Social Network Analysis and Mining (2023) 13:121
https://doi.org/10.1007/s13278-023-01131-7

ORIGINAL ARTICLE

Text mining of veterinary forums for epidemiological surveillance supplementation

Samuel Munaf[1,2] · Kevin Swingler[1] · Franz Brülisauer[3] · Anthony O'Hare[1] · George Gunn[2] · Aaron Reeves[4]

Received: 12 February 2023 / Revised: 7 July 2023 / Accepted: 8 September 2023 / Published online: 25 September 2023

Abstract
Web scraping and text mining are popular computer science methods deployed by public health researchers to augment traditional epidemiological surveillance. However, within veterinary disease surveillance, such techniques are still in the early stages of development and have not yet been fully utilised. This study presents an exploration into the utility of incorporating internet-based data to better understand smallholder farming communities within the UK, by using online text extraction and the subsequent mining of this data. Web scraping of the livestock fora was conducted, with text mining and topic modelling of data in search of common themes, words, and topics found within the text, in addition to temporal analysis through anomaly detection. Results revealed that some of the key areas in pig forum discussions included identification, age management, containment, and breeding and weaning practices. In discussions about poultry farming, a preference for free-range practices was expressed, along with a focus on feeding practices and addressing red mite infestations. Temporal topic modelling revealed an increase in conversations around pig containment and care, as well as poultry equipment maintenance. Moreover, anomaly detection was discovered to be particularly effective for tracking unusual spikes in forum activity, which may suggest new concerns or trends. Internet data can be a very effective tool in aiding traditional veterinary surveillance methods, but the requirement for human validation of said data is crucial. This opens avenues of research via the incorporation of other dynamic social media data, namely Twitter, in addition to location analysis to highlight spatial patterns.

Keywords Veterinary epidemiology · Infodemiology · Infoveillance · Smallholding · Web scraping · Text mining · Topic modelling · Anomaly detection

1 Introduction

Rapid advancements in social media platforms have provided the public with a vast array of topical information and the ability to communicate globally instantaneously (Park et al. 2021). Furthermore, computational methods related to text analysis and data mining of such internet data have promptly increased in both applicability and usefulness across a wide array of domains. Such methods have been adopted extensively in the public health field, especially regarding utilising patient triage comments to predict the progression of illness/disease (Tulloch et al. 2019). However, within the veterinary domain, its application remains rather scarce.

Forum-based intelligence is still relatively new even amongst human public health, with researchers adopting such methods to understand the contents of pregnancy and health forums (Id et al. 2020). In the veterinary domain, little to no literature exists to conduct similar analysis (Dórea et al. 2019), and this study looks to explore the potential benefits of using publicly available data sources as a means of understanding the areas of conversation amongst the farming population. Latent Dirichlet allocation (LDA) was applied for the automatic labelling of the data, for the purposes of gaining a better comprehension of the topics discussed.

✉ Samuel Munaf
a.s.munaf@stir.ac.uk

[1] Division of Computing Science and Mathematics, University of Stirling, Stirling, UK

[2] Centre for Epidemiology and Planetary Health, Department of Veterinary and Animal Sciences, Northern Faculty, Scotland's Rural College (SRUC), Inverness, UK

[3] SRUC Veterinary Services, Scotland's Rural College (SRUC), Inverness, UK

[4] Centre for Applied Public Health Research, RTI International, Raleigh, NC, USA

1.1 Social media as a potential surveillance tool

Over the past two decades, research indicates substantial paradigm shifts in epidemiological disease surveillance due to the rapid expansion of the technological enterprise, in the form of the internet and social media (Mavragani and Ochoa 2019). The public health domain has been able to reap the benefits of increased quantities of health research derived from the internet boom, by subsequently creating progressively innovative epidemiological models (Mavragani and Ochoa 2018). Moreover, the internet has been utilised to assist in the determination of disease outbreaks, monitor the proliferation of infectious disease, and appraise outbreaks in the population. The methods in which communication tools are used in traditional public health surveillance have also been augmented with the aid of internet surveys, forums, and social media polls, proving to be more time and cost-efficient than the conventional approaches of telephone/mail/face-to-face surveys (Young et al. 2018). The term "social media" has various sources such as Twitter, Facebook/Meta, Instagram, Quora, and other forums for public engagement whereby these platforms can all be employed to facilitate the collection of passive data for analysis (Gittelman et al. 2015).

The study of sentiment analysis involves analysing people's feelings and opinions towards a subject through computational methods (Nandwani and Verma 2021). Systematic reviews have highlighted that the rise of digitalisation has led to an immense growth in user-generated content on the internet, which includes people's opinions on various topics (Ligthart et al. 2021). Furthermore, the authors purported that the use of sentiment analysis enables tracking of public attitudes towards a specific entity, providing actionable knowledge for understanding, explaining, and predicting social phenomena.

Peak detection is one method for detecting events from social media. It works by identifying unusual user behaviour in a particular location and time frame (Comito et al. 2017). Users are likely to pay attention to events that cause these deviations. The two primary phases of peak detection methods are feature extraction and peak identification. Space–time features are extracted from social data during feature extraction. These include the amount of posts and users, as well as sentiment polarity within a specified region and time frame. Time series are used to model these features, showing the changes in social activity over time in various regions. The paper by Comito et al. contributes to the literature on event detection from social media by proposing a peak detection method that can capture relevant events from geo-tagged data (Comito et al. 2017).

1.2 Smallholdings in the UK

Within the UK, owners of commercial/large-scale farms are required to record and share animal health data with the respective governmental agencies (Correia-Gomes et al. 2017). These data include farm capacity, species type, and livestock movements; thus, granting a comprehensive picture of animals at risk, yet these data are not available for public consumption. Furthermore, commercial holdings are subject to regular veterinary supervision and thereby contribute to national scanning surveillance systems (UKSF 2019). In contrast, little is known about the behaviours and activities of smallholder/backyard farmers, and the data surrounding them are minimal. There are many more premises with smaller numbers of livestock than in larger farms (DEFRA 2023). Whilst pig holdings of any size are obliged by law to be registered and report animal movements including births and death, backyard flocks may not be recorded. Although voluntary registration is encouraged, flocks with fewer than fifty birds are not required to be registered (Extrapolation of Poultry Smallholding Data Report 2020). Omission of backyard poultry from governmental databases has an impact on control of avian influenza, and there have been calls to address this surveillance oversight within this demographic (Correia-Gomes et al. 2017).

Recent outbreaks of highly pathogenic avian influenza (HPAI) from 2019 within the UK demonstrate the risks posed to such poultry producers (Hill et al. 2019), as well as the potential risks that they pose to larger scale commercial producers. Social media, advice for and search engines like Google, are all potential methods used by smallholders to acquire information regarding animal husbandry, business, and health (Moreno-Ortiz et al. 2021). These tools have been used by researchers in recent years to reach this secluded cohort to provide advice on biosecurity, transportation, and adequate disposal of livestock (Correia-Gomes et al. 2017). Despite an uptake in recent efforts to integrate social media as a supplementary surveillance tool within public health, little to no research has been conducted to evaluate the efficacy within livestock systems.

1.3 Where do farmers get their information from?

Smallholders and backyard hobbyist are progressively shifting to digital information to supplement their livestock and biosecurity needs (Correia-Gomes et al. 2017). Contemporary research, in the form of questionnaires, has corroborated such findings and highlights online communities immersed within both forums and social

media alike. Governmental agencies such as the Animal and Plant Health agency (APHA) and the Department for Environment, Food, and Rural Affairs (DEFRA) oversee the regulation of livestock within the UK, in addition to disseminating information related to disease outbreaks and biosecurity measures for all livestock owners (APHA 2015).

