Received: 23 July 2021 | Revised: 9 December 2021 | Accepted: 2 January 2022

DOI: 10.1111/tbed.14454

ORIGINAL ARTICLE

Transboundary and Emerging Diseases WILEY

Web crawling of social media and related web platforms to analyze backyard poultry owners responses to the 2018–2020 Newcastle Disease (ND) outbreak in Southern California

Joseph Gendreau[1] | Shayne Ramsubeik[2] | Maurice Pitesky[1]

[1] Department of Population Health and Reproduction, School of Veterinary Medicine-Cooperative Extension, University of California, Davis, California

[2] California Animal Health and Food Safety Laboratory (CAHFS), Turlock, California

Correspondence
Maurice Pitesky, Department of Population Health and Reproduction, School of Veterinary Medicine-Cooperative Extension, University of California, Davis, CA, USA.
Email: mepitesky@ucdavis.edu

Funding information
California Department of Food and Agriculture

Abstract

As social media becomes an ever-increasing staple of everyday life and a growing percentage of people turn to community driven platforms as a primary source of information, the data created from these posts can provide a new source of information from which to better understand an event in near real time. The 2018–2020 outbreak of Newcastle Disease (ND) in Southern California is the third outbreak of ND in Southern California within a 50-year time span. These outbreaks are thought to be primarily driven by non-commercial poultry (i.e. backyard and game fowl) in the region. Here we employed a commercial ‘web crawling’ tool between June of 2018 and July of 2020 which encompassed the majority of the outbreak in order to collect all available online mentions of ‘virulent Newcastle Disease’ (vND), the terminology commonly used by the California Department of Food and Agriculture (CDFA), United States Department of Agriculture (USDA), and the general public, in relation to the outbreak. A total of 2498 posts in English and Spanish were returned using a Boolean logic-based string search. While the number of posts was relatively small, their impact as measured by the number of visitors to the website and the number of people viewing the post (where provided) was much larger. Posts with negative sentiment were found to have a larger audience relative to posts with a positive sentiment. In addition, posts with negative sentiment peaked in May of 2019 which preceded the formation of the anti-depopulation group Save Our Birds (SOB). As the usage and impact of social media grows, the ability to utilize tools to analyze social media may improve both response and outreach-based strategies for various disease outbreaks including vND in Southern California which has a large non-commercial poultry population.

KEYWORDS
2018–2020 vND outbreak, social media, virulent Newcastle Disease, web crawling

1 | INTRODUCTION

Newcastle Disease (ND) is a highly contagious and, in the case of velogenic genotypes (Hidaka et al., 2021), highly lethal disease of poultry caused by virulent Newcastle disease viruses (NDVs), which are viruses of the genus Avian orthoavulavirus 1 (AOAV-1), previously known as Avian paramyxovirus 1 (APMV-1). (Dimitrov et al., 2019), While ND is endemic to many parts of the world, it is considered an exotic animal disease in the United States (Dimitrov et al., 2019). Recent outbreaks of ND in Southern California in 1971–1973, 2002–2003 and most recently 2018–2020 have all required extensive culling of millions of birds and caused significant economic damage to the poultry industry (Rue et al., 2011). Past ND outbreaks in the region began in gamefowl (GF) illegally imported for fighting and exhibition purposes

(USDA-ARS, 2016), which are especially prevalent in the Southern California counties of Los Angeles, Riverside and San Bernardino that were geographically associated with the three most recent outbreaks noted (USDA-APHIS, 1978, 2018). In this paper, GF refers to cocks kept specifically for fighting or exhibition purposes, or for breeding other birds for these purposes, and does not include cocks that are kept as part of a flock for meat, eggs and/or recreational purposes not aforementioned. Bird fighting and other GF events typically result in the frequent movement of birds to different properties, even when quarantines are in place due in part to financial incentive. Additionally, a heavily engrained tradition of backyard poultry (BYP) ownership exists in urban and semirural areas of Southern California (USDA, 2013). For the purposes of this paper, BYP is defined as privately held birds kept on the same premises as a residence. BYP flocks in this region are particularly vulnerable to infectious disease spread due to high spatial overlap with neighboring flocks, in some cases only separated by fencing, lack of strong biosecurity practices, gaps in disease knowledge and low vaccination rates (Elkhoraibi et al., 2014). The spatial overlap between GF events and BYP flocks in the region, and in some cases proximity to commercial poultry operations, further increases the potential for disease spread (Garber et al., 2007). The illegal nature of bird fighting in the United States as well as the large number and wide distribution of poultry owners in Southern California has historically posed significant challenges with respect to public perception of the governmental response and mitigation of the actual ND outbreaks as well as affecting behavior change through social involvement.

