Human–Wildlife Interactions 15(1):95–110, Spring 2021 • digitalcommons.usu.edu/hwi

Machine learning as a tool for wildlife management and research: the case of wild pig-related content on Twitter

Lauren M. Jaebker, Department of Human Dimensions of Natural Resources, Warner College of Natural Resources, Colorado State University, 1480 Campus Delivery, Fort Collins, CO 80523, USA; and USDA, APHIS, Wildlife Services' National Wildlife Research Center, 4101 LaPorte Avenue, Fort Collins, CO 80525, USA *lauren.jaebker@colostate.edu*

Hailey E. McLean, Department of Human Dimensions of Natural Resources, Warner College of Natural Resources, Colorado State University, 1480 Campus Delivery, Fort Collins, CO 80523, USA; and USDA, APHIS, Wildlife Services' National Wildlife Research Center, 4101 LaPorte Avenue, Fort Collins, CO 80525, USA

Stephanie A. Shwiff, USDA, APHIS, Wildlife Services' National Wildlife Research Center, 4101 LaPorte Avenue, Fort Collins, CO 80525, USA

Keith M. Carlisle, Department of Human Dimensions of Natural Resources, Warner College of Natural Resources, Colorado State University, 1480 Campus Delivery, Fort Collins, CO 80523, USA; and USDA, APHIS, Wildlife Services' National Wildlife Research Center, 4101 LaPorte Avenue, Fort Collins, CO 80525, USA

Tara L. Teel, Department of Human Dimensions of Natural Resources, Warner College of Natural Resources, Colorado State University, 1480 Campus Delivery, Fort Collins, CO 80523, USA

Alan D. Bright, Department of Human Dimensions of Natural Resources, Warner College of Natural Resources, Colorado State University, 1480 Campus Delivery, Fort Collins, CO 80523, USA

Aaron M. Anderson, USDA, APHIS, Wildlife Services' National Wildlife Research Center, 4101 LaPorte Avenue, Fort Collins, CO 80525, USA

Abstract: Wild pigs (*Sus scrofa*) are a non-native, invasive species that cause considerable damage and transmit a variety of diseases to livestock, people, and wildlife. We explored Twitter, the most popular social media micro-blogging platform, to demonstrate how social media data can be leveraged to investigate social identity and sentiment toward wild pigs. In doing so, we employed a sophisticated machine learning approach to investigate: (1) the overall sentiment associated with the dataset, (2) online identities via user profile descriptions, and (3) the extent to which sentiment varied by online identity. Results indicated that the largest groups of online identity represented in our dataset were females and people whose occupation was in journalism and media communication. While the majority of our data indicated a negative sentiment toward wild pigs and other related search terms, users who identified with agriculture-related occupations had more favorable sentiment. Overall, this article is an important starting point for further investigation of the use of social media data and social identity in the context of wild pigs and other invasive species.

Key words: human–wildlife interactions, identity, machine learning, sentiment, social media, Twitter, wild hogs, wild pigs

Wild pigs (*Sus scrofa)*, also known as feral swine or wild hogs, are a non-native, invasive species in the United States that causes significant damage to agriculture (Anderson et al. 2016, McKee et al. 2020); negatively impacts ecosystems through their rooting and wallowing behaviors; and poses a risk of disease transmission to humans, livestock, and companion animals (Brown et al. 2019). First introduced by Spanish explorers in the sixteenth century, wild pig populations have increased in size and distribution and are now established in an estimated 35 U.S. states, with a total estimated population of up to 6.9 million individuals (Mayer and Brisbin 2008, Goedbloed et al. 2013, Lewis et al. 2019, Boyce et al. 2020). The growth of wild pig populations is partially attributed to their high intelligence, generalist diet, ability to acclimate to a wide range of regions and climatic conditions, and high fecundity (Bevins et al. 2014). Human activities, such as the translocation of wild pigs for sport hunting, have also contributed to the expansion of wild pig populations (Grady et al. 2019). As the vast majority of wild pigs are likely located on private lands, enlisting the support and cooperation of landowners in controlling wild pig populations is critical to stemming their growth. This points to the need for social science research that can inform effective public outreach and engagement on issues of wild pig damage and control.

For decades, surveys have been the most prevalent method to study human attitudes, perceptions, and behaviors within the social sciences (Chew and Eysenbach 2010, Sloan et al. 2015). Despite their popularity, surveys have a number of disadvantages and limitations. For example, they can be costly to design, implement, and analyze, and there may be lags in data acquisition, limiting their usage for timely issues (Chew and Eysenbach 2010). Additionally, surveys have several associated biases, such as social desirability bias, which occurs when participants want to please the researcher or appear virtuous (Grimm 2010). Although surveys are useful, there is still work that needs to be done to improve and supplement them.

To address the challenges and limitations of surveys and other traditional social science methods (e.g., interviews), "big data"—data gathered from many sources (e.g., transactional and naturally occurring) that are massive in volume and expansive over time (Sloan et al. 2015)—are increasingly being used to investigate a range of social phenomena. One commonly used source of big data is social media-generated content. Social media is a popular platform for disseminating and communicating information (Crooks et al. 2013). With its popularity, social media is also a source of data on a vast range of topics. Social scientists are increasingly leveraging this trove of data to evaluate and understand patterns of human–environment interactions and concerns (Song et al. 2020).

Twitter, the most popular social media microblogging platform, has an estimated 330 million monthly active users worldwide and generates billions of messages daily, making this social networking site an exceptional tool for studying diverse groups of people and their opinions (Tamburrini et al. 2015, Daume 2016, Kabakus and Simsek 2019). Twitter allows users to post short messages, known as "tweets," which can be up to 280 characters in length (Sansone et al. 2019). On average, 473,000 tweets are sent every minute, and 46% of Twitter users tweet daily (Madden et al. 2013). This results in an enormous quantity of data that holds great statistical power on the opinions of internet users with broad coverage across space and time (Tamburrini et al. 2015, Reyes-Menendez et al. 2018).

Twitter does, however, have certain limitations that may make it an inappropriate source for some studies. The platform provides little demographic information about users, for example, especially when compared to other social media platforms such as Facebook. Although metadata are available for Twitter users, these data may or may not include demographic information. As a result, Twitter is often regarded as a less reliable source of big data to study (Sloan et al. 2015). This skepticism for big data use—which is not limited to Twitter—may partially explain the paucity of innovative applications or tools that leverage big data in the context of natural resource-related research (Daume 2016). While social media data are not a replacement or proxy for data collected through traditional methods, these data can supplement other forms of data collection to enhance understanding of social phenomena. This raises practical and theoretical questions concerning how big data can be integrated into natural resource-related research. Such questions are of increasing urgency given the seriousness of today's conservation and environmental challenges (Rahman 2020). Given the stakes, big data content deserves both methodological exploration and assessment (Daume 2016).

In the context of research to inform wild pig management more specifically, Twitter data can contribute to this in several ways. First, wild pig-related experiences and perspectives shared on Twitter constitute a source of free or inexpensive data that can be leveraged to better understand the attitudes and management preferences of different stakeholder groups represented on the platform (Daume 2016). Second, a study by Chae et al. (2014) showed that social media data from citizens could be used by managers to inform quicker decision making during minor crises. Further, social media text mining was used in another study to detect and track diseases (Broniatowski et al. 2013), which could help wildlife agencies to

Table 1. Relevant metadata stored in this project.