Research indicates that farmers have historically relied on "agricultural experts" when seeking information regarding legislation, practices, biosecurity measures, and public health guidance (Rust et al. 2022). As social media became more prevalent amongst these communities, questions have been raised with regard to the sources of information deemed reliable, what type of information is disseminated, and how this is perceived through the communities. It is evident that particularly influential users within these communities are deemed as the go-to source for new information, and trust is lessening amongst academics and governmental agencies (Rust et al. 2022). Peer networks seem to be the most utilised source for knowledge exchange, with forums and social media groups becoming the central hub to mediate these exchanges. This adds complexity to the current understanding of guideline adherence and perceptions to public health measures by the farming communities, as information is severely affected by the subjective bias of the influential users, before trickling down to the individual farmers.

1.4 Text mining applications

The increases in livestock-based data have created opportunities to adopt machine learning tools on the textual information to derive insights and expand on the current body of knowledge. Text mining and natural language processing (NLP) offer opportunities to decompose both the syntax and semantic elements of sentences to find patterns within seemingly unstructured text (McGarry and McDonald 2017).

The previous work has highlighted the benefits of text mining methods on online health platforms, such as forums and blogs, to generate keyword frequencies related to public health. An extension to this analysis is the adopting of topic modelling (TMO) as a method of generating clusters of associated text segments or "topics" from the unstructured data. Topic models classify related words with analogous meanings using vectors of topic distributions amongst documents and word distributions of undiscovered topics (Park et al. 2021).

TMO is recognised as a powerful methodology in the field of text mining due to its performance and efficiency when processing copious amounts of text. It yields the ability to mitigate the common issues associated with word frequencies, namely infrequent/sparse words, synonyms, the existence of many possible meanings for words/phrases, and semantic hierarchical compositions (Doan et al. 2019). In continuation, a popular algorithm for performing TMO is latent Dirichlet allocation (LDA), which has proven to be an effective method in public health research, particularly when the data are derived from either health blogs or forums (Alessa and Faezipour 2018).

1.5 Study aim

This paper addresses the gap in existing veterinary research by combining the fields of computer science with veterinary epidemiology to supplement traditional surveillance methods, which are often time-lagged and cumbersome. Contemporary primary data via questionnaires have highlighted the use of social media platforms by smallholders to seek advice and engagement with other smallholders, therefore, creating an opportunity for researchers to utilise such data for the enhancement of livestock disease surveillance (Correia-Gomes et al. 2017).

We look to explore the effectiveness topic modelling to smallholding pig and poultry forums to build a greater level of epidemiological intelligence related to animal husbandry, biosecurity, locations, opinions, and attitudes.

1.6 Main contributions and research novelty

This study fills an interdisciplinary research gap by combining veterinary epidemiology and social media mining. The research introduces a data-oriented approach to examine the topics and discussions surrounding livestock health amongst small-scale farmers. The lack of the literature reinforces that such methods have yet to be fully tested within livestock animal data, and we believe that an opportunity to use social media as a platform to build passive livestock intelligence can be explored. This innovation is significant in veterinary epidemiology as it gathers real-world insights from grassroots level, which are usually overlooked in conventional research, thus enhancing the contextual understanding of animal health trends, challenges, and community responses.

2 Methods

2.1 Data extraction

The methodology for data extraction is shown in Fig. 1, with the workflow being performed for both subsections (poultry and pigs) simultaneously. R programming language through the RStudio integrated development environment was applied for the data extraction through web scraping in the package with RVest. This was done in conjunction with the CSS (cascading style sheet) selector tool in Google Chrome, which allows HTML tags within each page to be derived. Furthermore, the TM and TidyText package in R

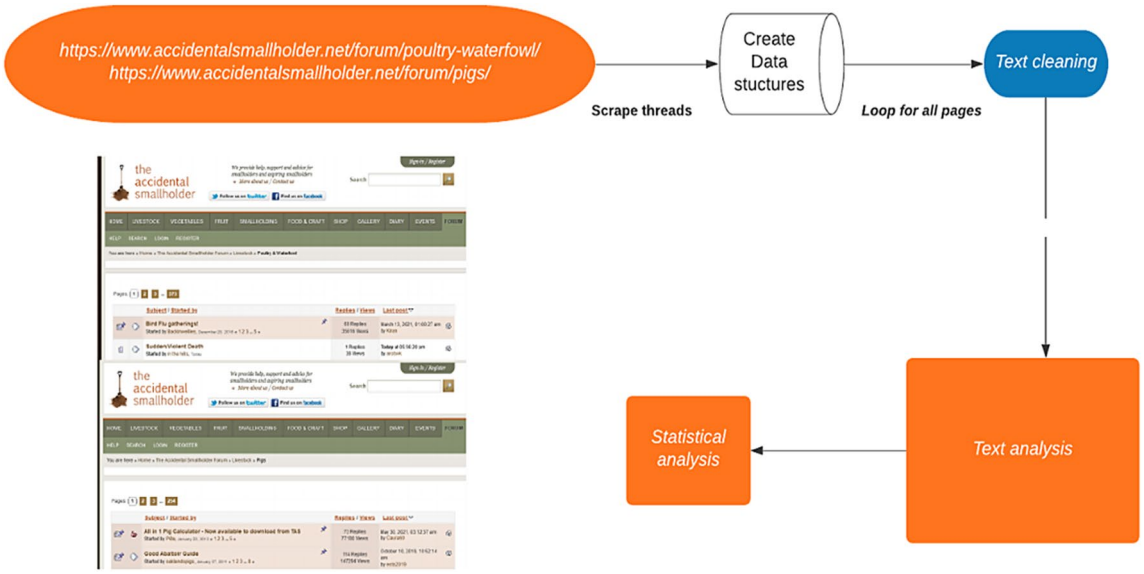

Fig. 1 Data extraction and analysis overview

and the Natural Language Toolkit (NLTK) in Python were used for text analysis and outputs section.

Web scraping the forums began with downloading individual user posts, each of which is stored as a single row in the final database. Individual posts are submitted by the user (who has their own username, and some optional details on display, e.g. location and job title), alongside a date–time stamp on the thread. The forum is an online community whereby smallholders and backyard livestock keepers can engage with each other and seek advice, tips, and useful information pertaining to their animals. It is publicly available, and no membership is needed to access the information within these forums.

The process of extracting the forums was split into two sections; firstly, each individual forum page was scraped from both the poultry and waterfowl subsection and the pig section from www.accidentalsmallholder.net. This included extracting the date of the post, the title, URL link for each thread, and engagement data such as likes and replies. This was iterated for all pages up to 100 (due to computational speed and not to inundate the website with excessive requests) and was replicated for both livestock subsections.

By selecting the time frame of 2017–2022, we strategically captured significant temporal variations and events that had an impact on the smallholding community. With a duration of 6 years, this time frame is adequate to perceive year-over-year changes and trends. Additionally, this time frame includes critical occurrences like the recent avian flu epidemic within the UK, that directly impacted poultry farming and the COVID-19 pandemic, which had far-reaching consequences that affected different aspects of small-scale farming such as supply chains and market demands.

This produced 281 URL links resulting from all the threads in the pig forum. The poultry data were substantially larger and produced 775 URL links, which may possibly be explained by the greater number of poultry keepers in this community and the recent avian flu outbreaks.

