

Mammalian H5N1 cases predict online discourse volume and emotion during the 2024–2025 H5N1 outbreak

Luizza Dalla Pria

ln326@cornell.edu

Cornell University <https://orcid.org/0009-0000-8748-5271>

Katherine Koebel

Department of Population Medicine and Diagnostic Sciences, College of Veterinary Medicine, Cornell University
<https://orcid.org/0000-0003-4910-2131>

Amelia Nelson

Cornell University

Renata Ivanek

Department of Population Medicine and Diagnostic Sciences, College of Veterinary Medicine, Cornell University
<https://orcid.org/0000-0001-6348-4709>

Article

Keywords:

Posted Date: October 18th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11270170/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Additional Declarations: There is **NO** Competing Interest.

Abstract

The surge of highly pathogenic avian influenza A (H5N1) infection in animals across the United States since 2024 has renewed pandemic concern, making public perception a critical determinant of effective risk communication. We analyzed 42,957 Reddit comments (January 2024–December 2025) to characterize the patterns in online discourse during the outbreak. Topic modeling identified themes of public health, pandemic comparisons, and food safety. Sentiment analysis identified fear as the dominant emotional response, while anger was the most event-reactive, correlating with reported cases in cats, poultry, and wild mammals. Autoregressive Integrated Moving Average with Exogenous Regressors (ARIMAX) models with blocked cross-validation identified reported cases in domestic cats and dairy cattle as predictors of weekly Reddit engagement, reducing forecasting error by 31% over baseline. Mammalian cases thus were associated with early-warning signals for public reaction during the 2024–2025 outbreak, offering a one-week window for proactive risk communication.

Introduction

Sustained outbreaks of highly pathogenic avian influenza (HPAI), caused by influenza A (H5N1) viruses, are escalating concern over food security, zoonotic spillover, and pandemic risk (Garg et al., 2025). In recent years, the virus has been detected across multiple animal populations – wild birds, poultry, dairy cattle, and various domestic and wild mammals – reflecting its expanding host range. Although human infections remain relatively rare (Garg et al., 2025), the emergence of cases in new animal hosts (Mallapaty, 2024) and the persistence of outbreaks in agricultural settings have increased public attention. Understanding how the public perceives and responds to such threats is important for risk communication and policy development.

Social media platforms offer a real-time lens on public reactions to emerging infectious diseases including H5N1 (Wilson et al., 2021), and online discussions both reflect and shape public perceptions of risk during health crises (Van Bavel et al., 2020). Prior H5N1-specific work has examined misinformation dynamics on X/Twitter (Cooper et al., 2025), but limited attention has been given to how H5N1-related online discourse evolves alongside epidemiological signals from different animal hosts, economic indicators, and news article volume, each of which may shape public attention and interpretation of risk.

Reddit is a well-suited venue for observing how such signals shape public discourse. It organizes user discussions into topic-based communities called subreddits. Reddit's US-heavy traffic (~ 48% of the platform's global traffic; Duong, 2024) and the relatively low likelihood of non-US engagement with a spatially distant outbreak make the platform well-suited for analyzing public discourse around a US-based outbreak. With over 130 million daily active users worldwide (Reddit, Inc., 2026), its user base skews younger and more male than the US general population, and users of topic-focused subreddits are typically already attentive to their subject matter (Amaya et al., 2021). Reddit discussions have previously been used to characterize public priorities during the COVID-19 pandemic through longitudinal topic modeling (Stokes et al., 2020), demonstrating that such analyses can surface meaningful patterns in public engagement with health crises.

Our objective was to characterize H5N1-related Reddit comments and identify the signals that shape them. We: (i) characterized the main themes and narratives in Reddit comments using topic modeling; (ii) assessed emotional responses and risk perception through sentiment and toxicity analyses; (iii) examined how online engagement aligns over time with external signals – media coverage, food-price trends, and confirmed H5N1 cases in mammals and birds; (iv) modeled the predictive relationships between these signals and weekly Reddit comment volume using Autoregressive Integrated Moving Average with Exogenous Regressors (ARIMAX) time-series models with blocked

cross-validation; and (v) performed a mechanism analysis to assess relationships between key phrases and topics with emotions and comment volume.

Results

In total, 42,957 comments were collected and retained for analysis from three H5N1-focused communities – r/H5N1_AvianFlu, r/BirdFlu, and r/BirdFluPreps –posted between January 1, 2024 and December 31, 2025. The start of the study window was chosen to precede the first reported case of H5N1 in dairy cattle in the United States on March 25, 2024 (USDA APHIS, 2026b), which heightened public attention. The dataset spans periods of varying outbreak activity and media attention, providing a longitudinal view of public engagement and perception during ongoing H5N1 developments.

Main Themes and Narratives

To identify the main themes in Reddit comments, two complementary approaches were applied: bigram frequency analysis (Xue et al., 2020) and Latent Dirichlet Allocation (LDA) topic modeling (Jelodar et al., 2019). Bigram analysis examines the most frequent pairs of consecutive words in the text, allowing for the identification of commonly co-occurring terms and recurring phrases. In contrast, LDA is an unsupervised probabilistic model that groups words into topics based on patterns of co-occurrence across text assigning each topic, a set of representative terms. Together, these methods provide both surface-level insights into frequently used expressions and a structured representation of underlying thematic patterns within the dataset.

Analysis of frequent bigrams reveals recurring word pairings that highlight dominant discussion patterns in the dataset (Fig. 1). Predictably, the most frequent bigram by a wide margin is "avian flu," combining 'bird flu,' 'avian influenza,' and other disease-referring terms. It is followed by "raw milk," "public health," "dairy cattle," and "wild bird," capturing discussions of food safety, health systems, agricultural transmission, and wildlife reservoirs. Additional frequently occurring phrases include "human human" and "human case," reflecting attention to human-to-human transmission and individual confirmed human cases; "flu vaccine" and "seasonal flu," reflecting comparisons with seasonal influenza and discussion of vaccination; and "mortality rate," "fatality rate," and "test positive," indicating attention to disease severity and surveillance outcomes.

Table 1

LDA topic modeling of Reddit comments. Three dominant topics were identified through Latent Dirichlet Allocation (LDA) topic modeling of the Reddit comments in the study window. For each topic, the table lists the representative terms output by LDA and the corresponding thematic interpretation.

Main Representative Terms	Theme
"flu", "bird", "human", "nt", "infect", "case", "peopl", "vaccin", "like", "would"	Human infection cases, disease characteristics, and vaccination
nt", "peopl", "get", "go", "think", "would", "covid", "one", "know"	General public discussion and comparisons with prior viral outbreaks (particularly COVID-19)
"milk", "farm", "test", "cow", "viru", "h5n1", "anim", "infect", "dairi", "state"	Dairy cattle, agricultural surveillance, and state level monitoring of confirmed infections

The LDA analysis reveals three dominant topic structures (Table 1). Together, these patterns define three main themes in the Reddit discussion surrounding HPAI. These themes reveal a public actively reflecting on the disease and drawing comparisons with the recent COVID-19 pandemic. The third theme in particular engages with One Health–adjacent dimensions of the outbreak – dairy-cattle infection, agricultural surveillance, and state-level monitoring – even without explicitly naming the framework.

Sentiment Analysis

Emotional responses were assessed using the Text2Emotion library, which maps text to five predefined emotional categories – fear, sadness, anger, happiness, and surprise – based on the presence and frequency of emotion-related words (Py_Team, 2020). Each comment received a score from 0 to 1 for each emotion, representing the relative intensity of that emotion in the text.

Average Text2Emotion scores across the corpus reveal a clear hierarchy of affective intensity (Fig. 2): fear is the dominant emotional register, followed by sadness, with surprise, happiness, and anger substantially lower. Per-comment distributions (Supplementary Fig. 1) show fear is consistently present and often dominant, while other emotions appear less frequently but with occasional strong outliers. Toxicity was then assessed with ToxicBERT – a transformer-based classifier fine-tuned on annotated online comments to detect harmful, offensive, or disruptive language (Hanu & Unitary team, 2020) – with per-comment scores shown in Fig. 3.

