Change* 12 (2022), 1114–1121.

[24] Christoph Fink, Anna Hausmann, and Enrico Di Minin. 2020. Online sentiment towards iconic species. *Biological Conservation* 241 (2020), 108289. https://doi.org/10.1016/j.biocon.2019.108289

[25] Christoph Fink, Tuuli Toivonen, Ricardo A. Correia, and Enrico Di Minin. 2021. Mapping the online songbird trade in Indonesia. *Applied Geography* 134 (2021), 102505. https://doi.org/10.1016/j.apgeog.2021.102505

[26] Gianfranco Gliozzo, Nathalie Pettorelli, and Mordechai (Muki) Haklay. 2016. Using crowdsourced imagery to detect cultural ecosystem services: A case study in South Wales. *Ecology & Society* 21 (3) (2016).

[27] Jan-Frederik Gräve. 2019. What KPIs Are Key? Evaluating Performance Metrics for Social Media Influencers. *Social Media + Society* 5, 3 (2019), 2056305119865475. https://doi.org/10.1177/2056305119865475 arXiv:https://doi.org/10.1177/2056305119865475

[28] Jeanine P.D. Guidry, Richard D. Waters, and Gregory D. Saxton. 2014. Moving social marketing beyond personal change to social change: Strategically using Twitter to mobilize supporters into vocal advocates. *Journal of Social Marketing* 4 (3) (2014), 240–260.

[29] Zellig Harris. 1954. Distributional structure. *Word* 10, 2-3 (1954), 146–162.

[30] Anna Hausmann, Tuuli Toivonen, Rob Slotow, Henrikki Tenkanen, Atte Moilanen, Vuokko Heikinheimo, and Enrico Di Minin. 2017. Social media data can be used to understand tourists' preferences for nature-based experiences in protected areas. *Conservation Letters* 11 (1) (2017), 249–258.

[31] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition* (Las Vegas, NV, USA) *(CVPR '16)*. IEEE, 770–778. https://doi.org/10.1109/CVPR.2016.90

[32] Jill E. Hopke and Luis E. Hestres. 2018. Visualizing the Paris Climate Talks on Twitter: Media and Climate Stakeholder Visual Social Media During COP21. *Social Media + Society* 4, 3 (2018), 2056305118782687. https://doi.org/10.1177/2056305118782687 arXiv:https://doi.org/10.1177/2056305118782687

[33] Kyle S. Van Houtan, Tyler Gagne, and Clinton N. Jenkins andLucas Joppa. 2020. Sentiment Analysis of Conservation Studies Captures Successes of Species Reintroductions. *Patterns* 1 (1) (2020).

[34] Y. Hu. 2018. Geo-Text Data and Data-Driven Geospatial Semantics. *Geography Compass* 12 (11) (2018).

[35] Liselot Hudders, Steffi De Jans, and Marijke De Veirman. 2021. The commercialization of social media stars: a literature review and conceptual framework on the strategic use of social media influencers. *International Journal of Advertising* 40, 3 (2021), 327–375. https://doi.org/10.1080/02650487.2020.1836925 arXiv:https://doi.org/10.1080/02650487.2020.1836925

[36] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision. In *International Conference on Machine Learning*. https://api.semanticscholar.org/CorpusID:231879586

[37] Nimish Joseph, Amir Sultan, Arpan Kumar Kar, and P. Vigneswara Ilavarasan. 2018. Machine Learning Approach to Analyze and Predict the Popularity of Tweets with Images. In *Proc. of the 17th Conference on e-Business, e-Services and e-Society (I3E)*. Kuwait City, Kuwait.

[38] Dan M. Kahan, Hank Jenkins-Smith, and Donald Braman. 2011. Cultural cognition of scientific consensus. *Journal of Risk Research* 14, 2 (2011), 147–174. https://doi.org/10.1080/13669877.2010.511246 arXiv:https://doi.org/10.1080/13669877.2010.511246

[39] Samantha Kay, Rory Mulcahy, and Joy Parkinson. 2020. When less is more: the impact of macro and micro social media influencers' disclosure. *Journal of Marketing Management* 36, 3-4 (2020), 248–278. https://doi.org/10.1080/0267257X.2020.1718740 arXiv:https://doi.org/10.1080/0267257X.2020.1718740

[40] Ritwik Kulkarni and Enrico Di Minin. 2021. Automated retrieval of information on threatened species from online sources using machine learning. *Methods in Ecology and Evolution* 12, 7 (2021), 1226–1239. https://doi.org/10.1111/2041-210X.13608

[41] Ritwik Kulkarni and Enrico Di Minin. 2023. Towards automatic detection of wildlife trade using machine vision models. *Biological Conservation* 279 (2023), 109924. https://doi.org/10.1016/j.biocon.2023.109924

[42] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. 2010. What is Twitter, a social network or a news media?. In *Proc. of the International Multimedia Conference (MM)*.