The second stage of the extraction involved using the URL links extracted to derive the entire discussion within each forum post. This is computationally expensive; hence, timeouts within the code were utilised as to not overwhelm the website with massive requests. Furthermore, this was iterated for each page within each forum post, in addition to all the URL links obtained earlier. A total of 4191 unique rows were produced in the pig forum and 5425 from the poultry forum.

This website was selected as the primary data source because it is considered the foremost forum for smallholdings in the UK. The website includes an extensive range of topics, such as animal husbandry, biosecurity, entrepreneurship, and general livestock management. At the time of the research, more than 57,000 active participants were registered on the forum. The platform is an invaluable repository for extracting information that is smallholder-centric and aggregating passive intelligence due to its vast and engaged community. Forums offer less noise than Twitter, as users discuss livestock-specific queries instead of general conversations or opinions on non-smallholding matters.

2.2 Data analysis

2.2.1 Pre-processing

Figure 2 displays the pre-processing procedure of the collected data, traversing through the text cleaning process by

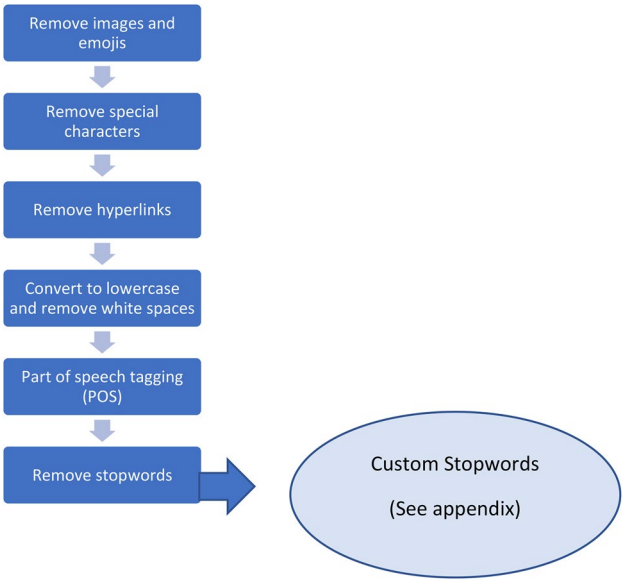

Fig. 2 Pre-processing procedure

cleaning all the noise found within the unstructured text. This includes removing any images and non-character symbols, namely emojis. In addition, we deleted the punctuation, hyperlinks, and white spaces generated from these reductions. Hyperlinks found within the text were also removed as they add no benefit to our study, as they mainly contained links to products for sale (e.g. second-hand farm equipment). Common stopwords in the NLTK package were removed, in addition to our own custom stopwords (see Appendix) which were deemed unrelated to the research questions.

2.2.2 Text mining and topic modelling

Word frequencies and bigrams were generated to visualise the common terms and word pairings amongst both forums.

LDA is a probabilistic framework to model the topic structures of the text, which calculates the probabilities of a word relating to a unique topic. It approximates the posterior distribution of the Bayesian probability model, which establishes the percentages of topic configurations in documents and the word configurations of the topics. The denotation of the topic can be calculated by the words with the largest probabilities for that topic (Park et al. 2021).

The optimum number of topics was calculated through the coherence score, which determines the similarities amongst the words with high probabilities for each topic. The ideal model has a low perplexity score, indicating better predictive accuracy, and a high coherence score, reflecting the semantic meaningfulness and logical consistency of its output (Zvornicanin 2021). The model's confidence in its predictions is reflected by low perplexity, whereas high coherence implies that the generated topics are easily interpretable and closely related.

2.2.3 Temporal analysis

2.2.3.1 Time-series topic modelling Subsequently, the dataset was separated into time bins (e.g. semi-annually) to facilitate a time-based analysis. LDA was employed within each bin to find out topic probabilities. This generative probabilistic model enabled identification of words with a statistical association, indicating a common topic. The frequency of these topics was examined across different time periods to detect temporal patterns and changes. Capturing trends and changes in discussions is crucial for understanding the dynamics of the subject matter, and time-series topic modelling could be more insightful than static topic modelling as it enables analysis of how topics evolve over time.

2.2.3.2 Anomaly detection The use of anomaly detection allowed for the identification of unusual temporal patterns in the time-series data. We applied an Isolation Forest algorithm in order to identify anomalies, which has proven to be effective in handling high-dimensional datasets by identifying isolated data points (Lesouple et al. 2021). The identification of sudden changes or spikes in topic prevalence can suggest important events or themes during specific time periods.

2.3 Computational requirements

Data extraction and analysis in the methodology section require computationally intensive tasks, such as web scraping, text processing, and topic modelling. Ensuring efficiency and feasibility requires addressing computational resource requirements and model run times. Each pig and poultry forum required almost 3 h to extract data due to time-out requests in the for loop. On an Intel Core i7 × 64 laptop, the modelling phase (calculating optimum topic numbers and running the topic models combined) needed approximately 20 min to complete.

Memory allocation and bandwidth are critical when web scraping, as it involves processing thousands of URL links and retrieving substantial amounts of data. For large datasets, it is recommended to use a multi-core processor to handle CPU-intensive tasks efficiently, such as text preprocessing and analysis using packages such as TM, TidyText, and NLTK, in order to improve performance. LDA topic modelling, as a probabilistic method, can be computationally intensive and may require substantial memory and CPU resources, particularly when optimising model parameters like the number of topics. Optimal performance can be achieved with a machine that has a powerful CPU and abundant RAM. Employing a GPU can further increase the

speed of LDA computations. Moreover, memory usage and parallel processing may be necessary for optimum results when integrating anomaly detection algorithms, such as Isolation Forest, into temporal analysis. Lastly, Amazon Web Services (AWS) or Azure are cloud computing services that are ideal for efficient handling of computational demands and scalability.

3 Results

3.1 Word frequencies—Pig

Using word frequencies and bigrams, we conducted an exploratory analysis of pig and poultry forums to uncover common topics and themes discussed in the forums.

The word cloud depicted in Fig. 3 illustrates the frequency distribution of words in the pig forum, where the size of a word in the image correlates to its frequency in the corpus. Terms such as rare breed, slap mark, and electric fence, which are easily identifiable, are further elaborated in the bigram analysis.

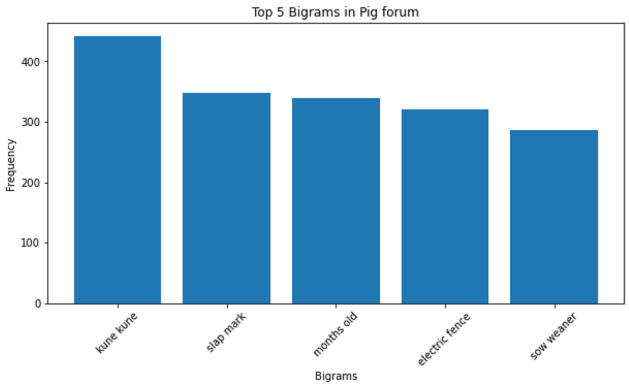

Fig. 4 Pig bigram

Figure 4 represents the results from the bigram analysis, with the top 5 word pairings being highlighted.

"Kune kune" is a bigram that identifies a domestic pig breed, characterised by small size and docility (Amalraj et al. 2018). The bigram's high frequency of 442 occurrences suggests that smallholder pig farmers might prefer these types of pigs due to their manageable size and

Fig. 3 Pig word cloud

temperament, which makes them ideal for small-scale farming.