In 2019, 90% of all adults in the United States reported using the Internet (Pew Research Center, 2019a), and 72% of them reported using at least one social media platform (Pew Research Center, 2019b). As social media, social networking and content sharing become more ubiquitous, the associated data generated creates catalogues of information people choose to provide about themselves (Kennedy & Moss, 2015). The ability to use these data to predict outcomes are well documented in multiple fields including elections (Tumasjan et al., 2010), marketability of consumer goods (Shimshoni et al., 2009) and even box office revenue (Asur & Huberman, 2010). With respect to predicting infectious diseases, Google Flu Trends (GFT) (2009–2015) attempted to harness search engine data as an epidemiological predictor of flu outbreaks with varying results (Kandula & Shaman, 2019). One analysis of the project revealed that GFT reduced errors in Centers for Disease Control (CDC) predictive models by up to 52.7% compared to CDC data alone (Preis & Moat, 2014). More recent efforts combine sentiment analysis with various supervised and unsupervised Machine Learning (ML) algorithms to predict influenza (Broniatowski et al., 2013), dengue (Othman & Danuri, 2016) and other vector borne diseases (Jain & Kumar, 2018).

Web crawling and web scraping broadly encompass the process of gathering and analyzing big data from multiple online news sources and social platforms to provide insight into public attitudes towards a subject. Commonly used associated data analysis techniques include sentiment analysis, opinion mining, other natural language processing (NLP) and social network analysis (SNA) (López et al., 2012). Sentiment analysis and opinion mining employ machine learning algorithms to understand emotions – both positivity/negativity and more complex emotions like fear, anger, and happiness – expressed by social media users (Alamoodi et al., 2020; Singh et al., 2018). When considered in the context of a foreign animal disease like ND that affects both commercial and non-commercial poultry in a highly urban area, the ability to use these types of data can be employed to better understand the epidemiology of an outbreak, current sentiment towards first responders and overall accuracy of shared information. These types of results in turn can then be used for active response during the outbreak. The aim of this study is to investigate correlations between the timeline of the California Department of Food and Agriculture (CDFA) and US Department of Agriculture (USDA) regulatory response to the 2018–2020 ND outbreak and overall social media activity and sentiment during the outbreak.

2 | MATERIALS AND METHODS

2.1 | Data collection

Social media data related to ND was collected using an enterprise customer relationship management (CRM) service (Brandwatch, Brighton, UK) which employs a web crawler and Application Program Interfaces (APIs) to nine social media platforms, including Twitter, to collect online posts using Boolean strings. Since the primary purpose of the CRM tool is brand management (i.e. monitoring and managing conversations about the user's products or services), certain API connections provided only return results for the user's page or account. For example, the CRM service only provides Facebook data once the user authenticates their brand account with Facebook. For this reason, only data from the Twitter, Reddit and Tumblr APIs are included in results from the CRM tool.

A standard Boolean string search was set up initially searching for posts including both English and Spanish in the United States, Central America and the Philippines from 30 June 2018 to 22 July 2020. The following Boolean string was used in the search:

> (((chicken OR rooster OR 'hen' OR 'Hen') NOT pox)
>
> AND
>
> ('virulent Newcastle Disease' OR (Newcastle NEAR/20 disease) OR 'vND outbreak' OR NDV OR vND OR vNDV))
>
> NOT
>
> (flu OR influenza OR COVID-19 OR corona OR coronavirus OR COVID OR Sars-CoV-2)

The terms 'virulent Newcastle Disease' and 'vND' were included here since communications from USDA Animal and Plant Health Inspection Service (APHIS), CDFA and various academic extension groups in California refer to the disease using these terms. As a result, the majority of online discussions use these terms as well and 'vND' will be used to refer to posts that use this terminology.

The CRM tool employed only returns the exact term when quotation marks are used. The NEAR/20 operator returns results that include the following term within 20 words of the preceding term, thus posts containing 'velogenic', 'viserotropic', 'exotic' etc. will automatically be included when using 'Newcastle NEAR/20 disease'. Terms not inside of quotation marks are not case sensitive and may be part of a larger word. The AND, OR and NOT operators function as standard logical operators. To demonstrate how these operators work, consider the example post containing the text 'My vet told me about the Newcastle outbreak going on in California. I am worried my chickens might catch the disease,' and the simplified Boolean string '(chicken OR hen) AND ((Newcastle NEAR/20 disease) OR vND).' The 'chicken OR hen' section would return logical True since the word 'chickens' contains 'chicken', and the '(Newcastle NEAR/20 disease) OR vND' section would return logical True since 'disease' appears within twenty words of 'Newcastle'. The resulting expression 'True AND True' returns logical True, and the example post would be added to the list of results. If the simplified Boolean string was modified as '((chicken OR hen) AND ((Newcastle NEAR/20 disease) OR vND)) NOT California', then the example post would not be added since the inclusion of 'California' in the post returns logical False. A similar search was performed in Spanish. The Boolean string search returned all available posts from sites crawled by the CRM service that include 'chicken', 'rooster' or 'hen' and some variations of 'vND' or 'Newcastle disease'. The search also specified the exclusion of 'chicken pox', 'influenza', and all terms relating to 'COVID-19' to reduce noise in the data collected.