[43] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. 1998. Gradient-based learning applied to document recognition. *Proc. IEEE* 86, 11 (1998), 2278–2324. https://doi.org/10.1109/5.726791

[44] Noam Levin, Salit Kark, and David Crandall. 2015. Where have all the people gone? Enhancing global conservation using night lights and social media. *Ecological Applications* 25 (8) (2015), 2153–2167.

[45] Noam Levin, Alex Mark Lechner, and Greg Brown. 2017. An evaluation of crowdsourced information for assessing the visitation and perceived importance of protected areas. *Applied Geography* 79 (2017), 115–126.

[46] Stephan Lewandowsky, Klaus Oberauer, and Gilles E. Gignac. 2013. NASA Faked the Moon Landing—Therefore, (Climate) Science Is a Hoax: An Anatomy of the Motivated Rejection of Science. *Psychological Science* 24, 5 (2013), 622–633. https://doi.org/10.1177/0956797612457686 arXiv:https://doi.org/10.1177/0956797612457686 PMID: 23531484.

[47] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557* (2019).

[48] Zhouhan Lin, Minwei Feng, Cícero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. 2017. A Structured Self-Attentive Sentence Embedding. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

[49] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. http://arxiv.org/abs/1907.11692

[50] Bianca E. Lopez, Nicholas R. Magliocca, and Andrew T. Crooks. 2019. Challenges and Opportunities of Social Media Data for Socio-Environmental Systems Research. *MDPI Land* 8(7) (2019).

[51] Jinna Lv, Wu Liu, Meng Zhang, He Gong, Bin Wu, and Huadong Ma. 2017. Multi-Feature Fusion for Predicting Social Media Popularity *(MM '17)*. Association for Computing Machinery, New York, NY, USA, 1883–1888. https://doi.org/10.1145/3123266.3127897

[52] Ilias N. Lymperopoulos. 2021. RC-Tweet: Modeling and predicting the popularity of tweets through the dynamics of a capacitor. *Expert Systems with Applications* 163 (2021).

[53] David W. Macdonald, Kim S. Jacobsen, Dawn Burnham, Paul J. Johnson, and Andrew J. Loveridge. 2016. Cecil: A Moment or a Movement? Analysis of Media Coverage of the Death of a Lion, Panthera, Leo. *MDPI Animals* 6(5) (2016).

[54] Maryam Mahdikhani. 2021. Predicting the popularity of tweets by analyzing public opinion and emotions in different stages of Covid-19 pandemic. *International Journal of Information Management Data Insights* 2 (1) (2021).

[55] Jonathan Mellon and Christopher Prosser. 2017. Twitter and Facebook are not representative of the general population: Political attitudes and demographics of British social media users. *Research and Politics* (2017).

[56] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *CoRR* abs/1301.3781 (2013). http://arxiv.org/abs/1301.3781

[57] Enrico Di Minin, Christoph Fink andTuomo Hiippala, and Henrikki Tenkanen. 2018. A framework for investigating illegal wildlife trade on social media with machine learning. *Conservation Biology* 33 (1) (2018).

[58] E. Di Minin, I. Fraser, R. Slotow, and D. C. MacMillan. 2012. Understanding heterogeneous preference of tourists for big game species: implications for conservation and management. *Animal Conservation* 16 (3) (2012), 249–258.

[59] E. Di Minin, H. Tenkanen, and T. Toivonen. 2015. Prospects and challenges for social media data in conservation science. *Frontiers in Environmental Science* 3 (2015).

[60] Alan Mislove, Sune Lehmann, Yong-Yeol Ahn, Jukka-Pekka Onnela, and J. Rosenquist. 2021. Understanding the Demographics of Twitter Users. *Proceedings of the International AAAI Conference on Web and Social Media* 5, 1 (Aug. 2021), 554–557.

[61] Nasir Naveed, Thomas Gottron, Jerome Kunegis, and Arifah Che Alhadi. 2011. Bad News Travel Fast: A Content-based Analysis of Interestingness on Twitter. In *Proc. of the 3rd International Web Science Conference (WebSci)*. Koblenz, Germany.