Furthermore, the term "slap mark" appears 348 times and refers to the pig's tattoo for identification. A slap mark in pig farming is composed of a distinctive blend of numbers and letters. Identification and tracking of pigs may be crucial for smallholder farmers, possibly for regulatory compliance or management purposes, as indicated by the prominence of this bigram (Agricultural and Rural Economy Directorate 2021).

With 339 instances, the bigram "months old" most likely pertains to the age of pigs. Age, which affects breeding, weaning, and market readiness, heavily influences pig farming (RSPCA 2022). The age of pigs is a frequent topic of discussion amongst those who are trying to optimise growth, ensure health, and decide when to sell or breed them.

The term "electric fence" is mentioned 320 times, highlighting the significance of containment in pig farming. Electric fences are a common choice for smallholder pig farmers as they are easy to install and effectively control pig movements (EFSA Panel on Animal Health and Welfare (AHAW) et al. 2021). Effective fencing solutions are crucial in managing and containing pigs, as indicated by the frequency of this bigram. This prevents losses and liability issues that could result from escapes.

Finally, "Sow weaner" was mentioned 287 times and signifies a young pig that has been weaned from a female pig called a sow. Discussions about breeding and weaning practices may be related to the term. Based on the frequency of this bigram, it can be inferred that breeding and weaning piglets are fundamental operational factors for smallholder pig farmers, which may impact herd productivity and financial feasibility (Harlizius et al. 2020).

In summary, the bigrams and word cloud shed light on the areas of concern and interest amongst smallholder pig farmers, such as choice of breed, animal identification, age management, containment strategies, and breeding and weaning practices. These areas are likely to have significant implications for the operational efficiency, productivity, and sustainability of smallholder pig farming. This sets the foundation for the terminology to be expected in the results of the upcoming topic modelling analysis.

3.2 Word frequencies—Poultry

Figure 5's poultry word cloud displays that, like the pig forum analysis, animal husbandry was the most prevalent subject. Figure 6 delves deeper into this through bigram analysis.

The bigram "free range" ranks first on the list with 1457 instances. Free range is a poultry farming method that allows birds to roam freely outdoors instead of being confined to enclosures. The rise in consumer interest in animal welfare and the belief that free-range products are more natural have made this practice more popular (Bray and Ankeny 2017).

A type of feed called layer pellets is likely what the bigram "Layers-Pellets", appearing 687 times, refers to. To ensure high-quality eggs, laying hens are given special feed called layer pellets (Sakomura et al. 2019). The term "layers" is commonly used to describe hens that are specifically bred for egg-laying. The conversations about layers and pellets are most likely to be focussed on the dietary requirements of laying hens and effective feeding techniques to optimise egg output.

"Red mite" is most likely used to describe the poultry red mite, a prevalent parasite that invades chickens, with 659 recorded instances. Poultry infested with these mites can experience irritation, anaemia, or death (Temple et al. 2020). Preventing and treating red mite infestations is a common topic in poultry farming discussions.

The term "Hatching eggs" is mentioned 441 times, indicating that it refers to conversations about eggs that are meant to be incubated and hatched into chicks. Some topics related to hatching eggs are proper incubation techniques, maintaining the optimal temperature and humidity, and caring for the chicks once they hatch.

Finally, 398 instances of the bigram "nest boxes" which are the designated places within a coop where laying hens can lay their eggs. Topics regarding nest boxes may cover their structure, ideal materials, quantity per hens, and sanitation techniques. Safe and comfortable nest boxes are essential in ensuring that hens lay eggs in a conducive environment, which ultimately affects the quality of the eggs (Hartcher and Jones 2017).

The poultry community's areas of interest and focus are highlighted through the bigrams. Various topics are covered from ethical farming practices, nutrition, bird health, egg production, and housing practicalities.

3.3 Topic modelling

The optimum number of topics was determined to be 4 from the coherence score, and the results from the LDA model are displayed in Table 1. The top terms were visualised, and a manually ascribed topic name was given to each topic number based on what we deemed to be a feasible name, which captures the essence of the top 10 terms.

Topic 0 was deemed to be related to containment and care, based on the prevalence of terms such as “electric fence”. Topic 1 pertained to feeding, as “feed, food, and feeding” where prominent, topic 2 seemed to be focussed on slaughter processes and housing considerations, and topic 3 centred around identification and processing.

Overlap between terms is possible as they may be relevant to multiple aspects of pig and poultry farming. Both topic 0 (pig containment and care) and topic 1 (feeding) contain

Fig. 5 Word cloud—Poultry

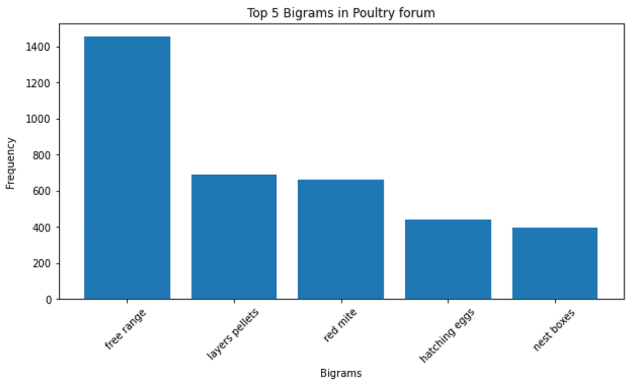

Fig. 6 Poultry bigrams

the word "feed" as it is a crucial factor in pig care and feeding practices. Abattoirs are a crucial element in both topic 2 and topic 3 since they are involved in pig identification and processing as well as the slaughter process.

The poultry forum was deemed to have an optimum topic number of 14, which may have been caused by the significantly larger corpus, greater number of users, and generally a more diverse range of topics discussed pertaining to poultry keeping. In addition to the avian flu outbreak in 2021, which significantly affected the housing measures enforced upon poultry keepers within the UK, as the government attempted to quell the spread of the disease.

Table 2 shows that poultry forums cover more topics, including species-level discussions (i.e. waterfowl, chicks, and goslings). Similar to Table 1, overlapping of terms in poultry farming can occur due to their relevance to multiple aspects of the topic. The importance of egg production in various aspects of poultry farming, including breeding, hatching, and selling, explains why the term "eggs" is present in multiple topics. Likewise, because of their essential role in poultry farming, the term "hens" appears in multiple discussions on the subject.

3.4 Temporal topic modelling

A deeper dive into the dataset through temporal topic modelling can uncover fluctuations over time periods in relation to topic weights.