For each post, the CRM service records the domain, post date, url, location information, title, partial body text, Boolean string match positions, monthly site visitors and other post metadata where available from APIs and website provided metadata. The CRM service provides sentiment and opinion mining data using proprietary NLP algorithms built into their platform. Estimates of relative post impact and reach are also provided but were not considered in this study. Due to filtering limitations of the web crawling service, the dataset was exported for data cleaning using the Brandwatch Consumer Research API (Brandwatch, 2021a) and the related API client for Python (Github, 2021). Cleaning was performed using Python (Python Software Foundation) with the Pandas package (Team, 2020) to remove as many non-relevant posts as possible without removing relevant posts. This was done by manually reading individual partial body text data and identifying words that were common to several irrelevant posts. These words were then compiled into lists, which were in turn concatenated into regular expressions (regex) with the Python 'or' operator ('|') separating words in the list. The data from the CRM service were read into a Pandas 'DataFrame', and the partial body text column was searched for the constructed regular expression, effectively creating a filter for the identified terms. Each resulting subset of data to be removed was manually checked for relevant posts before removal. A similar process was used to remove posts from domains that were consistently irrelevant. Additional non-relevant posts were removed manually.

2.2 | Data analysis

Trends in the social media data were compared to the significant events in the outbreak as identified by the CDFA vND website (CDFA, 2021). Peaks in post volume stratified by week were identified and compared to these significant events. Weeks were chosen as the time unit since post volume by month was found not to be granular enough to identify trends around individual significant dates, and post volume by day did not take into account the effect of 'retweets' and similar post magnification mechanisms in which multiple users discuss similar topics over several days. Post volume peaks occurring before ('leading') or after ('lagging') the CDFA announcement of a significant event along with the time delta between the peak and the event were recorded.

Sentiment analysis and opinion mining were performed with tools built into the CRM platform. Each post was classified as either positive, negative or neutral. Emotion qualifiers based on Ekman and Friesen's six emotional identifiers (anger, fear, disgust, joy, surprise and sadness) (Ekman & Friesen, 1971) were also assigned to posts where the post text was sufficiently long for analysis and emotion was detected. Both sentiment and emotional qualifiers were determined in the CRM platform by comparing words and phrases in the post title and body to rule-based classifiers established by the CRM company (pre-defined words and phrases that are known to convey positive or negative sentiment and/or emotion) and a proprietary NLP-based ML algorithm.(Brandwatch, 2021b; Hayes et al., 2021) .

Descriptive statistics were performed using Microsoft Excel (V16.0.13901.20400). Posts were grouped by 'post type' to determine the average of each post type over time and the total number of each type of post during the study period. Post type groups were also used to determine proportions of negative, neutral and positive sentiment by platform type. Impressions, a metric unique to Twitter posts that describes the total number times a Tweet has been viewed by Twitter users (Twitter, 2021) was recorded through the Twitter API. Twitter accounts were ranked by the average number of impressions per post. Average monthly site visitors were recorded for other sites using metadata. Figures were generated using the CRM platform.

3 | RESULTS

3.1 | Web crawling collection of data and processed results

The 2018–2020 ND outbreak in California was first detected by the California Animal Health and Food Safety (CAHFS) Laboratory in May of 2018, but the web crawling service employed has a backlog limit of 2 years. Therefore, the first 2 months of the outbreak were not captured using the web crawling service. A total of 2498 posts from 862 unique authors were returned using the Boolean string search. Searches with the same or equivalent terms in Spanish returned less than 150 relevant results, and the majority of those were from news reports. For these reasons, Spanish language postings were not included in this

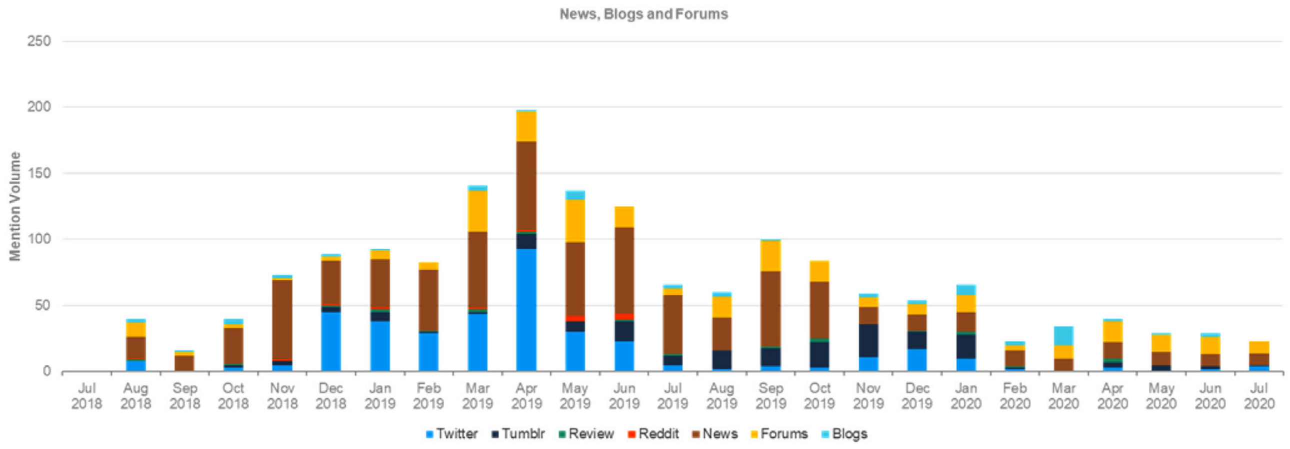

FIGURE 1 vND posts between 31 July 2018 and 23 July 2020 on various websites and social media platforms using a Boolean logic-based algorithm

study. Initial analysis of data collected without specifically excluding mentions of chicken pox, influenza, and COVID-19 resulted in multiple false positives. After further data cleaning, 1503 relevant posts were used for analysis. Post types were classified as blog, forum, news, Reddit, Tumblr and Twitter posts based on the post domain and the layout of the website (Figure 1). The news category accounted for the majority of posts (50.3%, 756 posts) and includes trade journals and scholarly articles in addition to traditional Internet news sources. Twitter posts were the second most prevalent post type (25.2%, 378 posts) followed by forum posts (18.4%, 277 posts), blog posts (2.8%, 42 posts), Tumblr posts (2.6%, 39 posts) and Reddit posts (0.7%, 11 posts). Seventy-four percent of forum posts were on backyardchickens.com, a popular hobbyist poultry keeper site.