Twitter given variable name	Variable label
text	Tweet text
user_desc	User defined description
lang	Language of tweet
created_at	Time tweet created
user_loc	User generated location
user_followers_count	Count of user followers
user_friends_count	Count of user frie
retweet_count	Count of retweets
user_name	Username

aid in detection of wild pig pathogens that may affect domestic pigs and other livestock. Web-based tools like Twitter that utilize machine learning algorithms allow managers to access the most recent and relevant information (e.g., human–wild pig interaction events) necessary to make rapid decisions (Humphries 2018).

The purpose of our study was to understand the sentiment expressed toward wild pigs by Twitter users in different online identity categories (e.g., outdoorsmen/women and agricultural occupations) using machine learning. Sentiment mirrors underlying emotions, which can be largely classified as positive, neutral, or negative (Becken et al. 2017). Research that examines online identities is valuable in a number of respects. First, the large number of users on Twitter allows us to study a wide array of individuals who might never participate in a survey, focus group, or other type of human subjects' study. Second, studying self-described online identities captures a wider variety of attitudes than doing so using traditional social science methods, like surveys. User descriptions within Twitter users' profiles are considered online identity expressions for this paper (Priante et al. 2016). The specific objectives were to: (1) identify the overall sentiment expressed toward wild pigs in relevant tweets and (2) determine the extent to which sentiment varied by online identity.

Methods

Data collection

The data collection targeted tweets posted to Twitter. These messages are accessible through Twitter's Application Program Interfaces (APIs). An API is a computing interface to a program (Twitter) that makes a connection with servers to retrieve specific information (www.dictionary.com). The information retrieved is based on a pre-defined set of conditions, or filters relating to the objects of interest. For this study, we used Twitter's premium package to access the search API, which retrieved tweets that matched a set of pre-defined key words, such as "wild pigs," "wild hogs," and "feral swine." Based on our reading of the API documentation, wildcards were not allowed. We investigated searching for both the singular and plural but found that many results were duplicated. To avoid this, we focused on the plural form, which was more commonly used. We used these different terms due to the lack of consistency in how scientists, wildlife managers, and the general public refer to the species (Keiter et al. 2016). All 3 key word phrases had the potential to capture a broad array of users' opinions about wild pigs and the online identities involved. Tweets selected for analysis met 2 criteria for inclusion. Identified tweets had to be (1) written in English and (2) relevant to the wild pigs/hogs/swine theme. Geolocation was not a criterion for inclusion, as geotagged tweets account for only about 1% of all messages sent via Twitter (Longley et al. 2015). We acknowledge that our sample has some content outside of the United States. However, limiting our dataset to retrieve just U.S. tweets would have reduced our sample size to an unusable number of examples. We evaluated tweets posted between May 1 and November 4, 2019. This time frame allowed ample time to sufficiently capture a large and diverse sample of tweets. For example, metadata fields downloaded from the API included not only the text within the tweet, but also timestamp, user profile description, username, user followers, user friend count, and retweet information (Table 1). Fields that were partially filled provided no utility to us (i.e., geolocation).

Python is, for the purpose of this paper, an open-source computer programming software that is used to improve quality, productivity, and integration (Lutz 2001). Python 3.7 and a collection of established libraries were used to scrape tweets specific to the identified key words and to analyze our data. Given the large quantity of data we collected, we used a 2-step process to address our research objectives. First, we manually labeled a random sample of tweets from our

Table 2. Examples of coded tweets.

Relevance and sentiment
Relevant
• Sounders of wild hogs are the reason I carry a firearm while riding my bicycle early in the morning in Arizona. I've had them charge me, but fortunately, I've not had to shoot at one yet.
• The meme went viral, but wild pigs are a serious threat.
• Wild pigs causing 'ecological disaster' as they spread rapidly across Canada, survey says.
Irrelevant
• @NetflixFilm @netflix wild hogs
• Are there any wild guinea pigs or do they only live as pets?
Positive sentiment
• yes, that's the malay name for bearded pigs. they are known to be gardeners of the forests; they reshape soil to help organic matter decomposition. these wild pigs provided meat for humans living in guaniah over the last 40,000 years.
• More project fear around the steady recovery of Europe's iconic wildlife. Wild boar play a crucial role in the healthy functioning of European ecosystems. Referring to them as wild hogs´ or a feral pig is a way of delegitimizing their place here.
Neutral sentiment
• WILD HOGs fleeing from flood waters on overtopped levee in St Marys Parish, LA from hwy 317!!!
• I don't guess I know the difference. Feral hogs aren't the same as wild pigs?
Negative sentiment
• A prime example is wild hogs. They impact habitats about the same as if you ran heavy equipment over it. They just decimate ground nesting birds and animals. They dirty water with mud and feces, and they're REALLY REALLY mean.
• There are numerous ways to deal with the issue of wild hogs, and assault rifles aren´t one of them.

larger collection using a pre-determined coding scheme for relevancy (n = 1,360), sentiment (n = 926), and online identities (n = 1,363). We based the size of this random sample on time constraints associated with this project and an examination of how classifier accuracy changed as we added more labeled examples. In particular, we stopped labeling when the returns to additional labeled data dropped to near-zero and we were in the region of strongly diminished returns. Next, we trained a series of machine learning algorithms on the labeled data and used them to classify and assign sentiment to all tweets. Machine learning, for the purpose of this paper and in the context of wild pigs, focuses on the computational process to aid in the understanding of basic algorithmic principles to train computers to learn from data (Blum 2007).

Measurement of key concepts and coding schemes

Relevance. We measured relevancy of tweets as being closely connected or appropriate to wild pigs. We used a binary (0,1) manual classification scheme to code a sample of tweets. An example of a non-relevant topic found throughout the manual coding stage included the movie *Wild Hogs*. Further, the entire search term was needed to evaluate relevancy accurately (i.e., "wild pig," not just "pig"). If the tweet was not fully comprehensible due to a lack of context or complete sentences, it was also considered irrelevant. Tweets with URL links and no other content were also excluded (e.g., "Wild pigs https://t.co/cEi0pyEqVC") for ease of measurement purposes. Specifically, we were interested in written sentiment toward wild pigs, not URL links that may include unrelated videos.

Sentiment. Sentiment analysis and opinion mining are forms of data analysis used to evaluate attitude expression within text (Fink et al. 2020). Our definition for sentiment scoring is derived from Becken et al. (2017), which includes a logical approach that transforms subjective text into meaningful information that can be analyzed to determine the emotional tone behind textual data for the purposes of understanding opinions.

Analysis of this kind comes with challenges, such as streamlining complex text so that a clear, overriding context can be recognized and inferring meaning from grammatical mistakes with ease (Becken et al. 2017). Although these challenges may be cumbersome, sentiment analyses have been used broadly across various disciplines to examine topics such as policy information, public health issues, disease outbreaks, and to communicate the importance of conservation science to professionals (Chew and Eysenbach 2010, Culnan et al. 2010, Merchant et al. 2011, Paul and Dredze 2011, Bombaci et al. 2016).

Tweets for this study were analyzed as an opinion toward the object of interest, wild pigs, and were chosen because we wanted to understand the emotional tone behind the tweet in order gain a full comprehension of opinions shared online. For the coding scheme, we used polarity 1, 0, and -1 for positive, neutral, and negative sentiment, respectively. We provide examples of manually coded tweets (Table 2).