The distribution of ToxicBERT toxicity scores is heavily concentrated at the low end of the 0-to-1 scale (Fig. 3): a primary mode sits at or very near 0, a smaller secondary cluster appears around 0.08–0.10. This bimodal pattern with relatively low toxicity scores indicates that the discourse in the three analyzed H5N1-focused subreddits is overwhelmingly civil, with very few comments approaching even mild levels of toxic language. This pattern likely reflects both genuine civility and the effects of subreddit moderation: comments removed by moderators for violating community rules were excluded from the corpus (see Methods), so highly toxic content caught prior to data collection is not represented in the distribution. A parallel analysis using the Perspective API (Jigsaw) confirmed the same pattern (Supplementary Fig. 2).

Aligning External Factors

To examine the external signals that may shape public engagement with the H5N1 outbreak (represented as the weekly Reddit discussion volume), seven weekly time series were assembled spanning the study window (January 2024 – December 2025): five animal-surveillance series (poultry, dairy cattle, cats, wild mammals and wild birds), retail egg prices, and media coverage. All series were aligned to a Monday-anchored weekly grid for comparison. Reddit comment volume was computed by summing weekly comment counts across the three H5N1-focused subreddits (r/H5N1_AvianFlu, r/birdflu, r/BirdFluPreps) over the study window. The Reddit series exhibits two distinct waves of engagement: the (i) first wave beginning in late March 2024 that climbs rapidly through the spring dairy spillover, and the (ii) second, larger wave through the winter 2024–2025 surge. Hereafter, we refer to these two waves as: (i) dairy cattle spillover (March–June 2024) and (ii) surge (October 2024 – February 2025). Over the two-year study window, weekly counts range from 1 to 2,682 comments per week (median = 219 across the 93 weeks of non-zero activity, excluding 12 weeks with zero activity).

The surveillance series were obtained from U.S. Department of Agriculture (USDA) Animal and Plant Health Inspection Service (APHIS), covering weekly reported confirmed HPAI cases in 4 APHIS databases: “commercial and backyard flocks” for poultry cases, “livestock” for dairy cattle cases, “mammals” for non-livestock mammal cases in domestic cats and various wild mammals, and “wild birds” for wild bird cases (USDA APHIS, 2026a). Reporting units differ by species per USDA APHIS convention: cattle and poultry cases are reported at the herd or flock (premises) level, whereas cases in domestic cats, wild mammals (foxes, mice, seals, raccoons, bears, and others), and wild birds are reported at the level of individual animals. A confirmed HPAI case was defined as a case of HPAI (H5N1 subtype) reported to and confirmed by USDA, based on samples submitted by producers, veterinarians, state animal-health officials, or wildlife-agency personnel following clinical suspicion or targeted surveillance. Across the 105-week study window, weekly reported case counts ranged from 0 to 13 for domestic cats, 0 to 49 for wild mammals, 0 to 125 for cattle herds (with non-zero values concentrated after the spring 2024 dairy spillover), 0 to 45 for poultry flocks, and 0 to

394 for wild birds. Among confirmed non-livestock mammal cases, domestic cats account for the largest single-species share with 154 reported cases (Supplementary Fig. 3).

Two contextual signals completed the variable set. Retail prices for U.S. Grade A large eggs were extracted from the USDA Economic Research Service Meat Price Spreads dataset (Vaiknoras, 2026), which provides monthly price observations and is available only through July 2025 and were aligned to weekly bins for analysis. Eggs were the only animal-based food category in the dataset to show substantial price movement across the study window (Supplementary Fig. 4); prices for the other tracked products — beef, pork, and chicken — remained essentially flat. Weekly egg prices ranged from \$2.52 to \$6.23 per dozen (from January 2024 to July 2025). Media coverage was quantified via the SerpAPI Google News interface: 4,892 unique English-language articles retrieved with 15 H5N1-related queries (see Methods), yielding weekly counts of 0 to 146 articles.

All evaluated external factors are visualized in Supplementary Fig. 5. Here we display the four variables retained by the ARIMAX analysis (described below): weekly Reddit comment volume (the response variable) and confirmed HPAI cases in domestic cats, dairy cattle, and wild mammals (the three exogenous predictors) (Fig. 4). To allow visual comparison across the heterogeneous units of the original series, each variable was $\log_{1p}$ -transformed (matching the transformation applied in the ARIMAX modeling), smoothed with a 4-week centered rolling mean to suppress short-window noise; min-max scaling for all factors in the model was tested, and equivalent results were seen (Supplementary Note 1). The shaded gray regions in Fig. 4 mark the two waves of Reddit engagement observed during the study window, during which all four highlighted variables reach local peaks.

ARIMAX Modeling

To evaluate the relationships between external signals and Reddit comments, we fit ARIMAX models with weekly log-transformed comment counts as the dependent variable and seven candidate exogenous variables (egg prices, news article volume, and confirmed H5N1 cases in poultry, cattle, domestic cats, wild mammals, and wild birds) at cross-correlation function (CCF)-identified lags. Per-variable in-sample results are reported in Supplementary Table 1.

Only domestic cats at lag 1 reached in-sample statistical significance ($\beta = +0.34$, 95% CI [0.09, 0.60], $p = 0.008$). Cattle and total non-livestock mammals (representing combined cat and wild mammal cases as reported in the APHIS “mammals” database) exhibited strong level correlations with Reddit volume (Spearman $\rho = 0.677$ and 0.407 respectively; both $p < 0.001$) but did not produce significant ARIMAX coefficients — reflecting shared long-run trends rather than week-to-week predictive structure. Egg prices, news volume, poultry alone, wild mammals alone, and wild birds showed neither significant Spearman correlations (all $p \leq 0.21$, all $p \geq 0.06$ for eggs and news; $\rho = -0.048$, $p = 0.628$ for wild birds) nor significant ARIMAX coefficients (all $p \geq 0.49$ for eggs and news; $p = 0.995$ for wild birds), and egg-price and news-volume specifications degraded out-of-sample root mean square error (RMSE) within their availability windows (Supplementary Table 1). Combined-model comparison under blocked cross-validation (Cerqueira et al., 2020) was therefore restricted to the three animal-case variables with at least partial in-sample signal — cats, cattle, and poultry (Table 2).

Out-of-sample forecasting performance was evaluated under blocked time-series cross-validation with Friedman and Bonferroni-corrected Wilcoxon signed-rank tests across the candidate specifications (Methods).

Table 2

ARIMAX forecasting performance under blocked cross-validation. Each row reports a candidate ARIMAX specification for the weekly Reddit comment-volume series, evaluated against the no-exogenous-variable baseline (ARIMA(0, 1, 1)) 10-week test folds, 5-week gap between training and test portions. Columns report the mean $\pm$ standard deviation (std) of test-fold root mean square error (RMSE) across the five folds, the percentage change in mean RMSE relative to the baseline, the worst-fold (maximum) RMSE observed across the five folds, and a qualitative verdict summarizing the specification's standing relative to the baseline. The best-performing specification (cats + cattle) is shown in bold.