Post types varied during the outbreak as shown in Figure 1. Initially, from July 2018 to November 2018, posts were primarily categorized as 'news' with a total of 151 news posts which accounted for 77.04% of all posts during this period (average of 30.2 news posts per month). News sites posted between one and four posts per month. Of the 196 posts discussing ND during this period, most of this subset clustered around late July near the beginning of the outbreak. The number of Twitter posts began to rival – and often exceed – news posts from December 2018 through June of 2019 (Figure 1). During this period, the average number of Twitter posts about ND increased from 3.6 per month to 41.7 per month, with a peak of 94 Twitter posts in April 2019 (Figure 1). News posts about ND also increased to an average of 44.8 posts per month. Forum activity also saw an increase to an average of 14 posts per month, primarily from backyardchickens.com (Figure 1). News and forum posts continued to be elevated from July through October of 2019, and then steadily declined to under 25 combined posts per month as the CDFA quarantine was lifted in June of 2020 (Figure 1).

Figure 2 shows the volume of online conversation during the study period with peaks in the collected data marked 'A' through 'F'. Peak 'F' is followed by a sustained increase in post volume compared to the two-year median weekly mention volume (11 mentions per week). This peak coincides with the first detected cases in Riverside County, and the majority of the posts are local news stories and retweets of a tweet linking an article on the outbreak (Figure 2). This peak sees the first negative sentiment reaction to mandatory depopulation reported by the Press Enterprise, a local newspaper in the Inland Empire (data not shown). Peak 'C' is primarily driven by several retweets, including one by the American Veterinary Medical Association, of an article posted by all Southern California News Group (SCNG) Newspaper Twitter accounts titled 'Chicken-killing Newcastle disease prompts widespread quarantines in Southern California'. Retweets of this news story continued into peak 'B', and were followed shortly after by another SCNG-wide news story 'Chicken-killing Newcastle disease prompts euthanasia orders for parts of Chino'. The words 'Chicken-killing' and 'euthanasia' caused negative overall sentiment during these peak weeks. Peak 'A' is primarily driven by retweets of an article by The Counter titled 'Backyard chickens hit hard by a long-gone, extremely contagious disease'. The first news reports and tweets of backyard bird owners taking issue with euthanasia efforts by CDFA occur during peak 'D', with words and phrases including 'protest', 'chicken slaughter', 'beg', and 'killing everyone's birds' driving a sharp spike in negative sentiment. The increase post volume at peak 'E' was primarily due to an article about designer chicken coops with a reference to ND published by 29 Tribune Publishing Company newspapers. This peak is also the first time during the ND outbreak that sentiment was significantly more positive than negative.

Of the websites that hosted posts in the dataset, yahoo.com reported the most monthly visitors at 1,992,603,000 per month followed by reddit.com (1,863,897,000), twitter.com (1,414,667,000) and tumblr.com (290,801,000). The most prominent BYP focused website, backyardchickens.com, hosted 205 posts during the outbreak and reported an average of 684,000 monthly visitors. While all websites included in this provided self-reported average monthly visitors, only Twitter provided data on the number of views for individual posts through the Twitter API impressions metric. During the outbreak, Twitter posts on average had 14380.9 impressions, and the median impressions was 2409. The Twitter post with the largest number of impression (236,817) was made by the Mercury News Twitter account