Online Identity. To measure online identity, we used the metadata field known as "user description," which is also known as "feed identity" in some literature (Walton and Rice 2013). Online identities are defined as hobbies/interests, occupations, or sociodemographic characteristics. Here, individuals can fill in a description about themselves, usually making statements about attitudes or beliefs, hobbies, and sometimes information relating to employment (Sloan et al. 2015). A codebook (Appendix 1) was created to include "broader codes," "finer codes," and "explanations" for each identity. The first author coded all sample tweets from the dataset. A combined inductive and deductive approach was used to formulate categories of online identities. Before analysis, the first author chose online identities most relevant to wild pig issues (i.e., farmer, rancher). Next, other identity categories were chosen after thorough examination of a subset of user description profile observations during the relevancy and sentiment analyses phases to ensure all identities were being captured within the identity analysis phase. While coding, an iterative process was integrated to capture all other non-predetermined identities. With each new online identity that emerged, the author continuously revised the coded data. Once the coding phase was finished, we were left with 8 overarching "broader" identities including occupational identity, gender and sexual orientation, spousal and parental, religious, political, ethnicity, interest/hobby, and membership/government identities. Explanations of these broader identities are as follows: (1) occupational identity: self-described based on career, profession, or occupations; (2) gender and sexual orientation identity: self-described based on gender and sexual orientation; (3) spousal and parental identity: self-described based on spousal and parental relationships, including grandparent identities; (4) religious identity: self-described based on membership in religious groups; (5) political identity: self-described based on political affiliation, parties/groups relating to politics; (6) ethnicity identity: self-described based on ethnic group relation; (7) interest/hobby identity: self-described based on activities, interests, or hobbies in which an individual participates or has an affinity; (8) membership/governmental identity: self-described membership affiliation with a governmental agency, organization, or university.

For each of the 8 broader categories, another coding scheme was created to narrow down, in more detail, subcategories associated with each identity, known as "finer codes." The finer-coded categories included 18 occupations, 4 genders, 3 spousal-related, 5 parent/grandparent-related, 7 political affiliations, 8 religious orientations, 7 sexual orientations, 6 ethnicities, 29 hobbies/interests, and 4 membership/governmental affiliations. Overall, when a code was unclear or did not match any of the categories for both the "broader codes" and "finer codes" categories during the manual labeling phase, that portion was left blank.

For some of the other categories, the Twitter user had to use the term within their user description to be considered in the analysis or use opposite or negative expression regarding a category. Lastly, if the individual placed emphasis on a hobby or interest, we made inferences on which category that individual will be placed using key words from the "finer codes" coding scheme. The hobbies categories included "sports" and "animal lover/advocate" for the mention of the National Football League and "lover of wildlife" expressions.

The dataset in which the manual coding scheme was created was then applied to our identity classifier, mentioned in the identity classification section. A binary classification scheme (0 = no, 1 = yes) was created if an individual fell into any of the categories. In many cases, individuals fell into multiple identity categories, meaning these categories were not mutually exclusive of each other. If the resulting category size was <0.2, we did not include the category in the remainder of the analysis. Additionally, if the category was directly irrelevant to wild pig-related issues (i.e., "pro-life" or "feminism"), it was removed. To aid in classification accuracy, some identities were bundled post-hoc into a single category (Appendix 2). For example, of the 3 "finer" coded categories, "outdoorsman/woman," "angler," and "hunter," all were bundled into 1 identity called "outdoorsman/women."

Relevance classification. The objective of this process was to label all 48,557 tweets to be able to evaluate demographics among relevant tweets. For this purpose, we used a machine learning approach in which we trained a classification algorithm on a smaller subset of manually labeled data and then used the trained algorithm to label all data. Our labeled training data consisted of 1,360 tweets that were randomly selected from each batch of search results with probabilities weighted by the size of each batch so that we obtained a set of sample tweets that was representative of all the tweets we collected. Of the tweets that were manually labeled, 70% were relevant.

After labeling our sample data, we investigated 5 simple algorithms (i.e., naïve bayes, support vector machine, logistic regression, standard multilayer perception, and random forest) based on a bag-of-words approach. To evaluate algorithm performance, we relied on k-fold cross-validation with 5 folds and 4 different accuracy metrics (accuracy, precision, recall, f1). In the k-fold procedure, we split the data into 5 parts, trained on 4 of those parts, and validated on the remaining part. The training process was repeated a total of 5 times such that each fold was used for validation exactly once. Accuracy metrics from each validation fold were retained and then averaged across the 5 validation folds to get an estimate of expected out-of-sample accuracy.

After evaluating the simple algorithms, we additionally investigated a more sophisticated approach that used word embeddings and a convolutional neural network (CNN). Word embeddings refer to vectors that represent the meaning of a word. These are typically extracted from algorithms that have been trained on very large amounts of text. As a result, word embeddings are available for nearly every English word. The advantage of using word embeddings is that words that appear in similar context tend to have similar embedding vectors. Furthermore, words that only appear in validation or test data will have known embeddings, and if the algorithm has been exposed to similar vectors in training, it can extract relevant information from words it has not seen in training. Finally, by representing each tweet as a sequential vector of word embeddings, we could exploit word order and additional context information to determine relevancy. We examined 2 different sources of word embeddings, including Stanford's GloVe embeddings that were trained specifically on tweets and Google's more general Word2Vec embeddings (https://code.google.com/archive/p/word2vec/; Pennington et al. 2014). Although a full presentation of CNNs is beyond the scope of this paper, they are a type of neural network that reduces the number of weights that need to be estimated (Le Cun et al. 1990). They are common in computer vision and natural language processing applications for this reason. Interested readers can find a complete background and presentation of CNNs in Goodfellow et al. (2017). Similar to various forms of discrete-choice regression models (e.g., multinomial logit), the output of CNNs in the context of a classification problem is a set of class probabilities where the classes correspond to the unique set of labels applied in the training data.

Sentiment evaluation. The process used for sentiment estimation was similar to what we used for relevance. We began by labeling the same 1,360 tweets with a measure of sentiment toward wild pigs that took the values -1, 0, or 1 for negative, neutral, and positive, respectively. However, we only included relevant tweets (n = 926) in the remainder of the training process because we wanted to estimate sentiment toward wild pigs specifically. Of the labeled tweets, about 43% were labeled negative, 43% neutral, and 14% positive toward wild pigs.

We then trained the same set of algorithms and evaluated the accuracy of each using the same k-fold cross-validation procedure.

After selecting the best performing algorithm, we retrained on all labeled data and then labeled each of the 48,557 tweets with the trained algorithm. Although we predicted a discrete measure (-1, 0, 1) of sentiment for each tweet, we also calculated expected sentiment for each tweet as:

$$E[sentiment] = -1 * \widehat{\Pr}(-1) + 0 * \widehat{\Pr}(0) + 1 * \widehat{\Pr}(1) = -\widehat{\Pr}(-1) + \widehat{\Pr}(1)$$

where $\widehat{\Pr}(\cdot)$ are the class probabilities given by the classification algorithm. This is a valuable measure because it better accounts for conflicting language in the tweet and any ambiguity in our labeling process. A number equal to 1 would imply our algorithm is certain that true sentiment lies somewhere >0. Likewise, a number equal to -1 implies that the algorithm is completely sure that true sentiment lies somewhere <0.