Model	RMSE (mean $\pm$ std)	% RMSE vs baseline	RMSE Worst Fold	Verdict
baseline (no exog)	403.8 $\pm$ 482.2	—	1,032.8	Reference (no exog)
cats only	373.9 $\pm$ 474.5	-7.4%	1,106.7	Modest improvement
cats + cattle	279.4 $\pm$ 314.1	-30.8%	650.6	Best model — lowest RMSE & lowest std
cats + poultry	748.2 $\pm$ 1316.5	+ 85.3%	3,076.1	Worse than baseline
cats + cattle + poultry	549.9 $\pm$ 884.9	+ 36.2%	2,093.5	Worse than baseline

Under blocked cross-validation (Table 2), the cats + cattle ARIMAX produced the largest mean RMSE reduction over the autoregressive baseline (mean RMSE 279.4 vs. 403.8 across folds, approximately 31% reduction) and the lowest cross-fold standard deviation, indicating both higher mean accuracy and greater stability across periods. Because forecasts were back-transformed from the log-transformed counts used for model fitting, the reported RMSE values are on the original weekly comment-count scale: a mean RMSE of 279.4 corresponds to an average forecast error of approximately 279 comments per week for the cats + cattle specification, compared with roughly 404 comments per week for the autoregressive baseline. Cats + cattle produced lower forecasting error than the baseline in every one of the five test folds — five out of five, with no exceptions. The cats-only specification produced a more modest 7.4% mean RMSE reduction, while specifications containing poultry performed substantially worse than baseline (RMSE 748.2 with cats + poultry and 549.9 with all three exogenous variables), confirming that poultry cases add noise rather than predictive signal under rigorous evaluation. This was further validated by observing the worst single fold within the RMSE to ensure the RMSE improvement was from a consistent change rather than 1–2 high performing, outlier folds.

To assess whether these differences are statistically meaningful, a Friedman omnibus test followed by pairwise Wilcoxon signed-rank tests with Bonferroni correction were applied. While neither test crossed the $\alpha = 0.05$ threshold (Friedman $\chi^2 = 6.78$, $p = 0.148$; cats + cattle vs. baseline: $W = 0$, raw $p = 0.0625$, Bonferroni-corrected $p = 0.625$), the Wilcoxon reaches the lowest p-value of 0.625 with cats + cattle, because this combination outperformed the baseline in every single fold — the strongest pairwise evidence possible at this sample size. The Friedman non-significance reflects the heterogeneous fold structure noted above: in burnout folds all five specifications converge to near-zero errors and rank essentially at random, diluting the formal test. A sensitivity analysis re-running these tests on only the two folds whose test windows fell entirely within the active outbreak period confirmed the dilution argument (see Supplementary Note 2).

Per-Emotion ARIMAX Modeling

To examine whether outbreak signals also shape the emotional composition of H5N1 discussion, we disaggregated Reddit comments by dominant Text2Emotion label into five weekly volume series (Fear: $n = 12,645$; Sad: $n = 9,326$; Surprise: $n = 4,869$; Happy: $n = 4,688$; Angry: $n = 3,116$; total $n = 34,644$ of 42,957 comments; 8,313 lacked a non-zero T2E score). Each series was modeled with the same ARIMA(0, 1, 1) baseline and the same exogenous lags as the aggregate model (cats lag 1, wild mammals lag 3, cattle lag 1, poultry lag 5) and evaluated under identical blocked

cross-validation. Pairwise Spearman correlations between the exogenous animal series and emotion-specific comment volumes (Table 2) provide a pre-modeling check on which signals track which emotions.

Table 3

Spearman rank correlations between weekly mean emotion intensity scores and weekly confirmed HPAI animal-cases. Each cell reports the Spearman rank correlation coefficient (ρ) and its two-sided p-value between the weekly mean Text2Emotion intensity score for the indicated emotion (rows) and the weekly count of confirmed HPAI cases in the indicated animal category (columns; USDA APHIS, 2026b), computed across the study window (January 2024 – December 2025). Statistically significant correlations ($p < 0.05$) and their p-values are shown in bold.

Emotion	Cats: Spearman's ρ	Cats: p- value	Poultry: Spearman's ρ	Poultry: p-value	Wild mammals: Spearman's ρ	Wild mammals: p-value	Cattle: Spearman's ρ	Cattle: p- value
Anger	0.29	< 0.005	0.35	< 0.001	0.27	0.009	0.10	0.30
Sadness	-0.23	0.02	-0.10	0.31	-0.07	0.47	0.17	0.09
Happiness	0.12	0.22	0.15	0.14	0.24	0.01	0.25	0.01
Fear	0.12	0.25	-0.08	0.41	0.004	0.96	0.01	0.87
Surprise	-0.01	0.85	-0.09	0.35	0.02	0.83	-0.08	0.42

Anger correlated most broadly across the animal signals, happiness with the two mammalian signals (wild mammals and cattle), and sadness only weakly and negatively with cats; fear and surprise showed no significant pairwise correlations (Table 3). Because Spearman ρ captures monotonic association but not lagged forecasting value, we next fit ARIMAX with the same lag structure and cross-validation protocol as the aggregate comment-volume model (Table 4).

Table 4

Per-emotion ARIMAX forecasting performance under blocked cross-validation. Each cell reports the mean $\pm$ standard deviation (std) of test-fold root mean square error (RMSE) across the five blocked cross-validation folds, the percentage change in mean RMSE relative to the no-exogenous-variable baseline for that emotion, and the worst-fold RMSE. The two co-winning specifications per emotion (cats + cattle and cats + wild mammals + cattle) are shown in bold.

Emotion		Baseline (no exogenous variables)	cats + wild mammals	cats + cattle	cats + wild mammals + cattle	cats + wild mammals + poultry
Anger	RMSE	32.2 $\pm$ 40.5	27.6 $\pm$ 33.2	21.9 $\pm$ 26.5	21.2 $\pm$ 25.5	32.4 $\pm$ 42.1
	% Change	Not Applicable	-14.3%	-32.0%	-34.0%	+ 0.7%
	RMSE: Worst Fold	91.4	69.2	59.8	56.7	95.8
Sadness	RMSE	78.5 $\pm$ 93.2	71.8 $\pm$ 90.2	56.9 $\pm$ 65.2	56.4 $\pm$ 64.7	102.0 $\pm$ 156.5
	% Change	Not Applicable	-8.6%	-27.5%	-28.2%	+ 29.9%
	RMSE: Worst Fold	193.4	209.3	136.4	136.6	370.9
Happiness	RMSE	45.4 $\pm$ 54.0	41.8 $\pm$ 49.9	30.8 $\pm$ 35.1	30.2 $\pm$ 33.9	62.0 $\pm$ 92.1
	% Change	Not Applicable	-7.9%	-32.1%	-33.5%	+ 36.4%
	RMSE: Worst Fold	117.2	111.1	81.8	78.5	218.7
Surprise	RMSE	44.6 $\pm$ 54.3	42.8 $\pm$ 61.5	31.1 $\pm$ 38.0	30.9 $\pm$ 37.9	71.2 $\pm$ 125.5
	% Change	Not Applicable	-4.1%	-30.3%	-30.7%	+ 59.5%
	RMSE: Worst Fold	106.9	146.0	87.1	87.5	293.0
Fear	RMSE	126.6 $\pm$ 151.5	108.6 $\pm$ 131.4	96.7 $\pm$ 109.8	95.2 $\pm$ 108.8	212.9 $\pm$ 357.3
	% Change	Not Applicable	-14.2%	-23.6%	-24.8%	+ 68.2%
	RMSE: Worst Fold	338.7	301.8	237.1	239.0	840.9

Table 4 reports blocked cross-validation forecasting performance for five candidate specifications. The two-variable cats + wild mammals model produced modest RMSE reductions across all five emotions (4.1%–14.3%). Adding cattle (cats + cattle, cats + wild mammals + cattle) produced substantially larger reductions of 23.6%–34.0%, with the two cattle-inclusive models statistically indistinguishable (mean differences $\leq$ 3.3% relative; Wilcoxon signed-rank raw p between 0.375 and 1.000). We adopt cats + cattle as the primary per-emotion specification for consistency with the aggregate comment-volume model and treat cats + wild mammals + cattle as an equivalent alternative. Worst-fold RMSE followed the same rank order, with the cats + wild mammals + poultry model 1.7–3.5 $\times$ worse than the co-winners in the same fold – consistent with the wild mammals $\times$ poultry multicollinearity ($p = 0.64$) documented in the preceding section, which drives overfitting and RMSE inflation of 0.7%–68.2% for that specification.