[62] Daniela Onita, Liviu P. Dinu, and Adriana Birlutiu. 2019. From Image to Text in Sentiment Analysis via Regression and Deep Learning. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*. INCOMA Ltd., Varna, Bulgaria, 862–868.

[63] S.K. Papworth, T.P.L. Nghiem, D. Chimalakonda, M.R.C. Posa, L.S. Wijedasa, D. Bickford, and L.R. Carrasco. 2015. Quantifying the role of online news in linking conservation research to Facebook and Twitter. *Conservation Biology* 29 (3) (2015), 825–833.

[64] Ankur Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. 2016. A Decomposable Attention Model for Natural Language Inference. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Austin, Texas, 2249–2255. https://doi.org/10.18653/v1/D16-1244

[65] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global Vectors for Word Representation.. In *EMNLP*, Vol. 14. 1532–1543.

[66] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, 8748–8763.

[67] A. Radford, Jeffrey Wu, R. Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners.

[68] Adrià Recasens, Jason Lin, João Carreira, Andrew Jaegle, Luyu Wang, Jean-Baptiste Alayrac, Pauline Luc, Antoine Miech, Lucas Smaira, Ross Hemsley, and Andrew Zisserman. 2023. Zorro: the masked multimodal transformer. *CoRR* abs/2301.09595 (2023).

[69] Joseph Redmon and Ali Farhadi. 2017. YOLO9000: Better, Faster, Stronger.. In *CVPR*. IEEE Computer Society, 6517–6525. http://dblp.uni-trier.de/db/conf/cvpr/cvpr2017.html#RedmonF17

[70] Lucía Rivadeneira, Jian-Bo Yang, and Manuel Lopez-Ibanez. 2020. Predicting tweet impact using a novel evidential reasoning prediction method. *Expert Systems With Applications* 169 (2020).

[71] JM Roberge. 2014. Using data from online social networks in conservation science: which species engage people the most on Twitter? *Biodiversity and Conservation* 23 (2014), 715–726.

[72] Javier Rodríguez-Vidal, Julio Gonzalo, Laura Plaza, and Henry Anaya Sánchez. 2019. Automatic detection of influencers in social networks: Authority versus domain signals. *Journal of the Association for Information Science and Technology* 70, 7 (2019), 675–684. https://doi.org/10.1002/asi.24156

[73] Derek Ruths and Juergen Pfeffer. 2014. Social Media for Large Studies of Behaviour. *Science* 346 (6213) (2014), 1063–1064.

[74] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. 2018. MobileNetV2: Inverted Residuals and Linear Bottlenecks. (2018). https://doi.org/10.48550/ARXIV.1801.04381

[75] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108* (2019).

[76] Stefan Siersdorfer, Enrico Minack, Fan Deng, and Jonathon Hare. 2010. Analysing and Predicting Sentiment of Images on the Social Web. In *Proc. of the International Multimedia Conference (MM)*.

[77] Karen Simonyan and Andrew Zisserman. 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. https://doi.org/10.48550/ARXIV.1409.1556

[78] Karen Sparck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation* 28, 1 (1972), 11–21.

[79] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. 2015. Rethinking the Inception Architecture for Computer Vision. *CoRR* abs/1512.00567 (2015). http://dblp.uni-trier.de/db/journals/corr/corr1512.html#SzegedyVISW15

[80] Farideh Tavazoee, Claudio Conversano, and Francesco Mola. 2020. Recurrent random forest for the assessment of popularity in social media: 2016 US election as a case study. *Knowledge and Information Systems* 62 (2020), 1847–1879.

[81] Henrikki Tenkanen, Enrico Di Minin, Vuokko Heikinheimo, Anna Hausmann, Marna Herbst, Liisa Kajala, and Tuuli Toivonen. 2017. Instagram, Flickr and Twitter: Assessing the usability of social media data for visitor monitoring in protected areas. *Nature Scientific Reports* 7 (2017).

[82] T Toivonen, V Heikinheimo, C Fink, A Hausmann, T Hiippala, O Jaerv, H Tenkanen, and E Di Minin. 2019. Social media data for conservation science: A methodological overview. *Biological Conservation* 233 (2019), 298–315.

[83] Quoc-Tuan Truong and Hady W. Lauw. 2017. Visual sentiment analysis for review images with item-oriented and user-oriented CNN. In *Proc. of the International Multimedia Conference (MM)*.

[84] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal Transformer for Unaligned Multimodal Language Sequences. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, Anna Korhonen, David R. Traum, and Lluís Màrquez (Eds.). Association for Computational Linguistics, 6558–6569.