Identity classification. Identity classification was a more challenging classification problem for several reasons. First, it relied on text in the user description field of the user's profile, and the amount of information in this field was often sparse. Second, there were 33 different categories that we used to classify users. We again relied on the same sample of tweets we used for the relevance and sentiment analyses. Most rows had zeros, but some rows had multiple identities. The multi-label nature of identity classification necessitated a modification to the output layer of our CNN. In the sentiment CNN, we used a SoftMax activation function in the final layer to ensure the probabilities assigned to -1, 0, and 1 summed to 1. In the identity problem, we had >2 classes, but we did not want to restrict the sum of probabilities to 1, as a user could belong to multiple identity categories. Because we essentially had 33 binary classification problems, we specified an output layer with 33 nodes, each with a sigmoid activation function. After final tuning, the algorithm was retrained and applied to all collected tweets.

Sentiment by identity classification. Finally, we computed identity-specific sentiment by averaging the sentiment of each tweet weighted by the estimated probability that the user belonged to the identity. Thus, for a given identity, tweets from users that we are more confident belong to the identity get weighted more heavily. The first step was to remove non-relevant tweets, which left us with 36,739 tweets. We then computed the weighted mean sentiment for identity according to

$$sentiment_i = \sum_{t=1}^{36,739} \widehat{Pr}(i)_t * E[sentiment_t]$$

which states that, when calculating the average sentiment of an identity, the sentiment of a given tweet is weighted by the probability that the author belonged to the identity in question.

Results

From May 1 to November 4, 2019, 48,557 total tweets were collected and stored in a database. A breakdown of the number of total tweets (not including relevancy) by search terms are as follows: "feral swine" (n = 3,622), "wild hogs" (n = 25,274), and "wild pigs" (n =19,661). The highest count of tweets was generated during the week of August 5, 2019 (Figure 1). This was due in large part to the "30–50 feral hogs" meme that went viral on August 4, 2019 (see Figure 1).

Relevance classification results

Our results indicated that the CNN based on Google's Word2Vec was the best performing algorithm with an accuracy approaching 90%. It was the best performer in 3 of the accuracy metrics, including the 2 general metrics. The final result of this classification exercise was that 93% of rows from the feral swine search, 69% of rows from the wild pig search, and 75% of rows from the wild hog search were labeled relevant. Thus, we substantially reduced the number of irrelevant tweets in our analysis through this exercise. Additionally, to ensure within coder reliability, the same dataset was manually coded twice: once on November 13, 2019 and again on December 16, 2019. The within coder reliability, Cronbach's alpha was 0.993, providing excellent internal consistency.

Sentiment estimation results

The Word2Vec-based CNN was again the best performing algorithm with an accuracy of 72.5%. The architecture and optimization methods were unchanged from the relevance

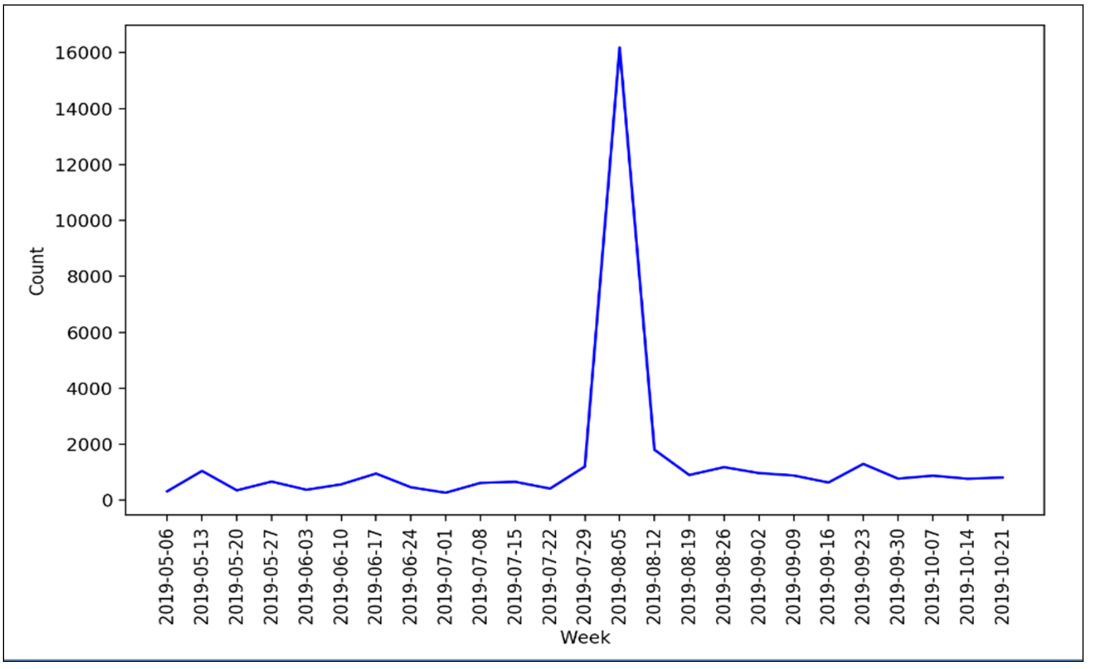

Figure 1. Tweet count over time. A tweet emerged on August 4, 2019 in response to Jason Isabell, a musician, about his opinion that "no one needs an assault weapon." The response tweet that went viral read, "Legit question for rural Americans - How do I kill the 30–50 feral hogs that run into my yard within 3–5 mins while my small kids play?" This resulted in numerous tweets regarding the term "30–50 feral hogs" (*Sus scrofa*).

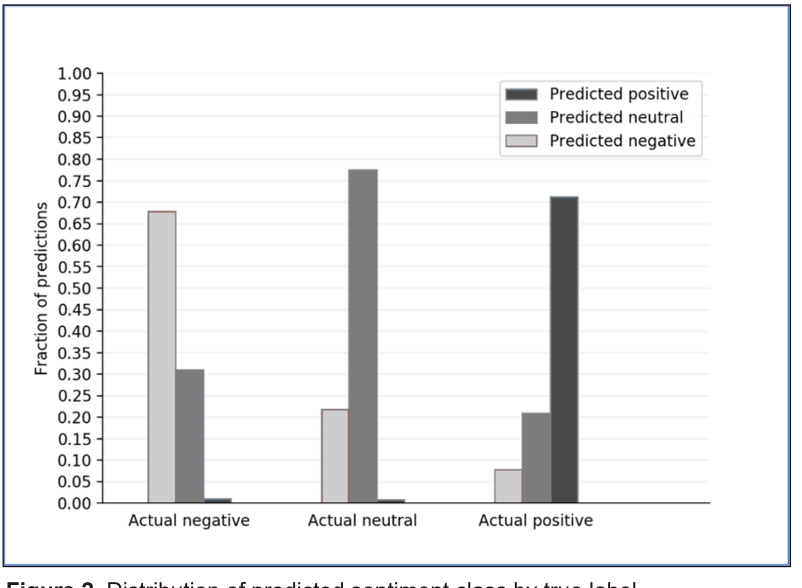

Figure 2. Distribution of predicted sentiment class by true label.

classifier with the exception of a slightly higher dropout rate (0.5 instead of 0.4). We also examined how incorrect predicted classes were distributed in the data (Figure 2).

Our classifier displayed the worst results on true positives. This was expected given the relatively small number of these examples in the training data. We also examined how expected sentiment varied across the data (Figure 3). These results largely mirror those displayed in Figure 2, with expected sentiment for true positives displaying the most variability.