Friedman tests reached $\alpha = 0.05$ for Anger ($\chi^2 = 10.857$, $p = 0.028$) and Sadness ($\chi^2 = 9.500$, $p = 0.049$) but not Fear, Happiness, or Surprise ($p \geq 0.218$). Bonferroni-corrected pairwise Wilcoxon contrasts ($m = 10$) did not cross $\alpha = 0.05$ for

any emotion, reflecting the burnout-fold dilution and the stricter correction imposed by the expanded candidate set. Both co-winner specifications nonetheless reached the $n = 5$ Wilcoxon floor ($W = 0$, raw $p = 0.0625$) against baseline in Anger and Sadness – the strongest pairwise evidence possible at this sample size – confirming they beat baseline in every fold for the two most event-reactive emotions.

Mechanism Analysis: Reddit Engagement with Animal-Case Reports

To further study the relationship between animal cases topics, comment volume, and emotions, five complementary tests were performed: (i) Spearman rank correlation at lags 0–2; (ii) an ARIMAX mechanism model with reported cases with topically-matched mention volume as the dependent variable; (iii) aggregate mentions-per-case ratios and log-log elasticities characterizing whether attention scales proportionally with case count; (iv) dominant-emotion base-rate distributions comparing each topically-matched subset against the corpus baseline to identify which emotions the subset over- or under-represents; and (v) bigram frequency analysis on the topically-matched subset to interpretively contextualize the discourse. Cattle-mentioning comments were analyzed under two keyword definitions – a strict set covering direct animal terms (cattle, cow, dairy cow) and a broad set that additionally captures the food-system framing (dairy, raw milk, milk) identified as prominent in the corpus's overall bigram analysis (see Supplementary Table 2 for full list).

Table 5

Reddit engagement mechanisms across three animal case signals. Five mechanism tests per topically-matched comment subset. ARIMAX fit only to strict cattle (direct match to USDA cattle cases); broad subset retained for descriptive comparison. Keyword lists in Methods. Spearman ρ and elasticity β all $p < 0.01$. Elasticity contrasts (ref. cattle strict): cats + 0.82 ($p = 0.037$), wild mammals – 0.12 ($p = 0.634$).

Metric	Cats	Cattle (strict)	Cattle (broad)	Wild mammals
Topically-matched comments, n (%)	1,653 (3.8%)	2,939 (6.8%)	5,265 (12.3%)	691 (1.6%)
Peak Spearman ρ (lag, weeks)	0.52 (lag 1)	0.71 (lag 0)	0.72 (lag 1)	0.28 (lag 0)
ARIMAX blocked-CV RMSE Δ vs. baseline	–19.8%	–2.1%	N/A	+ 6.8%
Ljung–Box p (lag 10)	0.68	0.95	N/A	0.76
Mentions per reported case	12.24	2.70	4.83	2.11
Log-log elasticity β [95% CI]	+ 1.503 [+ 0.696, + 2.311]	+ 0.683 [+ 0.395, + 0.971]	+ 0.788 [+ 0.477, + 1.099]	+ 0.561 [+ 0.168, + 0.955]
Over-represented emotions (vs. corpus baseline)	Anger 12.0% (9.0%), Fear 41.4% (36.5%)	Sadness 43.8% (26.9%)	Sadness 36.5% (26.9%)	Fear 49.1% (36.5%)
Under-represented emotions	Happy 8.7% (13.5%), Surprise 10.1% (14.1%)	Happy 5.4% (13.5%), Surprise 6.4% (14.1%), Anger 6.9% (9.0%)	Happy 8.1% (13.5%), Surprise 9.4% (14.1%)	Happy 5.0% (13.5%), Surprise 6.5% (14.1%)
Distinctive top bigrams (from top-25)	raw milk, raw food, cat food, pet food, indoor cat	raw milk, dairy cow, dairy farm, cow milk, farm worker	raw milk, dairy cow, drink raw, pasteurized milk, dairy herd	highly pathogenic, clade 2.3.4.4b, adapt mammal, sea lion, mink farm

The three signals show qualitatively distinct engagement mechanisms (Table 5). Domestic cats are an event-driven, attention-amplifying signal: mentions track reported cases at lag 1 with a significant conditional ARIMAX effect, a 19.8% blocked cross-validation RMSE reduction over baseline, and a super-proportional log-log elasticity — each additional case is associated with disproportionately more discussion. Cat-mentioning comments over-represent Anger and Fear, and bigram analysis surfaces a food-safety framing (raw milk, raw food, cat food, pet food). Cattle operate as a trend-driven, saturating signal: contemporaneous correlation is high, but the ARIMAX mechanism model shows no significant conditional effect after first-differencing under either keyword definition, and log-log elasticities are sub-proportional — each additional case adds less attention than the previous one. Cattle-mentioning comments are dramatically Sadness-dominant, and bigram analysis is food-system-dominated (raw milk is the second-most-frequent bigram in the strict subset). Wild mammals show the weakest and most saturated regime: the ARIMAX mechanism model fails to improve on baseline and instead inflates RMSE by 6.8%, and elasticity is statistically indistinguishable from cattle-strict in a formal three-way contrast ($p = 0.634$) while cats sit significantly above both ($p = 0.037$). Wild-mammal-mentioning comments are dramatically Fear-dominant (49.1%, the highest Fear share of any topic subset examined), with a technical, epidemiologically-oriented bigram framing (highly pathogenic, clade 2.3.4.4b, adapt mammal, sea lion, mink farm) suggesting science-attentive discussion of cross-species adaptation rather than event-specific reactivity.

Discussion

This study characterizes how the public engaged with the 2024–2025 H5N1 outbreak through Reddit discourse, and identifies the external signals — reported animal cases — that shaped both the volume and emotional composition of that engagement. Three findings emerge. First, the public discourse reflected the underlying One Health nature of the outbreak, spontaneously integrating human, animal, and environmental health themes without using the label itself. Second, comment volume and emotional composition were predicted by reports of infection in domestic cats and dairy cattle through mechanistically distinct pathways with distinct emotional signatures. Third, the integrated analytic pipeline demonstrated here — combining multi-signal ARIMAX, topic-level emotion modeling, and mechanism analysis — offers a generalizable framework for characterizing how outbreak signals shape online discourse.

First, the public discourse in the sampled Reddit communities reflected the underlying One Health nature of the H5N1 outbreak, integrating human health, animal health, and food safety themes without using the label itself. Food-safety framing dominated the corpus: *raw milk* was the second most-frequent bigram overall, and the topically matched comment subsets carried food safety framing across both companion-animal and livestock discussions — the cat-mentioning subset was anchored by *raw milk*, *raw food*, *cat food*, and *pet food*, while the cattle-mentioning subset was anchored by *raw milk*, *dairy cow*, *drink raw*, *pasteurized milk*, and *dairy herd* (Table 5). The emotional signature reinforced this framing: cattle-mentioning comments were dramatically Sadness-dominant (43.8% under the strict cattle keyword definition vs. 26.9% corpus baseline), a pattern more consistent with sustained concern about food-supply integrity than with acute crisis reactivity. This observed discourse tracks a real epidemiological hazard, as H5N1-infected dairy cow's milk contains high infectious titers (Burrough et al., 2024). Federal and state authorities increased raw-milk surveillance in response to the HPAI outbreak in dairy cattle through a mandatory testing order and interstate movement rules for lactating dairy cattle (USDA APHIS, 2024) and food safety warnings against raw milk consumption (CDC, 2024b; U.S. Food and Drug Administration, 2024). These initiatives were added onto the existing, largely decentralized regulation — raw milk consumption remains legal in many US states, with specific point-of-sale rules varying by jurisdiction (CDC, 2024a). The public's food safety framing therefore aligns with authority messaging on the outbreak, offering agencies a shared conceptual foundation for future risk communication. Whether this convergence reflects effective institutional messaging, spontaneous public conclusion, or a combination cannot be determined from our observational data alone.