Table 1 Topic modelling results—Pigs

Topic number	Top terms	Assigned topic name
0	Keep electric fence well back weaners water feed way know	Pig containment and care
1	Feed sow piglets food well weeks around boar feeding keep	Feeding
2	Straw back meat feed abattoir well butcher trailer ark used	Slaughter processes and housing considerations
3	Meat tag slaughter tags abattoir know weaners back people mark	Pig identification and processing

Table 2 Topic modelling results—Poultry

Topic number	Top terms	Assigned topic name
0	Birds hens eggs old meat laying keep around cockerel lay	Egg-laying and breeding
1	Birds used hens well eggs feeder quote sand work old	Equipment and maintenance
2	Geese bit may water see keep back run birds quote	Waterfowl
3	Eggs hens broody chicks water ducks hen put used weeks	Hatching and raising chicks
4	Hens birds eggs water keep hen days well run coop	General poultry care
5	Eggs ducks hens old well house ones birds back water	Housing and shelter
6	Rats run hens house keep around rat birds fox mite	Pest and predator management
7	Hens feed hen well birds cockerel really quote see weeks	Nutrition and feed
8	Birds hens eggs breed last white well quote bit hen	Poultry breeds
9	Hens ducks water hen food keep feed old bit days	Ducks and waterfowl care
10	Birds eggs geese hens back ducks free anyone quote feed	Free-range management
11	Chicks well eggs old geese feed grass fine weeks water	Raising chicks and goslings
12	Eggs hatch goose geese nest incubator hen days keep egg	Incubation and hatching
13	Eggs sell egg selling duck birds people hens around keep	Egg production and selling

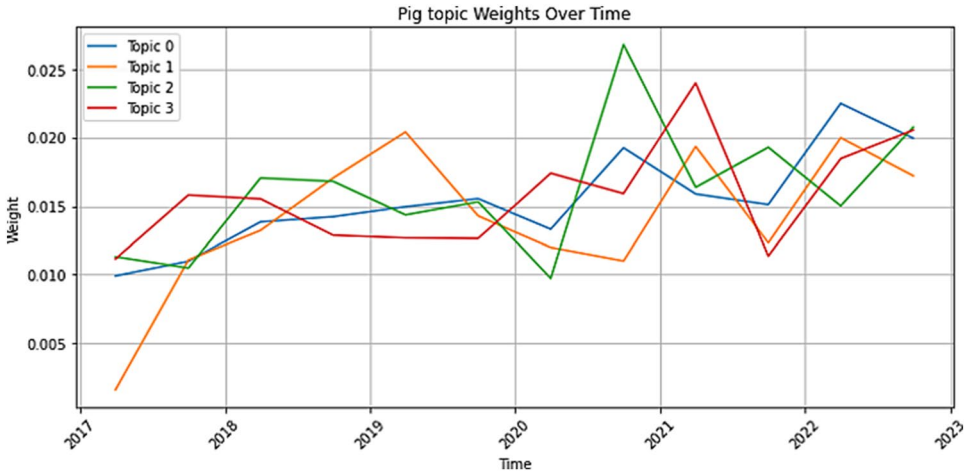

Fig. 7 Temporal topic model—Pigs

Figure 7 depicts changes in topic weights over time for the pig forum. Starting from topic 0, weight shows a general increase with minor fluctuations over time. At the beginning of 2017, the weight was low, but it reached its peak in March 2022. This indicates that the relevance or frequency of discussions pertaining to containment and care has generally increased over the given time frame. Conversely, topic 1 exhibited greater variation over time. It began at the lowest point in March 2017 and experienced a considerable rise by March 2019. Its weight decreased by September 2022 after some periods of decline and stabilisation. Feeding-related discussions seemed to have peaked in the middle of the timeline but have slightly decreased recently.

An interesting pattern emerged in topic 2, starting with moderate weight in March 2017, peaking in September 2020, and steadying towards the end of the timeline. We can infer from the results that there might have been a rise in discussions or relevance related to slaughter processes

around the 2020 peak. In March 2017, topic 3 started with a moderate weight and experienced notable fluctuations until it reached a significant peak in March 2021. This topic seemed to be the most variable in terms of weight change for discussions focused on identification and processing.

Figure 8 illustrates the changes in topic weights for the poultry dataset. The largest peaks were witnessed by topic 1 in mid-2019 (equipment and maintenance), topic 2 in early 2018 (Waterfowl), and topic 3 in mid-2021 (hatching and raising chicks). The majority of the topics all seem to be on the rise towards the end of 2022, which is in line with the extensive new poultry-keeping regulations enforced by the APHA as a result of the avian flu outbreak.

There could be multiple factors behind the changes in topic weights over this time period. These factors encompass seasonal practices, market demands, new technologies, regulatory alterations, diseases or veterinary methods, community engagement and education, external occurrences, and the surge in social media and information-sharing platforms.

3.5 Anomaly detection

Figure 9's results uncovered numerous anomalies, revealing dates with unusual pig livestock-related posts or discussions. Multiple high peaks were observed in June and July 2017, possibly indicating a seasonal trend or industry developments. In February 2020, there was another significant peak with 30 posts during the initial stages of the COVID-19 outbreak, which might be linked to the indirect consequences on pig farming. We also detected infrequent posts during some periods of low activity, possibly because of holidays or other events. These were more prevalent between 2017 and late 2019, with only a few low activity days occurring after 2020. Regular spikes in activity, with over 20 posts on various days across different years, may suggest a recurring event that drives increased discussions.

Similarly, Fig. 10 represents the results from the anomaly detection for the poultry dataset. This cohort witnessed greater variations in trends and a larger number of anomalies, than the pig cohort. The first quarter of 2017 witnessed the greatest number of anomalies, with a high number of posts possibly being explained due to the spring breeding season. There was a substantial increase in posts from March to May 2020, which could be attributed to the COVID-19 pandemic and an influx of newcomers seeking advice. Anomalies between February and March 2021 could have resulted from discussions on adapting or recovering from the pandemic or early spring-related conversations. Seasonal activities and new housing regulations in the poultry industry may be responsible for anomalies between April and June 2022.

Understanding these anomalies demands considering the bigger picture, including events related to public health and animal health. By comparing these anomalies with historical events, we can gain insights into the correlation between discussions and real-world events.

4 Discussion

This study provides new understanding of the concerns and priorities of small-scale farmers in the UK, through the analysis of forum data. Key themes in pig farming discussions were identified through bigram and topic modelling analysis. These themes include breed selection, animal identification, age management, containment strategies, and breeding and weaning practices. The frequent occurrence of the word pairing "Kune Kune" suggests that smallholders favour this breed of domestic pig, possibly because of its small size and gentle disposition. The importance of identification and containment in pig farming is reflected in terms like "slap mark" for pig identification tattoos and "electric fence".