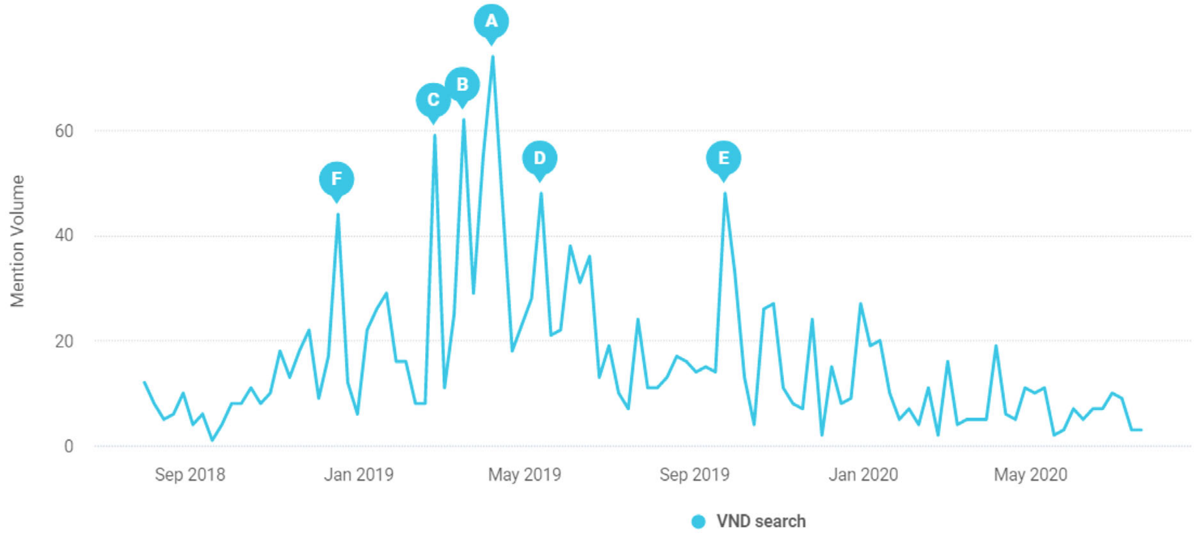

FIGURE 2 vND-based Twitter posts (not shares or views) collected between 31 July 2018 and 23 July 2020. Peaks C, B, A and E were associated with high negative sentiment (Figure3). The backyard poultry advocacy group Save our Birds (SOB) was formed 1 week after peak A. Peak 'F' coincides with the first detected cases of vND in Riverside County. Peak 'C' is followed by a sustained increase in post volume compared to the 2-year median weekly mention volume of 11 mentions per week

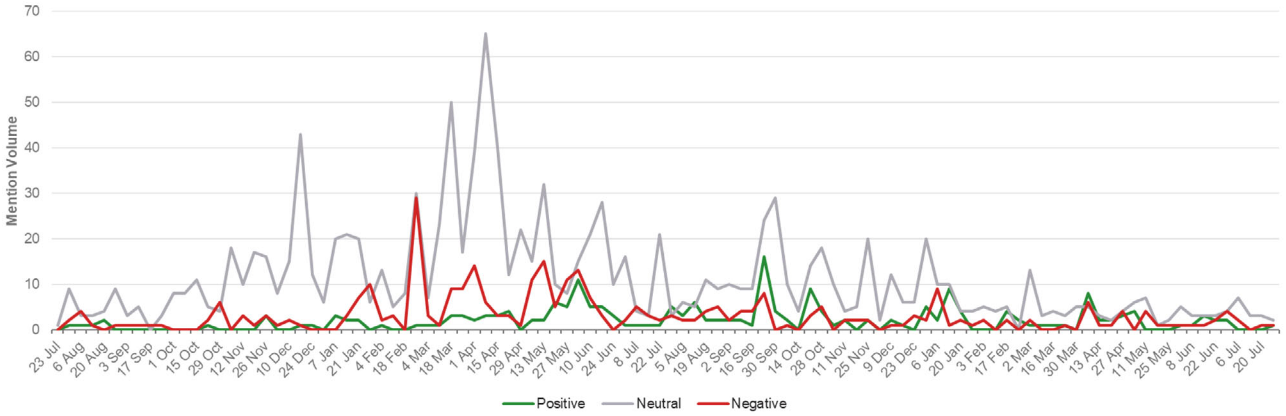

FIGURE 3 Positive, negative and neutral sentiment of posts per week from 31 July 2018 to 23 July 2020

on 5/31/2019 sharing a news article hosted on its website reporting 'more than 1.1 million chickens have been euthanized at 10 egg farms…' Among the Twitter accounts with the most average impressions were several Twitter accounts for daily newspapers that are members of the Southern California News Group, including the Press Enterprise (8 posts, 103,900 average impressions), the Sun (8 posts, 42479.3 average impressions), the San Gabriel Valley (SGV) Tribune (7 posts, 38,078 average impression), the Orange County Register (5 posts, 210248.4 average impressions) and the Daily Breeze (5 posts, 43843.4 average impressions). The same newspapers self-reported 456,000 (Press Enterprise, pe.com), 107,000 (the Sun, sunherald.com), 156,000 (SGV Tribune, sgvtribune.com), 2,716,000 (Orange County Register, ocregister.com) and 288,000 (Daily Breeze, dailybreeze.com) monthly visitors. Other prominent Twitter accounts included AVMAvets (8 posts, 49497.3 average impressions), NYFarmer (2 posts, 41475.5 average impressions) and UCANR (5 posts, 10844.6 average impressions).

The number of negative sentiment posts greatly exceeded the number of positive sentiment posts during most weeks (Figure 3), though the majority of posts during almost every week were classified as neutral (1059 posts). Twenty-point eight percent of posts were classified as negative (312 total) and 8.8% of the posts were classified as positive (132 total). Forum posts were the source of the largest number of negative posts (118 posts, 42.6% of total forums posts) and positive posts (71 posts, 25.63%). Of the post categories recorded, over 50% blog, news, Tumblr and Twitter posts were classified as neutral. Only forum and Reddit posts had more than 50% of total posts classified as either positive or negative. All post categories had more total negative posts than positive posts. Tumblr (17:1) and Twitter (7:1) had the highest proportion of negative posts to positive posts.