This observed discourse of an engaged online audience contrasts sharply with the general US public: survey evidence documents that most Americans express low H5N1 concern and hold poor raw-milk knowledge, with a majority unaware that pasteurized milk is safer than raw (Motta et al., 2026) – despite peer-reviewed outbreak evidence documenting 202 raw-milk-linked outbreaks, 2,645 illnesses, and 228 hospitalizations in the US between 1998 and 2018 (Koski et al., 2022), current Centers for Disease Control and Prevention (CDC) guidance (CDC, 2024b), and risk assessments identifying raw milk as a non-negligible source of H5N1 infection risk (Chen et al., 2025; Koebel et al., 2026). The Reddit subset we characterize does not share the low H5N1 concern of the general public: this community is concerned and integrates multiple One Health domains in its discussion without regularly using the term "One Health" itself. This contrast matters because public conceptions of zoonotic spillovers remain generally fragmented: in a recent large-scale Australian survey, only 14.3% of general respondents articulated a full One Health problem definition when asked about zoonotic spillovers, with 46.9% using a human-animal framing that omitted the environment and 27.7% using a human-only framing (Shanahan et al., 2026). Against this baseline, the engaged Reddit H5N1 community produced an integrated multi-domain system-level framing: human infection, cat cases, cattle cases, wildlife cases, dairy industry and even biological and virological terminology in the wild-mammal subset. The implication is that risk communication for future spillover events does not need to teach the One Health framework from scratch to this audience; it can build on the framing the audience is already using. Importantly, our analysis identifies what the audience discusses rather than assesses the accuracy of individual comments. Thus, practical communication approach could include naming the already used integrated framing as *One Health* and connecting existing food-system discussion to evidence-based risk mitigation strategies. This corresponds to a "yes, and" approach – acknowledging existing framing and co-building additional context rather than correcting – a strategy operationalized in the motivational-interviewing tradition (Gagneur et al., 2024).

Second, the two animal signals that predicted comment-volume dynamics produced distinct emotional signatures and distinct attention profiles. Cat-mentioning comments over-represented both anger (12.0% vs. 9.0% corpus baseline) and fear (41.4% vs. 36.5%), cat-inclusive per-emotion ARIMAX specifications significantly outperformed the no-exogenous baseline for anger (Table 4) – a temporally associated pattern that parallels prior social-media evidence from public health emergencies. Yu et al. (2021) documented similar pattern in Wuhan Weibo during the early COVID-19 outbreak, with both anger and fear triggered by specific milestone events (confirmation of human-to-human transmission, lockdown declaration, whistleblower death) and fear especially dominant in the early outbreak phase. Our cat-case announcements resemble a similar class of milestone event: discrete, temporally localized events that carry perceived immediate personal-risk implications. Cattle-mentioning comments, by contrast, were dramatically Sadness-dominant (43.8% under the strict cattle keyword definition vs. 26.9% baseline), and this pattern persisted across the outbreak timeline – consistent with the trend-driven, saturating attention dynamic we identified in the mechanism analysis, with prior evidence that sadness persists during and after peak outbreak phases (Ou et al., 2025), and with the broader pattern that cumulative anxiety accompanies sustained health crises (Wenchang et al., 2026). The two signals also differed sharply in attention profile: cats produced super-proportional log-log elasticity, each additional case generating more than a proportional increase in discussion, whereas cattle produced sub-proportional elasticity, with attention saturating as cases accumulated. This attention asymmetry is plausibly reflects the human-companion-animal bond: cat caregivers attribute more complex emotions and cognition to their cats in proportion to bond strength (Vonk & Bouma, 2024), so cat cases appear to land not as epidemiological events but as threats to sentient companions. Livestock lacks a comparable relational bond, as public concern for farm-animal welfare is culturally reinforced as lower than for pets. This differential is consistent with the psychological mechanisms of empathy that shape pet relationships, and is reinforced by the "meat paradox" – the cognitive dissonance that allows people to hold positive attitudes toward pets while accepting the killing of food animals (Prato-Previde et al., 2022). Our findings, viewed alongside this literature, suggest that pets appear to activate protective, personal-stakes frames, while livestock activate utility, agricultural-management frames.

At the corpus level, fear (36.5%) and sadness (26.9%) dominated the overall emotion distribution, matching the fear-dominant patterns observed across other outbreak-social-media studies — Xue et al. (2020) on COVID-19 Twitter, Yu et al. (2021) on Wuhan Weibo, and Ou et al. (2025) on influenza A in China — suggesting that outbreak discourse has a general fear/sadness baseline that recurs across diseases, platforms, and cultures. In our data, however, this baseline was dissociable from case-reactive dynamics: anger was the most broadly case-correlated emotion — significantly correlated with cat, poultry, and wild-mammal cases — while fear and surprise showed no significant case-signal correlations, representing largely stable emotional registers of the outbreak-attentive audience. Sadness correlated only weakly (and negatively) with cat cases, and happiness tracked mammalian signals (wild mammals and cattle). This suggests that cat-case announcements are temporally associated with spikes in anger during the H5N1 outbreak, and therefore represent a tractable window for crisis and emergency risk communication (Reynolds & Seeger, 2005). Crisis and emergency risk communication timed based on such stimuli should acknowledge the emotional response, provide accurate hazard framing (e.g., H5N1 risk from raw milk and raw pet food), and offer concrete protective actions (e.g., avoid raw milk, avoid raw pet food) to help bring public perception into alignment with actual risk, leaving audiences both calmer and better protected.

Third, beyond its H5N1-specific findings, this study contributes a generalizable methodological framework for characterizing how outbreak signals shape online discourse. The framework integrates four analytic layers into a single reproducible pipeline: (i) Spearman correlations between candidate exogenous variables and the response variable as a pre-modeling check; (ii) multi-signal ARIMAX with CCF-identified lags across parallel case-reporting streams; (iii) emotion-decomposed per-topic ARIMAX; and (iv) mechanism analysis via log-log elasticity, attention-per-case ratios, base-rate emotion distributions, and bigram frequency. The framework builds on prior work operationalizing the concept of social media as a tool for outbreak surveillance (Wilson et al., 2021), and extends established ARIMAX-family approaches that combine social-media signals with case-count time series — including Lamsal et al. (2022), who used Twitter conversation volume and sentiment-topic latent variables to forecast COVID-19 case counts in Australia (48.8–51.4% RMSE improvement over baseline), and Sigalo et al. (2023), who used Twitter vaccine-attitude sentiment as an exogenous input into vaccine-uptake forecasting models. Where existing infoveillance frameworks produce descriptive outputs (topic prevalence, sentiment distribution) or predictive outputs (nowcasting, forecasting), our framework introduces a mechanism-discriminating output, a quantitative signature (log-log elasticity coefficient + attention-per-case ratio) that classifies each exogenous signal as event-driven (super-proportional), trend-driven (sub-proportional), or saturated. This signature is transferable to any outbreak with parallel case-reporting infrastructure and online discourse, zoonotic or otherwise.

This study has three main limitations. First, the sampled Reddit community is not representative of the general US public, of all US residents, or even of all Reddit users; no demographic data on the users generating the comments analyzed here were available, and inferences drawn from this corpus should not be extrapolated to broader populations without appropriate caution. Second, the two-year observation window is short for time-series modeling on public-discussion datasets, limiting both the number of cross-validation folds available for out-of-sample evaluation and the ability to detect longer-period cycles in discourse dynamics. Third, and most fundamentally, the ARIMAX framework establishes predictive association rather than causation: a significant lagged relationship between reported animal cases and comment volume is consistent with cases driving discussion, but does not rule out a shared response to an unmeasured common cause, for example, policy events that may drive both case reporting and public attention in parallel.