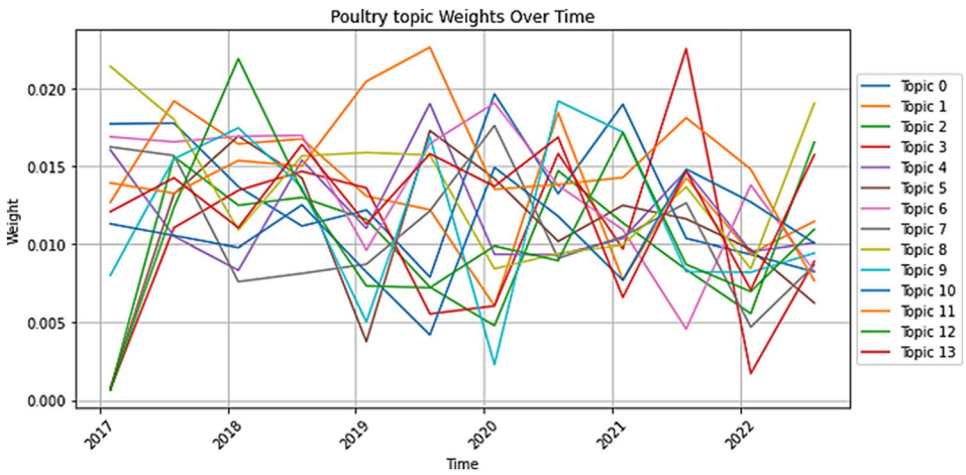

Fig. 8 Temporal topic model—Poultry

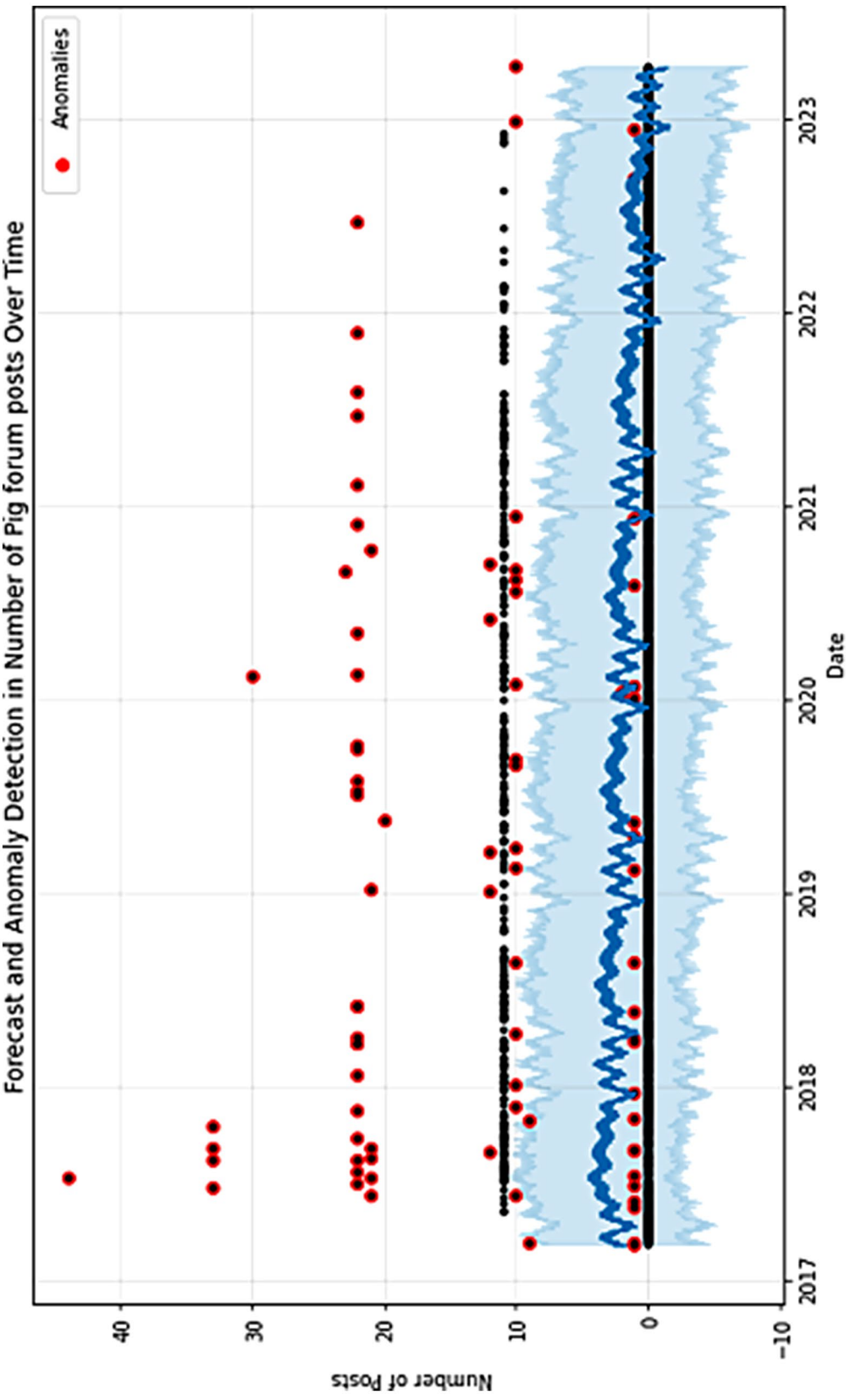

Fig. 9 Anomaly detection—Pigs

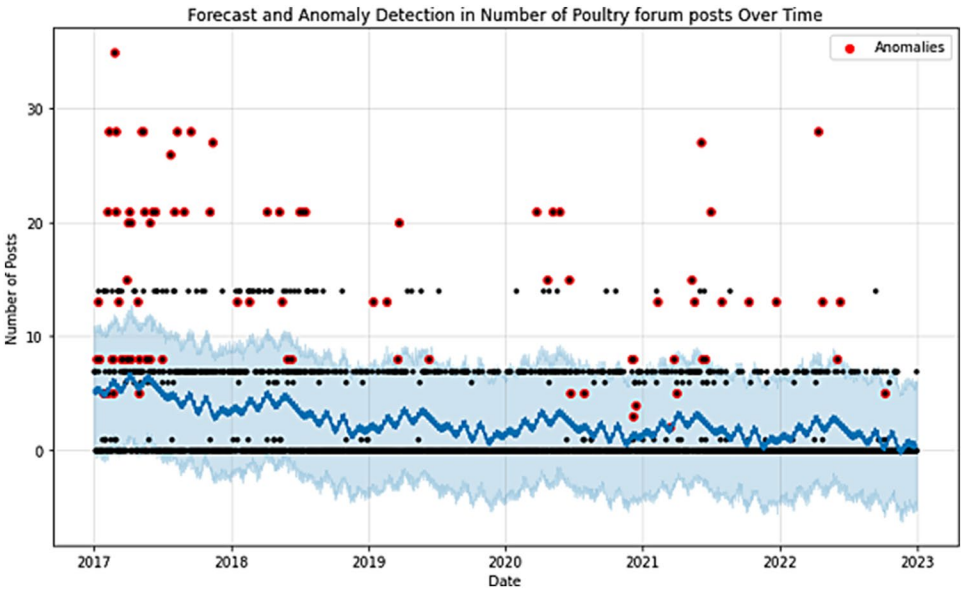

Fig. 10 Anomaly detection—Poultry

The poultry community talks focused on free-range methods, layer pellet nutrition, red mite infections, egg hatching, and nest boxes. The prevalence of the term "free range" as a bigram implies that poultry farming is highly focused on animal welfare and ethical farming practices. Discussions about layer pellets and red mites underscore the crucial role that nutrition, health, and parasite prevention play in optimising poultry production.

Whilst the application of topic modelling and temporal topic modelling within the veterinary domain still remains scarce, LDA applications, combined with outbreak detection methods, have been applied to identify disease in the UK dogs, and proven to be an accurate predictive tool (Noble et al. 2021). However, data on companion animals are far greater (200,000 + records) and more accurate than smallholding livestock; therefore, direct comparisons are difficult to make with our research in this context.

The results from our anomaly detection analysis need to be further built upon in order to create a feasible outbreak detection surveillance system, which can capture spikes in posting activity around certain topics. Early warning systems are made possible through detecting anomalies in user activity within a given time period, as witnessed in the significant increase in research during the COVID-19 pandemic (Botz et al. 2022). Spikes in posting frequency may provide opportunities to respond to disease outbreaks sooner, as highlighted by research conducted in endemic disease surveillance (Eze et al. 2023). Outlier detection using thresholds can be an effective instrument for epidemiological disease surveillance; however, a more dynamic source of information, such as Twitter, may be the ideal data imputation platform as it has a larger user base and is able to generate greater online traffic.

A differentiation needs to be made between publicly available data, such as this forum, and the more refined/granular information only available to the government. This includes precise farm locations, demographics, and personal information. For the general public, APHA regularly updates their surveillance dashboards with information related to livestock disease incidence (APHA 2023). In the case of avian livestock, this dashboard incorporates veterinary diagnoses from non-commercial, hobby, and small-scale flocks of chickens. This is updated on a monthly basis, with the main clinical signs being displayed in a frequency chart corresponding to the confirmed diagnoses during that time period. Data can be further filtered by the age of the species, in addition to county level of location. The government will have further information regarding these dashboards, including the addresses of registered pig and poultry holdings.