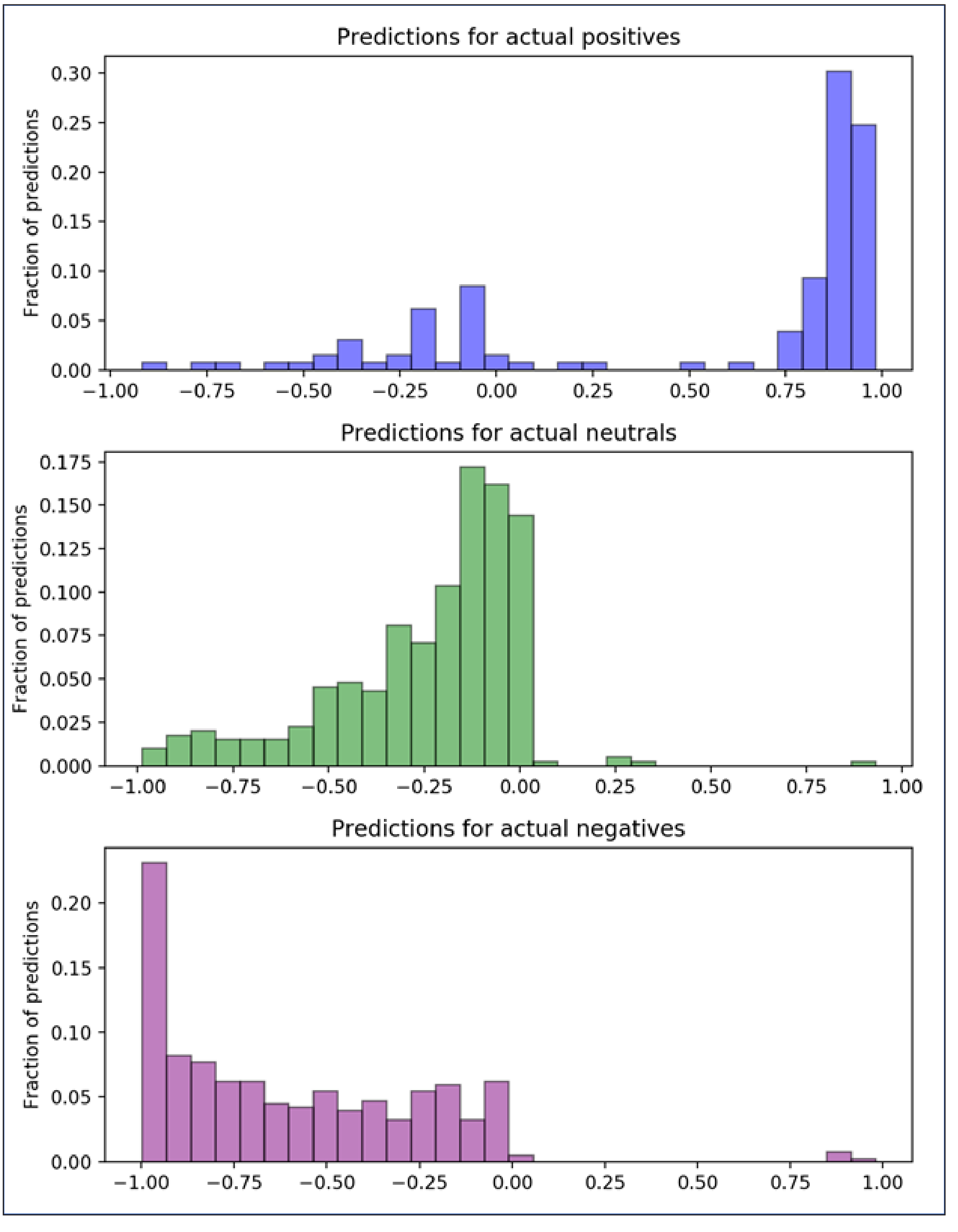

Figure 3. Distribution of expected sentiment by true sentiment label.

After labeling all 48,557 tweets, we plotted the distribution of sentiment across the original 3 search terms (Figure 4). The distributions for search terms "wild hog" and "wild pigs" are heavily skewed to the left, indicating more negative sentiment. The "feral swine" search term, though mostly negative, has a more symmetrical distribution.

Identity classification results

Two accuracy metrics were used: simple accuracy and exact match ratio (EMR). Because there were so many zeros (i.e., 1 = yes, 0 = no for an identity in question) in the labels, achieving high accuracy was straightforward. Our CNN achieved accuracy of about 98.5%. However, we note that this only marginally improved the

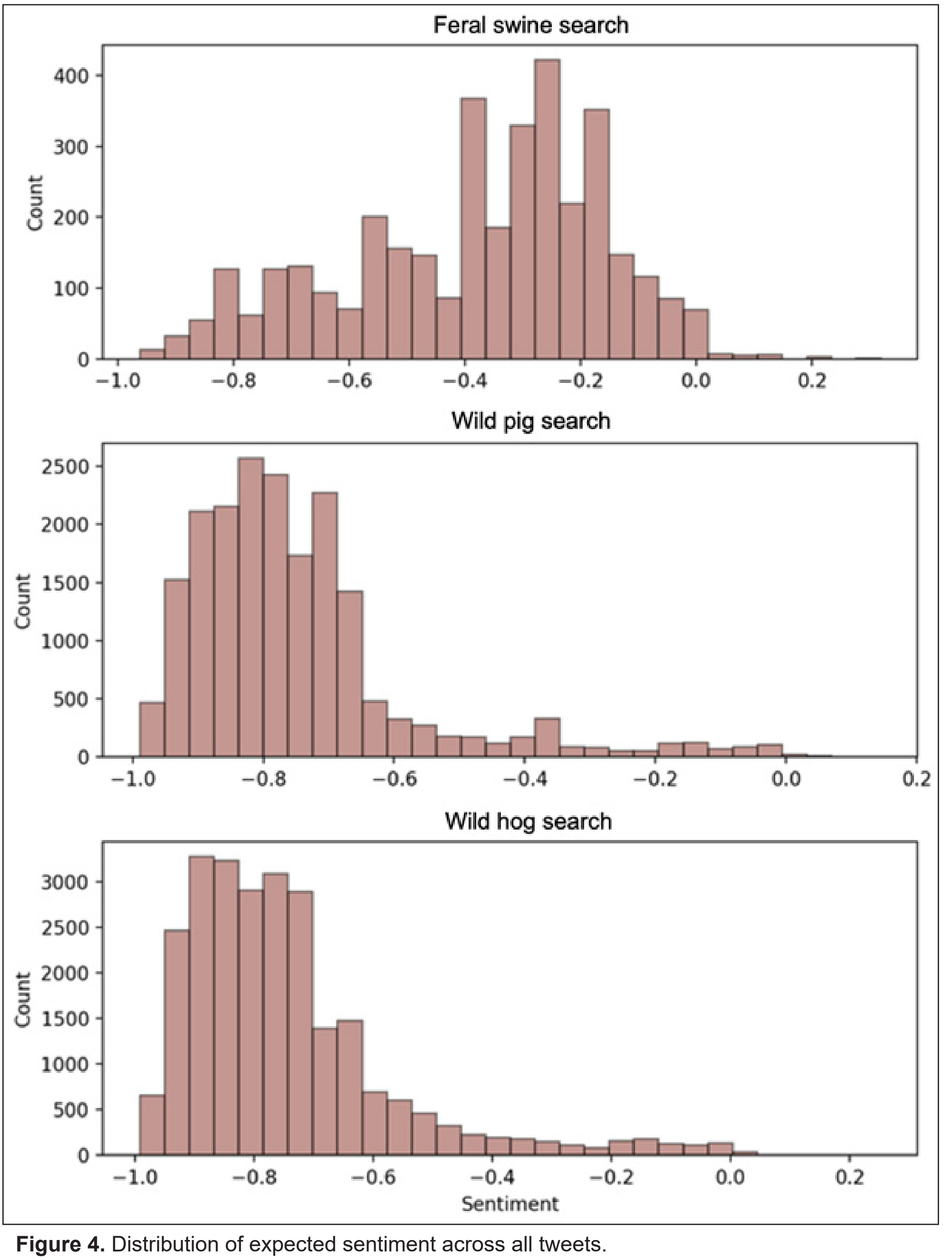

Figure 4. Distribution of expected sentiment across all tweets.