In summary, these findings characterize how a two-year Reddit corpus engaged with the 2024–2025 H5N1 outbreak and identify the external signals that shaped that engagement. The public discourse reflected the underlying One Health nature of the outbreak, spontaneously integrating human health, animal health, food safety, and environmental themes with food-safety concerns, most visibly the prominence of raw milk, anchoring the discussion. Comment

volume was predicted by reports of infection in domestic cats and dairy cattle through mechanistically distinct pathways: cats operated as an event-driven, super-proportional signal that also accompanied by case-reactive anger, while cattle operated as a trend-driven, saturating signal accompanied by sadness across the study timeline, a species-level asymmetry consistent with the human-companion-animal bond. Finally, the integrated multi-layer analytic pipeline demonstrated here — combining Spearman screening, multi-signal ARIMAX, emotion-decomposed per-topic ARIMAX, and mechanism analysis via elasticity, attention-per-case, emotion base rates, and bigrams — extends existing infoveillance methods with a mechanism-discriminating output that is transferable to any outbreak with parallel case-reporting infrastructure and online discourse.

Methods

Data Collection

Reddit discussion data were collected using the official Reddit API, including posts, comments, and associated metadata such as timestamps, subreddit identifier, comment score, and moderator flag. Three H5N1-focused communities were selected for inclusion: r/H5N1_AvianFlu (47,153 comments), r/BirdFlu (824 comments), and r/BirdFluPreps (1,836 comments). Together, these communities span 2010 through 2025; activity prior to 2024 was sparse, and the analysis was therefore restricted to a two-year window from January 1, 2024 to December 31, 2025 to focus on the active outbreak period. Data were collected in two passes — an initial extraction in August 2025 followed by a recollection in February 2026 to capture the complete 2025 window — and comments were deduplicated on identifier to ensure each comment was counted only once. Reddit relies on community moderators — volunteer users empowered to remove comments that violate community rules or contain harmful content — to maintain discussion quality, and comments flagged as removed by subreddit moderators in the API metadata were excluded from the analysis. After defensive deduplication on the comment identifier, exclusion of moderator-removed comments, and restriction to the study window, 42,957 comments were retained for analysis.

Media coverage data were collected using the SerpAPI Google News interface (Developer plan). The dataset comprises 4,892 unique English-language articles published between January 2024 and December 2025. Articles were retrieved by repeated queries using the following 14 search terms: "bird flu," "avian influenza," "H5N1," "avian flu," "highly pathogenic avian influenza," "HPAI," "low pathogenic avian influenza," "LPAI," "H5 avian influenza," "H7N9," "H5Nx," "poultry influenza," "wild bird influenza," "zoonotic influenza," and "bird influenza." Pagination was continued until no additional results were returned for any query, ensuring complete coverage of the indexed corpus within the study window. The collected metadata included publication date, source outlet, article title, and snippet text. Articles were deduplicated on URL and aligned to weekly Monday-anchored bins for all downstream analyses.

Animal-based food price data were obtained from the USDA Economic Research Service Meat Price Spreads dataset (Vaiknoras, 2026), which provides monthly price observations for a range of animal-derived products. Retail and wholesale prices for "Eggs, Grade A large" were extracted as the focal economic series. The dataset was available only through July 2025; coverage was not extended because weekly Reddit comment volume in the final months of 2025 was too low for eggs to add predictive value. Monthly observations were upsampled to weekly resolution by forward-filling within each calendar month, so that every week inherits its containing month's price and is anchored to the same Monday grid used for the other series. This alignment was required for the inclusion of egg prices as an exogenous regressor in the ARIMAX models, which assume that the target and all exogenous variables share a common sampling frequency; forward-filling rather than linear interpolation was used to avoid introducing artificial precision the source data does not contain.

Datasets of reported cases of H5N1 infection in animal populations were obtained from publicly available federal databases maintained by the USDA APHIS (USDA APHIS, 2026b). Four databases reporting cases were used: (i) “commercial and backyard poultry”, (ii) “dairy cattle”, (iii) “mammals” (reporting on non-livestock cases, including domestic cats as well as wild and captive wild mammals), and (iv) “wild birds”. Each case represents one reported infection event for the epidemiological unit of concern, which was individual flock for poultry, individual herd for dairy cattle and individual animal for non-livestock mammals and wild birds, with associated date, state, and species or production-type metadata. The non-livestock mammal cases were further subdivided into domestic cats and wild mammals (foxes, mice, seals, raccoons, bobcats, and other free-living species), reflecting the differences between exposures in domestic and wild animals. All case datasets were filtered to the January 1, 2024 – December 31, 2025 study window and aggregated to weekly counts on the same Monday-anchored grid as the Reddit time series.

Confirmed human H5N1 cases were excluded as a candidate exogenous variable: the US CDC recorded 71 cases over the 105-week study window (CDC, 2025) (mean 0.68/week, ~ 50% zero-case weeks), a sparsity level at which log-transformed Gaussian ARIMAX becomes unreliable (Cameron & Trivedi, 2013).

Reddit Text Preprocessing

All Reddit post and comment text was preprocessed prior to natural-language analyses using the Python NLTK and SpaCy libraries. Preprocessing steps included lowercasing, tokenization, and removal of URLs, punctuation, numerals, emojis, and English stopwords, followed by lemmatization to reduce inflectional variants to their root forms.

Topic Modeling

Two complementary methods were applied to characterize dominant themes in the Reddit discussions. First, n-gram frequency analysis was performed on the preprocessed token stream to identify the most commonly occurring bigrams. Second, semantically equivalent bigrams were mapped to a single canonical term prior to tabulation to prevent the dispersion of frequency counts across interchangeable expressions: bigrams denoting the disease (e.g., ‘avian influenza,’ ‘avian flu,’ ‘bird flu’) were mapped to ‘avian flu,’ and bigrams denoting dairy livestock (e.g., ‘dairy cow,’ ‘dairy cattle’) were mapped to ‘dairy cattle.’ Stemmed terms were also converted to readable word forms; for example “mani peopl” was displayed as “many people.” In addition, high-frequency conversational expressions that carry no substantive H5N1-related information (e.g., ‘feel like,’ ‘gonna’) were filtered from the bigram outputs. Second, LDA was applied using the Gensim library. Models were fitted with three, four, and five candidate topics and compared on the basis of topic coherence and substantive interpretability of the dominant terms in each topic. The three-topic specification was retained for reporting because it produced the most distinctive and interpretable thematic groupings.

Sentiment Classification

Comment-level sentiment was assessed using the Text2Emotion Python library (Py_Team, 2020), which maps text to five basic emotion categories — happiness, sadness, anger, fear, and surprise. Each comment was assigned a continuous score between 0 and 1 for each emotion, reflecting the relative intensity of that emotion in the text. The five per-comment emotion scores were used in two ways downstream: (i) aggregated to weekly means to characterize the evolving emotional register of the discussion, and (ii) used to assign each comment to its dominant emotion category (the argmax across the five scores) for subsequent per-emotion volume modeling.

Toxicity Classification

Toxicity was scored with ToxicBERT, a BERT-based classifier fine-tuned on Reddit and Wikipedia comments to detect insults, threats, and identity attacks, outputting a 0–1 probability per comment (Hanu & Unitary team, 2020). As a robustness check, toxicity was additionally scored using the Perspective API, which rates text on a 0–1 scale for

perceived rudeness using semantic context (Jigsaw, 2017). Both methods agreed, indicating low toxicity across the discourse (Fig. 3 and Supplementary Fig. 2).

Data Processing for Modeling

To examine temporal relationships between Reddit engagement and external indicators, all variables were converted to weekly resolution on a common Monday-anchored grid spanning the study window. Reddit comment counts, animal-case counts, and news articles were summed per week. Egg prices, originally monthly, were up sampled to weekly resolution by forward-filling within each calendar month. Pairwise relationships between Reddit comment volume and each exogenous variable were quantified using Spearman rank correlation (`scipy.stats.spearmanr`), with the two-sided p-value reported alongside the correlation coefficient ρ . The same procedure was applied to compute pairwise Spearman correlations between weekly mean Text2Emotion intensity scores and the animal-case exogenous variables for the sentiment temporal analysis. Spearman p-values are reported as approximate because the standard test assumes independent observations and is anticonservative on autocorrelated time series.