4.1 Implications of findings and further research

The insights gained from forum data analysis can benefit various stakeholders. Firstly, policymakers and regulators can use these insights to craft policies that are more aligned with the needs of smallholder communities, such as simplifying compliance requirements. Secondly, APHA and DEFRA can provide knowledge and support to these communities in developing effective containment strategies by customising their services to their needs and interests. Finally, with these insights, farmers can now grasp the trends within their community and modify their practices accordingly, including the exploration of free-range poultry farming as a viable alternative.

The role of the government in this instance is to use this confirmed data, along with any passive data they collect, and

implement disease interventions and biosecurity measures if they deem the threat level to be a cause for concern. As previously mentioned, the validity of passive data collection through internet-based medium is still in question, however simply having this information as a supplementary tool can be used to bolster the surveillance tools currently in place. The literature clearly highlights the shift in the methods of which smallholders obtain their livestock information (Correia-Gomes and Sparks 2020), therefore augmenting the current dashboards to also include social media insights allows us to build a stronger intelligence repertoire.

Extensions of this analyses can be conducted on other sub-forums, including equine, cattle, and sheep. Supplementary work can be conducted examining the relationships between the users themselves, through the application of social network analysis. Additionally, similar to the work conducted by researchers regarding peak detection and sentiment analysis, these methods can also be incorporated as an extension of the work conducted in this study, to further bolster our findings (Comito et al. 2017).

4.2 Limitations

The limitations of conducting any form of web scraping and infoveillance from internet data are apparent in both selection bias and small sample sizes. The ability for a few influential users to control the narrative of these forums is something which is often overlooked, and the only method to mitigate these effects is through conducting network analysis. Forum data also may only represent a fraction of the community, as many may not participate online. In addition, despite topic modelling being an insightful technique, it may not represent the true essence of the discussions, and therefore, human expertise within the field of veterinary epidemiology is required for efficacy and quality checking of the results.

Furthermore, animal health information discussed in the fora relate to clinical signs as observed by their owners. Whilst this is valuable information, the veterinary input is missing. Moreover, many conditions require laboratory testing to be confirmed. In contrast, diagnoses published on APHA surveillance dashboards for a range of livestock have been generated following stringent diagnostic criteria.

5 Conclusion

This study intended to demonstrate the application of topic modelling algorithms on veterinary forum data, with the aims of being a gateway for further studies in the efficacy of combining data science techniques in the veterinary domain. Through the innovative use of temporal topic modelling, it captures the dynamic nature of discussions over time. Topic modelling results and the high frequency of bigrams suggest that smallholders are most concerned with regulatory compliance and day-to-day management, particularly in relation to specific breeds, animal identification, containment, and feeding practices. Another noteworthy point for consideration is the rise in discussions about free-range practices in poultry farming reflects a broader social trend towards ethical and sustainable farming practices.

The findings have implications for policy-making, extension services, and smallholder practices and highlight the potential of forum data as a valuable resource for understanding and supporting smallholder communities.

This paper has highlighted the implementation of one of the many methods available within the topic modelling field. LDA is amongst the most common applications within this field and has proven to be a successful tool when applied to public health data (Egger and Yu 2022). Contemporary models, namely Top2Vec and BERTopic, have shown to perform effectively with sparse, unstructured social media data, with the need of further research being crucial to compare performance amongst these various models. A full comparison of all these models applied to this data is beyond the scope of this study.

By reproducing the techniques used in public health and epidemiological studies for human health surveillance, this study creates a foundation for more in-depth research within livestock animals, as parallel studies within companion animal surveillance have proven to be effective (Noble et al. 2021). As the world continues to face challenges such as climate change, food security, and shifts in consumer preferences, understanding the dynamics within smallholding communities becomes increasingly vital. This research shows that harnessing the wealth of information available in online forums and employing sophisticated analytical techniques can provide meaningful insights that can contribute to the development of more sustainable and resilient farming systems.

Appendix

Custom stopwords list

Pigs –

'pigs','need','much','like','could','thanks','also','think','first','one','two','time','re','pig','hi', 'hello', 'want', 'obtain', 'look', 'hii','ive','got','use','number','get','would','day','one','good','year','thought','year'

Poultry –

'poultry','chicken','need','much','like','could','thanks','also','think','first','one','two','time','re','chickens','hi', 'hello', 'want', 'obtain', 'look', 'hii','ive','got','use','number','get','would','day','one','good','year','thought','year'

Acknowledgements Not applicable.

Author contributions All authors contributed to the study conception and design. Material preparation, data collection, and analysis were performed by SM. The first draft of the manuscript was written by SM, and all authors commented on the previous versions of the manuscript. All authors read and approved the final manuscript.

Funding This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Availability of data and materials All data were scraped from the public domain and can be requested by contacting the corresponding author.

Declarations

Competing interests The authors declare no competing interests.

Ethics statement All data were anonymised prior to analysis. No user identifiable data were scraped, and all texts were aggregated and analysed together; hence, no individual can be identified from the results. Ethical approval was granted by the University of Stirling's General University Ethics Panel (GUEP) and conformed to the research integrity policies. All methods were carried out in accordance with relevant guidelines and regulations surrounding social media scraping and analysis provided by the UK Research and Innovation (UKRI). Further information can be found here: https://www.ukri.org/councils/esrc/guidance-for-applicants/research-ethics-guidance/internet-mediated-research/.

Informed consent The need for informed consent was waived by the University of Stirling's General University Ethics Panel (GUEP).

Consent for publication Not applicable.

References

Agricultural and Rural economy directorate (2021) Livestock identification and traceability: guidance. http://www.gov.scot/publications/livestock-identification-and-traceability-guidance/

Alessa A, Faezipour M (2018) A review of influenza detection and prediction through social networking sites. Theoret Biol Med Modell. https://doi.org/10.1186/s12976-017-0074-5

Amalraj A, Matthijs A, Schoos A, Neirynck W, De Coensel E, Bernaerdt E, Van Soom A, Maes D (2018) Health and management of hobby pigs: A review. VLAAMS DIERGENEESKUNDIG TIJDSCHRIFT 87(6):6

APHA (2015) GB emerging threats quarterly report: pig diseases pig: disease surveillance reports, 22(1), 0–16

APHA (2023) Avian dashboard. https://public.tableau.com/app/profile/siu.apha/viz/AvianDashboard/Overview

Botz J, Wang D, Lambert N, Wagner N, Génin M, Thommes E, Madan S, Coudeville L, Fröhlich H (2022) Modeling approaches for early warning and monitoring of pandemic situations as well as decision support. Front Public Health. https://doi.org/10.3389/fpubh.2022.994949

Bray HJ, Ankeny RA (2017) Happy chickens lay tastier eggs: motivations for buying free-range eggs in Australia. Anthrozoös 30(2):213–226. https://doi.org/10.1080/08927936.2017.1310986

Comito C, Falcone D, Talia D (2017) a peak detection method to uncover events from social media. IEEE Int Conf Data Sci Adv Anal (DSAA) 2017:459–467. https://doi.org/10.1109/DSAA.2017.69

Correia-Gomes C, Sparks N (2020) Exploring the attitudes of backyard poultry keepers to health and biosecurity. Prevent Vet Med 174:104812. https://doi.org/10.1016/j.prevetmed.2019.104812