accuracy of labeling all tweets zero. Thus, EMR is a better metric. In the case of EMR, a row is deemed correct if all classes are correctly predicted for that row. Our EMR was 67.6%. This relatively low EMR is less problematic than it may appear given we were not interested in discrete identity labels but rather the probability that a user belongs to each identity.

The identities with the largest representation were "female" and "journalism and media communication." About 6% of users fell into each of these categories, making the sample unbalanced. Averaging across groups, about 1.7% of all users fell into a given group. In general, it was unlikely from our sample for an individual to fall into one of the identity categories.

Table 3. Sentiment by identity categories.

Identity category	Sentiment average	Category size
Academic occupations	-0.74	499.99
Agriculture occupations	-0.63	178.29
Animal advocate interests/hobbies	-0.73	362.91
Anti-environment interests/hobbies	-0.61	1.00
Armed Forces occupations	-0.74	55.57
Asian	-0.77	11.44
Black/African American	-0.76	3.34
Business and computers occupations	-0.65	109.56
Conservative ideology	-0.73	159.59
Criminal justice, political science, and legal aid occupations	-0.72	26.32
Female	-0.72	1,735.71
Government organizations	-0.78	0.24
Healthcare occupations	-0.72	23.61
Journalism and media communication occupations	-0.72	1,580.20
Latino	-0.74	2.44
LGBTQ	-0.72	304.57
Liberal ideology	-0.67	44.93
Life and natural science occupations	-0.71	30.79
Male	-0.72	1,506.78
Native American	-0.76	3.55
Natural resources occupations	-0.77	231.83
Outdoorsman/women interests/hobbies	-0.77	16.69
Parent	-0.74	620.67
Politics occupations	-0.72	29.02
Pro-environment interests/hobbies	-0.75	89.58
Pro-guns interests/hobbies	-0.75	18.17
Religious	-0.75	133.47
Spouse	-0.77	472.87
They	-0.73	232.50
White	-0.65	3.46
Wild pigs interests/hobbies	-0.57	68.17

Sentiment by identity results

The average sentiment for each identity category along with category size are shown (Table 3). Category size should be interpreted as an indication of sample size; it is the sum, across all users, of the probability of belonging to the identity. Sentiment toward wild pigs is measured on a -1 to 1 scale, -1 being negative and 1 being positive. As shown, the most negative sentiment toward wild pigs (-0.78) includes Twitter users that affiliate with governmental organizations. On the opposite end, the least negative sentiment toward wild pigs (-0.57 and -0.61) are Twitter users that mention wild pigs as a part of their hobbies or interests, as well as users that included anti-environmental descriptions. However, the sizes of those identity categories are small, with sentiment scores of 68.17 and 1.00, respectively. The second-least negative sentiment toward wild pigs was the agriculture identity, which has a category size of 178.29. This means that individuals who identify with agriculture-related occupations, like farmer or rancher, viewed wild pigs relatively more favorably. The largest representation of identities on Twitter in our sample were female, male, journalism and media communication occupations, parent, spouse, and academic occupations.

Discussion

We developed a methodological tool that harnesses large datasets using machine learning techniques, which we believe could help researchers more easily investigate and examine content related to human–wildlife interactions in the future. After extracting relevant data from Twitter, we applied the tool to evaluate sentiment and online identities pertaining to tweets about a natural resource issue of critical concern to management: invasive wild pigs. Of the extracted total tweets, 70% remained relevant after applying the machine learning algorithm. This step was essential because it allowed a filtration process to occur, clearing out all tweets that did not relate to our research objectives. Of the online identities examined with this new tool, the sample was highly unbalanced, indicating that although the machine learning algorithm exhibited a fairly high degree of discriminatory power, there is still opportunity moving forward to fine-tune the classifier to detect a greater number of online identities on Twitter. Ultimately, this tool provides an efficient method for analyzing large sets of social media data to better understand social phenomena. By combining this type of method with more traditional social science research methods (e.g., surveys and interviews), researchers can explore the role social media has on human sentiment and online identity.

In conducting a sentiment analysis, we determined that the majority of the tweets in our dataset were more negative than positive. In particular, the distribution of sentiment for the search terms "wild hog" and "wild pigs" was heavily skewed toward a negative sentiment. The "feral swine" search term, although negative, had a wider distribution of polarity, which may be explained by the identities of Twitter users who applied the term. The term "feral swine" was not commonly mentioned by the majority of users. Instead, the term was primarily used by academics and individuals from government agencies. These findings suggest that there may be potential confusion about the words used to describe wild pigs between the general public and the scientific and management communities, highlighting the importance of using commonly understood terminology in communication and outreach efforts relating to wild pig management.

In terms of the online identities of Twitter users in our dataset, we found that the largest groups were females and users working in journalism and media communication and in academia. Individuals from academia had slightly more negative sentiment toward wild pigs. Interestingly, we found that users who identified with agriculture-related occupations had more favorable sentiment toward wild pigs. This is in contrast to an earlier survey study, which found that the majority of farmers, ranchers, and landowners held negative attitudes toward wild pigs in Texas, USA (Adams et al. 2005). We speculate that this disparity may be partly due to geographical differences among farmers and ranchers who participated in the survey study and the Twitter users in our study. For example, farmers and ranchers in Texas may have more negative attitudes toward wild pigs because of the higher wild pig densities and associated damages than farmers and ranchers located in areas with lower wild pig densities. Additionally, many of the farmers and ranchers represented in our dataset under the agriculture identity might not be directly impacted by wild pigs and, therefore, might express a more positive sentiment toward wild pigs.

Previous research has found that people engage in social media when they encounter or learn of an event that is outside of their daily norm (Cassa et al. 2013). This type of engagement was evident in our dataset with the "30–50 feral hog" meme that went viral during the sampling timeframe. The feral hog meme also contributed to a wide array of identities found on Twitter that may not have been detected otherwise. This example suggests that managers could use social media, as well as the tool we introduced, to track trends regarding invasive species on social media to then inform outreach and management efforts.