For the cross-variable temporal overview in Fig. 4 (and the full eight-variable version in Supplementary Fig. 5), each weekly series was $\log_{10}$ -transformed, smoothed with a 4-week centered rolling mean to suppress short-window noise, and then min–max normalized to the interval $[0, 1]$ using $(x - x_{\min}) / (x_{\max} - x_{\min})$ to render heterogeneous units comparable on a common axis. The $\log_{10}$ transformation matched the treatment of the variables in the ARIMAX modeling; the smoothing and normalization were applied for visualization only and were not propagated to the ARIMAX modeling, which used the unsmoothed, non-normalized log-transformed series throughout.

ARIMAX Modeling

A univariate ARIMA baseline was identified for the weekly Reddit comment-volume series using the Box–Jenkins methodology. All weekly series used in the modeling pipeline — the Reddit comment-volume series (dependent variable) and all candidate exogenous regressors — were transformed using $\log(1 + \text{count})$ to stabilize the variance of the right-skewed distributions while preserving the meaning of zero-count weeks. Stationarity of the log-transformed comment-volume series was assessed using both the Augmented Dickey–Fuller (ADF) test, whose null hypothesis is the presence of a unit root, and the Kwiatkowski–Phillips–Schmidt–Shin (KPSS) test, whose null hypothesis is stationarity. The level series was non-stationary by both tests (ADF $p = 0.32$, KPSS $p = 0.04$), while the first-differenced series was stationary (ADF $p < 0.001$, KPSS $p = 0.10$), establishing the integration order $d = 1$. Autoregressive (p) and moving-average (q) orders were selected by inspection of the autocorrelation and partial autocorrelation functions of the differenced series: the ACF exhibited a clean cutoff at lag 1 and the PACF showed gradual decay across multiple lags, the diagnostic signature of a pure MA(1) process. The resulting specification, ARIMA(0, 1, 1), was retained as the baseline for all subsequent ARIMAX comparisons.

The baseline ARIMA(0, 1, 1) was extended to ARIMAX by introducing log-transformed weekly exogenous regressors at lags identified through CCF analysis. For each candidate exogenous variable — egg prices (retail and wholesale), news article volume, and confirmed H5N1 cases in poultry, cattle, domestic cats, wild mammals and wild birds — the CCF between the log-transformed exog series and the log-transformed Reddit comment volume was computed at lags from 0 to 8 weeks. The optimal lag for each variable was selected as the smallest lag at which the correlation was maximal in magnitude and consistent with the theoretical expectation that the exogenous signal leads the target. The resulting lag choices were: cats lag 1, cattle lag 1, wild mammals lag 3, and poultry lag 5. All ARIMAX models were fitted using the SARIMAX class from the statsmodels Python library, with `enforce_stationarity = False` and `enforce_invertibility = False`. Model performance was evaluated through both in-sample fit (coefficient z-tests and 95% confidence intervals; Ljung–Box test on residuals at lag 10) and out-of-sample forecasting accuracy under blocked time-series cross-validation. Pairwise Spearman correlations between candidate exogenous variables were also computed to detect

predictors carrying overlapping information; any pair correlating above $p = 0.6$ was considered too redundant to include together in the same model. Additionally, across all specifications reported in Table 2, the Ljung–Box test at lag 10 failed to reject residual independence ($p > 0.4$), confirming that the ARIMA(0, 1, 1) structure adequately captured the autoregressive dynamics of the comment-volume series.

Per-Emotion ARIMAX Modeling

To assess whether external signals predict emotion-specific dimensions of online engagement beyond the aggregate volume measure, ARIMAX models were also fitted separately for each emotion-dominant comment volume series. Five weekly count series were constructed by assigning each comment to its dominant Text2Emotion category and aggregating the resulting counts on the same Monday-anchored grid. Each emotion-specific series was log-transformed and modeled with the same ARIMA(0, 1, 1) order identified for the aggregate volume series, with the same lag-selected exogenous regressors. Four multivariate specifications were evaluated alongside the no-exogenous baseline: two two-exog models – cats + wild mammals and cats + cattle (the latter matching the primary comment-volume specification) – and two three-exog extensions of the cats + wild mammals base adding either cattle (lag 1) or poultry (lag 5). For both the primary cats + cattle specification and the cats + wild mammals + cattle alternative co-winner, the Ljung–Box test at lag 10 failed to reject residual independence in all five emotion-dominant series (cats + cattle: p between 0.29 and 0.75; cats + wild mammals + cattle: p between 0.34 and 0.76), confirming that the same ARIMA(0, 1, 1) structure adequately captured each emotion-specific series.

Blocked Time-Series Cross-Validation

Out-of-sample forecasting performance was evaluated using blocked time-series cross-validation, implemented with the TimeSeriesSplit utility from scikit-learn. Five non-overlapping ten-week test folds were defined across the study window, each preceded by a five-week gap to prevent autocorrelation leakage from training observations into the test period. Training windows expanded across folds, with the model refitted on the training portion of each fold and forecasts generated for the held-out ten-week period. Root-mean-square error (RMSE) and mean absolute error (MAE) were computed on the back-transformed scale of weekly comment counts (using the inverse log transformation $\exp(\text{prediction}) - 1$) so that reported error metrics are directly interpretable in comments per week. For each model $\times$ fold combination, the per-fold RMSE was retained; the mean, standard deviation (std), and worst-fold (maximum) RMSE across the five folds were reported as the headline performance metrics, with the full per-fold breakdowns shown in Supplementary Fig. 6 (comment-volume comparison) and Supplementary Fig. 7 (per-emotion comparison). To assess whether the differences in per-fold RMSE between specifications were statistically meaningful, a Friedman omnibus test, followed by pairwise Wilcoxon signed-rank tests with Bonferroni correction, were applied to the per-fold RMSE values, following the non-parametric framework recommended for comparing multiple models evaluated on the same set of cross-validation folds. The framework was applied uniformly to the comment-volume model comparison and to each of the five per-emotion model comparisons; for the per-emotion case, no outer family-wise correction was applied across emotions, consistent with the Demšar (2006) framework in which each evaluation context is treated as an independent unit of comparison. Blocked cross-validation was preferred over a single end-of-series holdout because the original 12-week holdout window covered a burnout period, during which all candidate models converged trivially to near-zero forecasts and produced indistinguishable error metrics, masking real differences in forecasting performance.

Mechanism Analysis

To characterize how the ARIMAX-identified exogenous signals shape Reddit engagement at a mechanism level, weekly counts of topically-matched comments were computed for each of the three animal categories retained in the ARIMAX specifications (domestic cats, dairy cattle, wild mammals) and analyzed against the corresponding APHIS reported case counts. Topically-matched comments were identified via case-insensitive, word-boundary–anchored regular

expressions applied to the raw comment text; full keyword lists for cats, cattle (strict), cattle (broad), and wild mammals are provided in Supplementary Table 2. APHIS reported wild-mammal cases were computed by filtering the “mammals” database to exclude the 'Domestic cat' species value; all other mammal species were treated as wild-mammal cases. Weekly mention counts were aggregated on the same Monday-anchored grid used for the main ARIMAX analyses.