4 | DISCUSSION

The results of this study demonstrate how valuable insights on the public's understanding of an outbreak event can be gained from monitoring relevant social media posts. Interestingly the subject and sentiment of posts differed most based on platform. For example, niche forums, like backyardchickens.com typically contained questions about the ND outbreak and information for other poultry owners including preventative measures updates from the CDFA and the state veterinarian. These posts tended not to lean extremely towards positive or negative sentiment. This result suggests that information is being effectively disseminated to many backyard stakeholders. However, forums like backyardchickens.com (684,000 total website visits per month) have relatively few users compared to large social media platforms like Twitter (1.4 billion total website visits per month), though only a small fraction of those Twitter site visitors are discussing or reading about poultry and ND. Individual Twitter accounts that do discuss ND, particularly local news agency accounts such as The Press-Enterprise based in Riverside, CA, can have over 100,000 Twitter users that view each post they make. Consequently, posts and reposts by these large social media accounts have the capacity to disproportionally influence public perception. Hence even though information may be effectively disseminated to many backyard stakeholders via targeted websites, that information and sentiment is being overwhelmed by highly trafficked social media. To this point, several months of Tweets that were classified as negative or shared news stories that displayed a largely negative sentiment towards the outbreak and how it was being handled preceded the formation of the 'Save our Birds' Facebook group, whose messaging was largely focused on breaking the ND quarantine and anti-depopulation due to fear and a perceived lack of transparency from regulators (Facebook, 2021) . In past human disease outbreaks, similar social media discourse has garnered distrust of scientific expertise and public health agency responses (Laurent-Simpson & Lo, 2019) as well as spread misinformation (Cinelli et al., 2020),which in turn have made effective outbreak mitigation challenging. While providing information and recommendations to stakeholders continues to be a vital step in stemming the spread of outbreaks, including future outbreaks of ND, the proliferation of online information sharing requires public health agencies and extension specialists to take a more active role in ensuring the public supports their efforts. This includes both proactive transparency when communicating prevention measures to the public as well as reactive responses to negative posts and misinformation.

Although the data captured on social media provide a relatively new source of information for stakeholder responders at the local, state, federal and university level it is important to recognize that biases such as age and race occur across different social media platforms (Mislove et al., 2011; Ruths & Pfeffer, 2014). Acknowledging these biases and attempting to correct for them with proper study design will likely require access to underlying user data. Other limitations include our lack of data for the first 2 months of the study. Using different Boolean strings associated with clinical signs associated with vND it would have been interesting to see if there was any social media 'chatter' consistent with vND associated clinical signs that would allow for the exploration of this Enterprise CRM tool for a purpose similar to GFT (Kandula & Shaman, 2019). Finally, as noted in the methods section, Facebook and Instagram were not included in this study. The commercial web crawling service used only provides access to posts on or related to Facebook and Instagram accounts managed by the service user and authenticated through the web crawling service. Since this study focused on activity in existing groups and platforms, Facebook and Instagram data could not be collected. Future studies, could utilize tools like Crowdtange (crowdtangle.com) offer some tools to analyze Facebook, but they are linked to a specific registered Facebook account as opposed to an entire social media platform and hence would only provide a relatively biased sliver of the actual content. The Facebook API itself does offer access to all publicly available content on the platform, but private groups are only accessible if the user that has authenticated with the API is already a member.

Enterprise CRM tools such as the one used in this study offer a tool to monitor and provide generalized analysis of social media and web data. These tools also offer the ability to link extension or public health agency social media accounts to directly address misinformation or negative posts when detected. However, some significant limitations were noted in this CRM service including the web crawlers limited ability to perform sentiment analysis and its inability to track classified postings including Craigslist and Facebook Marketplace. All NLP analyses and analyses describing the engagement of the general public with posts that were included in the results from the CRM tool used proprietary algorithms, thus there is a lack of transparency in exactly how these parameters were calculated. Recent studies have also shown the accuracy of ML-based sentiment analysis and emotional classification used by the CRM tool is poor compared to human determination of sentiment an emotion, but comparable to other commonly used lexical methods (Hayes et al., 2021). Since the software is designed as a marketing tool and relies on metadata provided by crawled websites for calculations, the results suggest a bias towards providing metrics more consistently for websites with higher traffic. Additionally, the CRM tool is optimized for brand management as opposed to general surveillance. It is possible that the low signal to noise ratio (1.51 on topic posts) when searching for terms that can be used in multiple contexts (e.g. chicken in reference to political candidates, fast food etc.) is due to the CRM tool not being optimized for this use case. Ultimately, the fundamental difference in mission between the CRM tool and epidemiological surveillance of social media for ND means that while the CRM is useful for preliminary establishment of primary centers of conversation and general NLP tasks, the CRM tool and tools like it are likely not viable for long term use in academic and public health application. Once the presence of online conversation is established, further targeted web crawling and using purpose-built NLP packages in R or Python for 'beyond polarity' sentiment analysis and web crawling results that include context may provide additional insight and, in terms of sentiment analysis, accuracy and reproducibility. Furthermore, the ability to compare the number of users that posts have reached across platforms in a meaningful way beyond average website visitors per month (i.e. the number of people reading an individual article or post) would provide better understanding of the most effective platforms for information dissemination and monitoring. Other services, such as SimilarWeb (SimilarWeb, 2021), can provide more insight into non-social media site visitor information, but these data are still typically not stratified by individual web page. Comparing the number of reply posts to forum posts to the number of social media post comments (and/or retweets in the case of Twitter) would better describe the relative amount of interaction on posts of each platform. Since the CRM software employed here does not provide these data, additional web crawling software would need to be developed.