Nevertheless, there are limitations with social media research, including this study, which could potentially be addressed in future research. First, one of the criteria for inclusion in our analysis was that each tweet had to be written in English. However, inclusion of non-English language tweets in future studies could increase the robustness and generalizability of findings, particularly as they relate to wild pig issues in non-English speaking countries. Second, the individuals who actively engage on Twitter (i.e., by tweeting) may not be representative of Twitter users more broadly, as many users may only monitor tweets or use the platform sporadically. Thus, it is important to note that our research should not be interpreted as capturing the full array of sentiment and online identities that may exist relative to wild pigs. Third, because one of our objectives was to study online identities, we chose to include as many identities that emerged from our dataset as possible. Because of the large number of identities we identified and the relatively small number of individuals within any given identity, it was more difficult for the classifier to predict the probability of a Twitter user falling into a particular classification. Future research that streamlines the number of identity groups by focusing on online identities most salient to the issue of interest, in this case wild pigs, may therefore be warranted. Exploring other social media platforms that have more readily available demographic information should also be considered.

This research provides an important starting point for further investigation of the use of social media data in the context of natural resource-related issues. The tool we developed could lend itself to investigating other social phenomena on Twitter about wild pigs (Savage et al. 2013, Sloan et al. 2015). Sloan et al. (2015), for example, recommends using the user description field to investigate archives of tweets to determine hobbies and thus identify money spent on goods. In the context of wild pig management, researchers could explore this avenue to estimate money spent on wild pig-related activities (e.g., hunting) or economic losses due to wild pig damages. These categories of expenditures and losses could be further categorized by online identity. With refinements to the tool we developed, it could also be used to focus on geographic areas where wild pig populations are being monitored for management purposes. For example, Becken et al. (2017) used Twitter data to monitor the environment and human sentiment on the Great Barrier Reef (Queensland, Australia). They showed that collective knowledge provided from Twitter can complement traditional management strategies of monitoring important ecological areas. Lastly, this tool could be refined to identify contextual themes in social media data. For example, a more refined tool could identify what topics users are tweeting about in regard to wild pigs (e.g., concern, damage, hunting, gear used for hunts, etc.).

Social media provides a vast amount of largely untapped data for investigating questions relating to human–wildlife interactions, including those involving wild pigs. It is our hope that researchers will use and refine the methods we developed in this study to explore such questions. Innovations in techniques for analyzing large social media datasets may ultimately contribute to innovative solutions for managing some of the most intractable human–wildlife problems.

Acknowledgments

We wish to thank the U.S. Department of Agriculture (USDA), Animal and Plant Health Inspection Service, Wildlife Services, National Feral Swine Damage Management Program, and the National Wildlife Research Center for making this research possible. Additional support was provided by Colorado State University. The findings and conclusions in this publication are those of the authors and should not be constructed to represent any official USDA or U.S. Government determination or policy. Comments provided by J. Tomeček, HWI associate editor, and 2 anonymous reviewers greatly improved an earlier version of our paper.

Appendices

Appendix 1 and 2 can be viewed as supplemental file downloads at https://digitalcommons.usu.edu/hwi/vol15/iss1/16.

Literature cited

Adams, C. E., B. J. Higginbotham, D. Rollins, R. B. Taylor, R. Skiles, M. Mapston, and S. Turman. 2005. Regional perspectives and opportunities for feral hog management in Texas. Wildlife Society Bulletin 33:1312–1320.

Anderson, A., C. Slootmaker, E. Harper, J. Holderieath, and S. A. Shwiff. 2016. Economic estimates of feral swine damage and control in 11 US states. Crop Protection 89:89–94.

Becken, S., B. Stantic, J. Chen, A. R. Alaei, and R. M. Connolly. 2017. Monitoring the environment and human sentiment on the Great Barrier Reef: assessing the potential of collective sensing. Journal of Environmental Management 203:87–97.

Bevins, S. N., K. Pedersen, M. W. Lutman, T. Gidlewski, and T. J. Deliberto. 2014. Consequences associated with the recent range expansion of nonnative feral swine. BioScience 64:291–299.

Blum, A. 2007. Machine learning theory. Carnegie Melon University, School of Computer Science, 26. Carnegie Melon University, Pittsburgh, Pennsylvania, USA.

Bombaci, S. P., C. M. Farr, H. T. Gallo, A. M. Mangan, L. T. Stinson, M. Kaushik, and L. Pejchar. 2016. Using Twitter to communicate conservation science from a professional conference. Conservation Biology 30:216–225.

Boyce, C. M., K. C. VerCauteren, and J. C. Beasley. 2020. Timing and extent of crop damage by wild pigs (*Sus scrofa Linnaeus*) to corn and peanut fields. Crop Protection 133:105131.

Broniatowski, D. A., M. J. Paul, and M. Dredze. 2013. National and local influenza surveillance through Twitter: an analysis of the 2012–2013 influenza epidemic. PLOS ONE 8(12): e83672.

Brown, V. R., M. C. Marlow, R. M. Maison, T. Gidlewski, R. Bowen, and A. Bosco-Lauth. 2019. Current status and future recommendations for feral swine disease surveillance in the United States. Journal of Animal Science 97:2279–2282.

Cassa, C. A., R. Chunara, K. Mandl, and J. S. Brownstein. 2013. Twitter as a sentinel in emergency situations: lessons from the Boston marathon explosions. PLOS Currents 5.

Chae, J., D. Thom, Y. Jang, S. Kim, T. Ertl, and D. S. Ebert. 2014. Public behavior response analysis in disaster events utilizing visual analytics of microblog data. Computers & Graphics 38:51–60.

Chew, C., and G. Eysenbach. 2010. Pandemics in the age of Twitter: content analysis of Tweets during the 2009 H1N1 outbreak. PLOS ONE 5(11): e14118.

Crooks, A., A. Croitoru, A. Stefanidis, and J. Radzikowski. 2013. #Earthquake: Twitter as a distributed sensor system. Transactions in GIS 17:124–147.

Culnan, M. J., P. J. McHugh, and J. I. Zubillaga. 2010. How large US companies can use Twitter and other social media to gain business value. MIS Quarterly Executive 9:243–259.

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

Fink, C., A. Hausmann, and E. Di Minin. 2020. Online sentiment towards iconic species. Biological Conservation 241:108289.

Goedbloed, D., H. J. Megens, P. Van Hooft, J. M. Herrero-Medrano, W. Lutz, P. Alexandri, R. P. M. A. Crooijmans, M. Groenen, S. E. Van Wieren, R. C. Ydenberg, and H. H. R. Prins. 2013. Genome-wide single nucleotide polymorphism analysis reveals recent genetic introgression from domestic pigs into northwest European wild boar populations. Molecular Ecology 22:856–866.

Goodfellow, I., Y. Bengio, and A. Courville. 2017. Deep learning. MIT Press, Cambridge, Massachuettes, USA.

Grady, M. J., E. E. Harper, K. M. Carlisle, K. H. Ernst, and S. A. Shwiff. 2019. Assessing public support for restrictions on transport of invasive wild pigs (*Sus scrofa*) in the United States. Journal of Environmental Management 237:488–494.

Grimm, P. 2010. Social desirability bias. *In* J. Sheth and N. K. Malhotra, editors. Wiley international encyclopedia of marketing. Part 2: marketing research. John Wiley & Sons, Inc., Hoboken, New Jersey, USA.

Humphries, G. R. 2018. How the internet can know what you want before you do: web-based machine learning applications for wildlife management. Pages 335–351 *in* G. Humphries, D. Magness, and F. Huettmann, editors. Machine learning for ecology and sustainable natural resource management. Springer, Cham, Switzerland.

Kabakus, A. T., and M. Simsek. 2019. An analysis of the characteristics of verified Twitter users. Sakarya University Journal of Computer and Information Sciences 2:180–186.