[85] Aman Tyagi, Matthew Babcock, Kathleen M. Carley, and Douglas C. Sicker. 2020. Polarizing Tweets on Climate Change. In *Social, Cultural, and Behavioral Modeling: 13th International Conference, SBP-BRiMS 2020, Washington, DC, USA, October 18–21, 2020, Proceedings* (Washington, DC, USA). Springer-Verlag, Berlin, Heidelberg, 107–117. https://doi.org/10.1007/978-3-030-61255-9_11

[86] Tuomas Väisänen, Vuokko Heikinheimo, Tuomo Hiippala, and Tuuli Toivonen. 2021. Exploring human–nature interactions in national parks with social media photographs and computer vision. *Conservation Biology* 35, 2 (2021), 424–436. https://doi.org/10.1111/cobi.13704

[87] Sander van der Linden, Anthony Leiserowitz, Seth Rosenthal, and Edward Maibach. 2017. Inoculating the Public against Misinformation about Climate Change. *Global Challenges* 1, 2 (2017), 1600008. https://doi.org/10.1002/gch2.201600008

[88] Boris T. van Zanten, Derek B. Van Berkel, Ross K. Meentemeyer, Jordan W. Smith, Koen F. Tieskens, and Peter H. Verburg. 2016. Continental-scale quantification of landscape values using social media data. *PNAS: Environmental Sciences* 113 (46) (2016), 12974–12979.

[89] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems 30*. Curran Associates, Inc.

[90] Yilin Wang, Suhang Wang, Jiliang Tang, Huan Liu, and Baoxin Li. 2015. Unsupervised Sentiment Analysis for Social Media Images *(IJCAI'15)*. AAAI Press, 2378–2379.

[91] Stanley Wasserman and Katherine Faust. 1994. *Social network analysis: Methods and applications*. Vol. 8. Cambridge university press. http://scholar.google.com/scholar.bib?q=info:gET6m8icitMJ:scholar.google.com/&output=citation&hl=en&as_sdt=0,5&as_vis=1&ct=citation&cd=0

[92] Louise Willemen, Andrew J. Cottam, Evangelia G. Drakou, and Neil D. Burgess. 2015. Using social media to measure the contribution of red list species to the nature-based tourism potential of African Protected areas. *PLOS ONE* (2015).

[93] HTP Williams, JR McMurray, T Kurz, and FH Lambert. 2015. Network analysis reveals open forums and echo chambers in social media discussions of climate change. *Global Environmental Change* 32 (2015), 126–138. https://doi.org/10.1016/j.gloenvcha.2015.03.006

[94] Ricarda Winkelmann, J.F. Donges, E. Keith Smith, Manjana Milkoreit, Christina Eder, Jobst Heitzig, Alexia Katsanidou, Marc Wiedermann, Nico Wunderling, and Timothy M. Lenton. 2022. Social tipping processes towards climate action: A conceptual framework. *Ecological Economics* 192

(2022).

[95] Spencer A. Wood, Anne D. Guerry, Jessica M. Silver, and Martin Lacayo. 2013. Using social media to quantify nature-based tourism and recreation. *Nature Scientific Reports* 3 (2013).

[96] Ifeoluwa Wuraola, Nina Dethlefs, and Daniel Marciniak. 2023. Linguistic Pattern Analysis in the Climate Change-Related Tweets from UK and Nigeria. In *Proceedings of the CLASP Conference on Learning with Small Data (LSD)* (Gothenburg, Sweden) *(CLASP)*.

[97] Qing Xu, Mingxiang Cai, and Tim K Mackey. 2020. The illegal wildlife digital market: an analysis of Chinese wildlife marketing and sale on Facebook. *Environmental Conservation* 47, 3 (2020), 206–212. https://doi.org/10.1017/S0376892920000235

[98] KC Yang, E. Ferrara, and F. Menczer. 2022. Botometer 101: social bot practicum for computational social scientists. *J Comput Soc Sc* 5 (2022), 1511–1528.

[99] Amir Zadeha and Ramesh Shardab. 2022. How Can Our Tweets Go Viral? Point-Process Modelling of Brand Content. *Information & Management* 59 (2) (2022).

[100] Tauhid Zaman, Emily B. Fox, and Eric T. Bradlow. 2014. A Bayesian Approach for Predicting the Popularity of Tweets. *The Annals of Applied Statistics* 8 (3) (2014), 1583–1611.