The social involvement aspect of disease is often neglected in disease control and epidemiological investigations. In outbreaks of disease in BYP populations, this area needs more focus since the social aspect of behavior change and culture play a significant role in outbreak mitigation. Using web crawling tools to monitor social media and other web data for content related to specific disease outbreaks and analyzing the sentiment of these posts is a novel and effective method of engaging with stakeholders during disease outbreaks. Currently there is a lack of understanding regarding the relationship between social media, other web data and infectious disease outbreaks in animals. For diseases like ND in urban areas, extension-based stakeholders could better understand overall sentiment, baseline knowledge and accuracy of information to better target outreach and extension messaging spatiotemporally. Despite the shortcomings of the CRM service employed in this study, the methods described allow for improved insights about knowledge and public perception, and they allow for direct social media response to posts that address mitigation efforts negatively. From an extension perspective, these types of tools could be used to better understand what topics need further outreach efforts to the general population and to stakeholders such as BYP owners and other known avian species which are potential carriers of ND and are commonly found in Southern California including psittacines and racing and roller pigeons. Furthermore, if the methodology described is repeated for future outbreaks, trends may emerge that will allow extension and regulatory stakeholders to proactively plan and improve social engagement strategies. Therefore, the ability to leverage these types of tools should be considered in future outbreaks with the goal of complimenting current extension efforts with the general public.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ETHICAL STATEMENT

No samples were collected from animals and no surveys were gathered from human subjects for this study.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

ORCID

Maurice Pitesky https://orcid.org/0000-0003-0084-6404

REFERENCES

Alamoodi, A. H., Zaidan, B. B., Zaidan, A. A., Albahri, O. S., Mohammed, K. I., Malik, R. Q., Almahdi, E. M., Chyad, M. A., Tareq, Z., Albahri, A. S., Hameed, H., & Alaa, M. (2020). Sentiment analysis and its applications in fighting COVID-19 and infectious diseases: A systematic review. *Expert Systems with Applications*, *167*, 114155.

Asur, S. & Huberman, B. A. (2010). Predicting the future with social media. Paper presented at the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology; Toronto, Canada.

Brandwatch. (2021a). Getting started with Brandwatch. Retrieved from https://developers.brandwatch.com/docs/getting-started

Brandwatch. (2021b). Sentiment analysis. Retrieved from https://www.brandwatch.com/wp-content/uploads/brandwatch/Brandwatch-Sentiment-Analysis.pdf

Broniatowski, D. A., Paul, M. J., & Dredze, M. (2013). National and local influenza surveillance through Twitter: An analysis of the 2012–2013 influenza epidemic. *Plos One*, *8*(12), e83672.

CDFA. (2021). Virulent Newcastle disease. Retrieved from https://www.cdfa.ca.gov/ahfss/Animal_Health/Newcastle_Disease_Info.html

Cinelli, M., Quattrociocchi, W., Galeazzi, A., Valensise, C. M., Brugnoli, E., Schmidt, A. L., Zola, P., Zollo, F., & Scala, A. (2020). The COVID-19 social media infodemic. *Scientific Reports*, *10*(1), 1–10.

Dimitrov, K. M., Abolnik, C., Afonso, C. L., Albina, E., Bahl, J., Berg, M., Briand, F. -X., Brown, I. H., Choi, K. -S., Chvala, I., Diel, D. G., Durr, P. A., Ferreira, H. L., Fusaro, A., Gil, P., Goujgoulova, G. V., Grund, C., Hicks, J. T., Joannis, T. M., … Wong, F. Y. K. (2019). Updated unified phylogenetic classification system and revised nomenclature for Newcastle disease virus. *Infection, Genetics and Evolution*, *74*, 103917.

Dimitrov, K. M., Ferreira, H. L., Pantin-Jackwood, M. J., Taylor, T. L., Goraichuk, I. V., Crossley, B. M., Killian, M. L., Bergeson, N. H., Torchetti, M. K., Afonso, C. L., & Suarez, D. L. (2019). Pathogenicity and transmission of virulent Newcastle disease virus from the 2018–2019 California outbreak and related viruses in young and adult chickens. *Virology*, *531*, 203–218.

Ekman, P., & Friesen, W. V. (1971). Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology*, *17*(2), 124.

Elkhoraibi, C., Blatchford, R. A., Pitesky, M. E., & Mench, J. A. (2014). Backyard chickens in the United States: A survey of flock owners. *Poultry Science*, *93*(11), 2920–2931.

Facebook. (2021). SOB save our birds. Retrieved from https://www.facebook.com/groups/SOBSaveOurBirds/about

Garber, L., Hill, G., Rodriguez, J., Gregory, G., & Voelker, L. (2007). Non-commercial poultry industries: Surveys of backyard and gamefowl breeder flocks in the United States. *Preventive Veterinary Medicine*, *80*(2–3), 120–128.