Correia-Gomes C, Henry MK, Auty HK, Gunn GJ (2017) Exploring the role of small-scale livestock keepers for national biosecurity—the pig case. Prev Vet Med 145:7–15. https://doi.org/10.1016/j.prevetmed.2017.06.005

DEFRA (2023) Agricultural facts: England regional profiles. https://www.gov.uk/government/statistics/agricultural-facts-england-regional-profiles/agricultural-facts-england-regional-profiles-guidance-note

Doan S, Yang EW, Tilak SS, Li PW, Zisook DS, Torii M (2019) Extracting health-related causality from twitter messages using natural language processing. BMC Med Informat Decision Mak. https://doi.org/10.1186/s12911-019-0785-0

Dórea FC, Vial F, Hammar K, Lindberg A, Lambrix P, Blomqvist E, Revie CW (2019) Drivers for the development of an Animal Health Surveillance Ontology (AHSO). Prevent Vet Med 166:39–48. https://doi.org/10.1016/j.prevetmed.2019.03.002

EFSA Panel on Animal Health and Welfare (AHAW), Nielsen SS, Alvarez J, Bicout DJ, Calistri P, Canali E, Drewe JA, Garin-Bastuji B, Gonzales Rojas JL, Herskin M, Miranda Chueca MÁ, Michel V, Padalino B, Pasquali P, Roberts HC, Sihvonen LH, Spoolder H, Stahl K, Velarde A, Gortázar Schmidt C (2021) African swine fever and outdoor farming of pigs. EFSA J 19(6):e06639. https://doi.org/10.2903/j.efsa.2021.6639

Egger R, Yu J (2022) A topic modeling comparison between LDA, NMF, Top2Vec, and BERTopic to demystify twitter posts. Front Sociol 7:886498. https://doi.org/10.3389/fsoc.2022.886498

Extrapolation of Poultry Smallholding Data Report. (2020). 5.

Eze PU, Geard N, Mueller I, Chades I (2023) Anomaly detection in endemic disease surveillance data using machine learning techniques. Healthcare 11(13):13. https://doi.org/10.3390/healthcare11131896

Gittelman S, Lange V, Gotway Crawford CA, Okoro CA, Lieb E, Dhingra SS, Trimarchi E (2015) A new source of data for public health surveillance: Facebook likes. J Med Internet Res. https://doi.org/10.2196/jmir.3970

Harlizius B, Mathur P, Knol EF (2020) Breeding for resilience: new opportunities in a modern pig breeding program. J Anim Sci 98(Supplement 1):S150–S154. https://doi.org/10.1093/jas/skaa141

Hartcher KM, Jones B (2017) The welfare of layer hens in cage and cage-free housing systems. World's Poult Sci J 73(4):767–782. https://doi.org/10.1017/S0043933917000812

Hill A, Gillings S, Alexander B, Adam B, Andrew CB, Snow L, Ashton A, Charles B, Irvine RM (2019) Quantifying the spatial risk of Avian Influenza introduction into British poultry by wild birds. Sci Rep. https://doi.org/10.1038/s41598-019-56165-9

Id AW, Davoudi A, Weissenbacher D, Choi R, Id KOC, Cummings H, Gonzalez-hernandez G (2020) Pregnancy and health in the

age of the Internet: a content analysis of online " birth club " forums. PloS one. https://doi.org/10.1371/journal.pone.0230947
Lesouple J, Baudoin C, Spigai M, Tourneret J-Y (2021) Generalized isolation forest for anomaly detection. Pattern Recogn Lett 149:109–119. https://doi.org/10.1016/j.patrec.2021.05.022
Ligthart A, Catal C, Tekinerdogan B (2021) Systematic reviews in sentiment analysis: a tertiary study. Artif Intell Rev 54(7):4997–5053. https://doi.org/10.1007/s10462-021-09973-3
Mavragani A, Ochoa G (2018) Infoveillance of infectious diseases in USA: STDs, tuberculosis, and hepatitis. J Big Data. https://doi.org/10.1186/s40537-018-0140-9
Mavragani A, Ochoa G (2019) Google trends in infodemiology and infoveillance: methodology framework. J Med Internet Res. https://doi.org/10.2196/13439
McGarry K, McDonald S (2017) Computational methods for text mining user posts on a popular gaming forum for identifying user experience issues. In: HCI 2017: digital make believe - proceedings of the 31st international BCS human computer interaction conference, HCI 2017, 2017-July, 1–6. https://doi.org/10.14236/ewic/HCI2017.100
Moreno-Ortiz C, Peterson D, Collart A, Downey L, Seal S, Gallardo R (2021) Small farmers' use of social media and other channels for marketing their agricultural products. J Extension 59(4):1–8
Nandwani P, Verma R (2021) A review on sentiment analysis and emotion detection from text. Soc Netw Anal Min 11(1):81. https://doi.org/10.1007/s13278-021-00776-6
Noble P-JM, Appleton C, Radford AD, Nenadic G (2021) Using topic modelling for unsupervised annotation of electronic health records to identify an outbreak of disease in UK dogs. PLOS ONE 16(12):e0260402. https://doi.org/10.1371/journal.pone.0260402
Park S, Kim-knauss Y, Sim J, Sim J (2021) Leveraging text mining approach to identify what people want to know about mental disorders from online inquiry platforms. Front Public Health 9(October):1–9. https://doi.org/10.3389/fpubh.2021.759802
RSPCA (2022) Welfare of Pigs. https://www.rspca.org.uk/documents/1494939/7712578/FAD-Pigs-Information-Sheet-2022.pdf/9def23d9-c86f-e16f-39ed-9023b68924a6?t=1673619310960
Rust NA, Stankovics P, Jarvis RM, Morris-Trainor Z, de Vries JR, Ingram J, Mills J, Glikman JA, Parkinson J, Toth Z, Hansda R, McMorran R, Glass J, Reed MS (2022) Have farmers had enough of experts? Environ Manage 69(1):31–44. https://doi.org/10.1007/s00267-021-01546-y
Sakomura NK, Reis MDP, Ferreira NT, Gous RM (2019) Modeling egg production as a means of optimizing dietary nutrient contents for laying hens. Anim Front 9(2):45–51. https://doi.org/10.1093/af/vfz010
Temple D, Manteca X, Escribano D, Salas M, Mainau E, Zschiesche E, Petersen I, Dolz R, Thomas E (2020) Assessment of laying-bird welfare following acaricidal treatment of a commercial flock naturally infested with the poultry red mite (Dermanyssus gallinae). Plos One 15(11):e0241608. https://doi.org/10.1371/journal.pone.0241608
Tulloch JSP, Vivancos R, Christley RM, Radford AD, Warner JC (2019) X Mapping tweets to a known disease epidemiology; a case study of Lyme disease in the United Kingdom and Republic of Ireland. J Biomed Informatics: X 4(1):100060. https://doi.org/10.1016/j.yjbinx.2019.100060
UKSF (2019) The UK approach to animal health surveillance (p. 12). https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/869173/uksf-animal-health-surveillance.pdf
Young SD, Torrone EA, Urata J, Aral SO (2018) Using search engine data as a tool to predict syphilis. Epidemiology 29(4):574–578. https://doi.org/10.1097/EDE.0000000000000836
Zvornicanin E (2021) When coherence score is good or bad in topic modeling? Baeldung on computer science. https://www.baeldung.com/cs/topic-modeling-coherence-score