Keiter, D. A., J. J. Mayer, and J. C. Beasley. 2016. What is in a “common” name? A call for consistent terminology for nonnative *Sus scrofa*. Wildlife Society Bulletin 40:384–387.

Le Cun, Y., J. S. Denker, and S. A. Solla. 1990. Optimal brain damage. Pages 598–605 *in* D. Touretzky, editor. Proceedings of the IEEE Conference on Neural Information Processing Systems, Denver, Colorado, USA.

Lewis, J. S., J. L. Corn, J. J. Mayer, T. R. Jordan, M. L. Farnsworth, C. L. Burdett, K. C. VerCauteren, S. J. Sweeny, and R. S. Miller. 2019. Historical, current, and potential population size estimates of invasive wild pigs (*Sus scrofa*) in the United States. Biological Invasions 21:2373–2384.

Longley, P. A., M. Adnan, and G. Lansley. 2015. The geotemporal demographics of Twitter usage. Environment and Planning A: Economy and Space 47:465–484.

Lutz, M. 2001. Programming python: powerful object-oriented programming. O'Reilly Media, Inc., Sebastopol, California, USA.

Madden, M., A. Lenhart, S. Cortesi, U. Gasser, M. Duggan, A. Smith, and M. Beaton. 2013. Teens, social media, and privacy. Pew Research Center 21:2–86.

Mayer, J. J., and I. L. Brisbin 2008. Wild pigs in the United States: their history, comparative morphology, and current status. University of Georgia Press, Athens, Georgia, USA.

McKee, S., A. Anderson, K. Carlisle, and S. A. Shwiff. 2020. Economic estimates of invasive wild pig damage to crops in 12 US states. Crop Protection 132:105–105.

Merchant, R. M., S. Elmer, and N. Lurie. 2011. Integrating social media into emergency-preparedness efforts. New England Journal of Medicine 365:289–291.

Paul, M. J., and M. Dredze. 2011. You are what you Tweet: analyzing Twitter for public health. International AAAI Conference on Web and Social Media 20:265–272.

Pennington, J., R. Socher, and D. Manning. 2014. GloVe: Global vectors for word representation. Pages 1532–1543 *in* A. Moschitti, B. Pang, and W. Daelemans, editors. Proceedings of the Conference on Empirical Methods in Natural Language Processing, Doha, Qatar.

Priante, A., D. Hiemstra, T. Van Den Broek, A. Saeed, M. Ehrenhard, and A. Need. 2016. #WhoAmI in 160 characters? Classifying social identities based on twitter profile descriptions. Page 55–65 *in* D. Bamman, A. S. Doğruöz, J. Eisenstein, D. Hovy, D. Jurgens, B. O'Connor, A. Oh, O. Tsur, and S. Volkova, editors. Proceedings of the First Workshop on NLP and Computational Social Science, Austin, Texas, USA.

Rahman, M. M. 2020. Environmental degradation: the role of electricity consumption, economic growth and globalisation. Journal of Environmental Management 253:109742.

Reyes-Menendez, A., J. Saura, and C. Alvarez-Alonso. 2018. Understanding #WorldEnvironmentDay user opinions in Twitter: a topic-based sentiment analysis approach. International Journal of Environmental Research and Public Health 15:2537.

Sansone, A., A. Cignarelli, G. Ciocca, C. Pozza, F. Giorgino, F. Romanelli, and E. A. Jannini. 2019. The sentiment analysis of tweets as a new tool to measure public perception of male erectile and ejaculatory dysfunctions. Sexual Medicine 7:464–471.

Savage, M., F. Devine, N. Cunningham, M. Taylor, Y. Li, J. Hjellbrekke, and A. Miles. 2013. A new model of social class? Findings from the BBC's Great British Class Survey experiment. Sociology 47:219–250.

Sloan, L., J. Morgan, P. Burnap, and M. Williams. 2015. Who tweets? Deriving the demographic characteristics of age, occupation and social class from Twitter user meta-data. PLOS ONE 10(3): e0115545.

Song, X. P., D. R. Richards, and P. Y. Tan. 2020. Using social media user attributes to understand human–environment interactions at urban parks. Scientific Reports 10:808.

Tamburrini, N., M. Cinnirella, V. A. Jansen, and J. Bryden. 2015. Twitter users change word usage according to conversation-partner social identity. Social Networks 40:84–89.

Walton, S. C., and R. E. Rice. 2013. Mediated disclosure on Twitter: the roles of gender and identity in boundary impermeability, valence, disclosure, and stage. Computers in Human Behavior 29:1465–1474.

Associate Editor: John M. Tomeček

Lauren M. Jaebker is a research associate in the Department of Human Dimensions of Natural Resources at Colorado State University (CSU) and contracted human dimensions specialist at the USDA-APHIS National Wildlife Research Center in Fort Collins, Colorado. She received a B.S. degree in fish, wildlife, and conservation biology and an M.S. degree in human dimensions of natural resources from CSU. Much of her research focuses on mixed-method approaches to understand and explain human–wildlife interactions.

Hailey E. McLean is a research associate at Colorado State University (CSU) and a contracted human dimensions specialist at the National Wildlife Research Center in Fort Collins, Colorado. She received a B.S. degree in fish, wildlife, and conservation biology and an M.S. degree in human dimensions of natural resources from CSU. Her research centers on using interdisciplinary science frameworks to understand and explain human–wildlife interactions. The goal of her research is to aid in informing wildlife management in the United States.

Stephanie A. Shwiff is a supervisory research economist and project leader at USDA's National Wildlife Research Center and affiliate faculty at the University of Colorado–Boulder, where she teaches a class on benefit-cost analysis of food systems in the Masters of the Environment Program. At the National Wildlife Research Center, she leads the research project titled "Economics, operations research, and social dimensions of wildlife management." Her research is aimed at better understanding the economic impact of wildlife diseases and damage as well as developing, assessing, and improving wildlife management techniques.

Keith M. Carlisle is a research associate in the Department of Human Dimensions of Natural Resources, Warner College of Natural Resources, at Colorado State University in Fort Collins, Colorado. He also serves as a human dimensions specialist at the USDA-APHIS National Wildlife Research Center in Fort Collins.

Tara L. Teel is a professor in the Department of Human Dimensions of Natural Resources at Colorado State University (CSU). Much of her research has been devoted to examining human–wildlife relationships in the United States and globally, with a particular focus on employing concepts and methodologies from social psychology to study human values toward wildlife, attitudes and behaviors toward wildlife-related issues, and the social factors underlying human–wildlife conflict. She works closely with conservation agencies and organizations in the application of social science to inform their planning, management, and communication efforts. She received her Ph.D. degree in human dimensions of natural resources from CSU and M.S. and B.S. degrees in fisheries and wildlife management from Utah State University.

Alan D. Bright is a professor in the Department of Human Dimensions of Natural Resources at Colorado State University. His research has focused on a variety of human dimensions of natural resources issues, including social psychological aspects of recreation and tourism behavior as well as public values and attitudes toward natural resource management strategies such as wildland fire management and the creation of defensible space in the wildland–urban interface.

Aaron M. Anderson is a data scientist with a skill set that spans statistical methods for causal inference, frontier machine learning methods for prediction, digital signal and natural language processing, and data engineering. With >10 years of experience working as an individual contributor and leader in both pure research and software development, he has managed and delivered major projects in the fields of economics, wildlife management, epidemiology, criminology, and sports analytics.