Github. (2021). Python client library for the brandwatch consumer research API. Retrieved from https://github.com/BrandwatchLtd/bcr-api

Hayes, J. L., Britt, B. C., Evans, W., Rush, S. W., Towery, N. A., & Adamson, A. C. (2021). Can social media listening platforms' artificial intelligence be trusted? Examining the accuracy of Crimson Hexagon's (Now Brandwatch Consumer Research's) AI-driven analyses. *Journal of Advertising*, *50*(1), 81–91.

Hidaka, C., Soda, K., Nomura, F., Kashiwabara, Y., Ito, H., & Ito, T. (2021). The chicken-derived velogenic Newcastle disease virus can acquire high pathogenicity in domestic ducks via serial passaging. *Avian Pathology*, 1–12. Advance online publication.

Jain, V. K., & Kumar, S. (2018). Effective surveillance and predictive mapping of mosquito-borne diseases using social media. *Journal of Computational Science*, *25*, 406–415.

Kandula, S., & Shaman, J. (2019). Reappraising the utility of Google flu trends. *PLoS Computational Biology*, *15*(8), e1007258.

Kennedy, H., & Moss, G. (2015). Known or knowing publics? Social media data mining and the question of public agency. *Big Data & Society*, *2*(2), 2053951715611145.

López, R., Tejada, J., & Thelwall, M. (2012). Spanish sentistrength as a tool for opinion mining peruvian facebook and twitter. *Artificial Intelligence Driven Solutions to Business and Engineering Problems*, 82.

Laurent-Simpson, A., & Lo, C. C. (2019). Risk society online: Zika virus, social media and distrust in the Centers for Disease Control and Prevention. *Sociology of health & illness*, *41*(7), 1270–1288.

Mislove, A., Lehmann, S., Ahn, Y., Onnela, J. & Rosenquist, J. (2011). Understanding the demographics of Twitter users. ICWSM'11: Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media; Barcelona, Spain.

Othman, M. K. & Danuri, M. S. N. M. (2016). Proposed conceptual framework of dengue active surveillance system (DASS) in Malaysia. Paper presented at the 2016 International Conference on Information and Communication Technology (ICICTM); Kuala Lumpur, Malaysia.

Pew Research Center. (2019a) Internet/broadband fact sheet. Retrieved from https://www.pewresearch.org/internet/fact-sheet/internet-broadband/

Pew Research Center. (2019b) Social media fact sheet. Retrieved from https://www.pewresearch.org/internet/fact-sheet/social-media/

Preis, T., & Moat, H. S. (2014). Adaptive nowcasting of influenza outbreaks using Google searches. *Royal Society Open Science*, *1*(2), 140095.

Python Software Foundation. Python Language Reference, v. A. a. Retrieved from http://www.python.org

Rue, C. A., Susta, L., Cornax, I., Brown, C. C., Kapczynski, D. R., Suarez, D. L., King, D. J., Miller, P. J., & Afonso, C. L. (2011). Virulent Newcastle disease virus elicits a strong innate immune response in chickens. *Journal of General Virology*, *92*(4), 931–939.

Ruths, D., & Pfeffer, J. (2014). Social media for large studies of behavior. *Science*, *346*(6213), 1063–1064.

Shimshoni, Y., Efron, N., & Matias, Y. (2009). On the predictability of search trends. Google, Israel Labs.

SimilarWeb. (2021). SimilarWeb. Retrieved from https://www.similarweb.com/

Singh, R., Singh, R., & Bhatia, A. (2018). Sentiment analysis using Machine Learning technique to predict outbreaks and epidemics. *International Journal of Advanced Science and Research*, *3*(2), 19–24.

Team, T. P. D. (2020). pandas-dev/pandas: Pandas. Retrieved from https://doi.org/10.5281/zenodo.3509134

Tumasjan, A., Sprenger, T., Sandner, P. & Welpe, I. (2010). Predicting elections with twitter: What 140 characters reveal about political sentiment. Paper presented at the Proceedings of the International AAAI Conference on Web and Social Media; Washington D.C. USA.

Twitter. (2021). Engagement API. Retrieved from https://developer.twitter.com/en/docs/twitter-api/enterprise/engagement-api/overview

USDA-APHIS (1978). Eradication of exotic newcastle disease in Southern California 1971–1974, APHIS-91-34 Technical Report.

USDA-APHIS. (2018). *Epidemiological analyses of virulent newcastle disease in backyard birds in California, December 2018*. Retrieved from https://www.aphis.usda.gov/animal_health/downloads/animal_diseases/ai/epi-analyses-vnd-in-backyard-birds-in-california-dec.pdf

USDA-ARS. (2016). Exotic and emerging avian viral diseases research: Newcastle disease. Retrieved from https://www.ars.usda.gov/southeast-area/athens-ga/us-national-poultry-research-center/exotic-emerging-avian-viral-diseases-research/docs/newcastle-disease/

USDA. (2013). *Urban chicken ownership in four U.S. cities*. Retrieved from https://www.aphis.usda.gov/animal_health/nahms/poultry/downloads/poultry10/Poultry10_is_Human-chicken.pdf

How to cite this article: Gendreau, J., Ramsubeik, S., & Pitesky, M. (2022). Web crawling of social media and related web platforms to analyze backyard poultry owners responses to the 2018–2020 Newcastle Disease (ND) outbreak in Southern California. *Transboundary and Emerging Diseases*, 1–8. https://doi.org/10.1111/tbed.14454