For each animal population, five complementary tests were applied: (i) Spearman’s rank correlation analysis, (ii) ARIMAX predictive fit, (iii) mentions-per-case ratio and log-log elasticity, (iv) emotion distributions and corresponding analysis, (v) bigram frequency of topics. For (i), the `scipy.stats.spearmanr` package was used to compute the relationship between weekly cases and weekly animal-related mentions at lags 0, 1, and 2. For (ii), an ARIMAX model was fit using the same ARIMA(0, 1, 1) order as the main comment-volume ARIMAX model, with log-transformed cases at the animal’s ARIMAX-identified lag (cats: lag 1; cattle: lag 1; wild mammals: lag 3) as the exogenous predictor of log-transformed weekly mention volume, and compared against a no-exogenous baseline via likelihood-ratio test, coefficient significance, Ljung–Box residual diagnostics at lag 10, and blocked cross-validation using the same five-fold protocol applied in the main ARIMAX analyses. For (iii), an aggregate mentions-per-case ratio (total study-window mentions ÷ total study-window cases) and a log-log ordinary least squares (OLS) elasticity β estimated from $\log_{10}(\text{mentions}) \sim \log_{10}(\text{cases})$ were computed on the subset of weeks with at least one case; when comparing multiple animal categories, a stacked OLS regression with animal-type interaction terms was used to formally test elasticity differences. For (iv), each comment with detectable emotion content (non-zero score for at least one of the five T2E categories) was assigned to its dominant emotion category (the `argmax` across the five scores), and the percentage distribution within each topically-matched subset was compared against the corpus baseline. For (v), bigram frequency analysis was performed on the topically-matched comment subsets following the same procedure as the corpus-wide bigram analysis (see Reddit Text Preprocessing subsection), with the top 15 bigrams per subset retained for interpretive contextualization. These combined outputs can be used to classify each exogenous signal by its attention-mechanism profile — event-driven (super-proportional), trend-driven (sub-proportional), or saturated.

Robustness and Sensitivity Analyses

Three additional analyses were conducted to test the robustness of the ARIMAX modeling framework. First, to verify that the min-max normalization used in the cross-variable visualization (Fig. 4 and Supplementary Fig. 5) does not affect the modeling conclusions, the comment-volume and per-emotion ARIMAX blocked-cross-validation pipelines were rerun with min-max scaling applied on top of the $\log_{10}$ transformation to all input variables, with forecasts back-transformed to the original comment-count scale before RMSE calculation (Supplementary Note 1). Second, to empirically verify that the non-significance of the Friedman omnibus and pairwise Wilcoxon tests reflected fold-count-dependent statistical power (“burnout-fold dilution”) rather than an absent effect, both tests were re-applied to the subset of folds whose test-window mean weekly comment volume exceeded the near-zero threshold (Supplementary Note 2; Supplementary Fig. 8). Third, after identifying the best additive specification, first-order interactions, cats × cattle, were tested in the highest performing ARIMAX model (comment-volume cats + cattle) as well as for both per-emotion co-winner specifications (cats + cattle and cats + wild mammals + cattle). Each test compared the extended model against its additive baseline via in-sample likelihood-ratio test and blocked cross-validation using the same five-fold protocol as the main model comparisons, with family-wise Bonferroni correction applied to each test family (Supplementary Note 3).

Software and Reproducibility

All analyses were conducted in Python 3.13.15. Using the following libraries: `pandas` and `numpy` for data manipulation; `matplotlib` and `seaborn` for visualization; `NLTK` and `SpaCy` for text preprocessing; `Gensim` for Latent Dirichlet Allocation;

Text2Emotion for sentiment classification; the Hugging Face Transformers implementation of ToxicBERT for toxicity classification; statsmodels for ARIMA, ARIMAX, and stationarity testing; scikit-learn (TimeSeriesSplit) for blocked time-series cross-validation; and scipy.stats for Spearman correlation testing, Friedman omnibus tests (scipy.stats.friedmanchisquare), and pairwise Wilcoxon signed-rank tests (scipy.stats.wilcoxon). Code is available via the project repository.

Declarations

Competing Interests. The authors declare no competing interests.

Author Contributions

L.D.P. conceived the study, designed and implemented the analysis pipeline, collected and processed the Reddit and epidemiological data, performed the topic modeling, sentiment and toxicity classification, and ARIMAX time-series analyses, generated all figures and tables, and wrote the initial draft of the manuscript. A.L.N. provided mentorship on the computational and statistical methodology, including the ARIMAX modeling and cross-validation framework, and contributed to manuscript review and editing. K.J.K. provided expert guidance on H5N1 epidemiology, host range, and contributed to manuscript review and editing. R.I. supervised the whole project, contributed to study conceptualization and analytical design, interpretation of findings, and contributed to manuscript review and editing. All authors read and approved the final manuscript.

Ethics Declaration

This study was reviewed by the Cornell University Institutional Review Board (IRB) and received an exemption determination (Protocol #IRB0150253, dated October 23, 2025).

Consent to Participate

This study did not involve direct interaction with human participants, and the analyzed data contained no identifiable private information. All data consisted of publicly available Reddit comments collected via the official Reddit Data API in accordance with Reddit's Terms of Service and User Agreement. Prior to manuscript submission, the authors contacted Reddit to describe the planned research use of the collected data, and Reddit confirmed its authorization for use in manuscript publication (email correspondence dated October 7, 2025). User identifiers (usernames or user IDs) were not retrieved as part of the data collection, and no verbatim user comments are reproduced in the manuscript. Accordingly, individual informed consent and consent for publication were not required for this study.

Data Availability

The Reddit comment corpus and the Google News snippet corpus collected for this study cannot be redistributed by the authors: Reddit's API Terms of Service and User Agreement prohibit the public release of user-generated content, and SerpAPI's Terms of Service restrict redistribution of search results obtained through their platform. To enable independent reproduction, the project repository (see Code availability) provides the full data-collection scripts together with documentation on how to obtain the required Reddit API and SerpAPI credentials. All external databases used as exogenous variables in this study are publicly available at their respective sources. Reported confirmed H5N1 cases in poultry, dairy cattle, mammals, and wild birds are available from the USDA APHIS HPAI Detections dashboards

(<https://www.aphis.usda.gov/livestock-poultry-disease/avian/avian-influenza/hpai-detections>). Retail egg price data are available from the USDA (<https://www.ers.usda.gov/data-products/meat-price-spreads>).

Code Availability

All code used in this study is openly available at <https://github.com/IvanekLab/Reddit-NLP-ARIMAX-Project> and archived with a permanent DOI at Zenodo (10.5281/zenodo.23070435). The repository includes: (i) the Python scripts used to collect the Reddit comment corpus and the Google News snippet corpus, with documentation on how to register for the required Reddit API and SerpAPI keys; (ii) the text preprocessing, topic modeling, sentiment classification (Text2Emotion), and toxicity classification (ToxicBERT, Perspective API (Jigsaw)) pipelines; (iii) the ARIMAX modeling and blocked cross-validation code; and (iv) the figure- and table-generation scripts. Analyses were performed in Python 3 using the libraries listed in the *Software and Reproducibility* subsection of the Methods.

References

1. Amaya, A., Bach, R., Keusch, F., & Kreuter, F. (2021). New Data Sources in Social Science Research: Things to Know Before Working With Reddit Data. *Social Science Computer Review*, 39(5), 943–960. <https://doi.org/10.1177/0894439319893305>
2. Burrough, E. R., Magstadt, D. R., Petersen, B., Timmermans, S. J., Gauger, P. C., Zhang, J., Siepker, C., Mainenti, M., Li, G., Thompson, A. C., Gorden, P. J., Plummer, P. J., & Main, R. (2024). Highly pathogenic avian influenza A(H5N1) clade 2.3.4.4b virus infection in domestic dairy cattle and cats, United States, 2024. *Emerging Infectious Diseases*, 30(7), 1335–1343. <https://doi.org/10.3201/eid3007.240508>
3. Cameron, A. C., & Trivedi, P. K. (2013). *Regression analysis of count data* (2nd ed.). Cambridge University Press. <https://doi.org/10.1017/CBO9781139013567>
4. CDC. (2024a). *Map of State Laws on the Sale of Unpasteurized Cow's Milk, 2012–2019*. <https://www.cdc.gov/php/publications/unpasteurized-cow-milk.html>
5. CDC. (2024b). *Talking to Patients about Unpasteurized (Raw) Milk and Highly Pathogenic Avian Influenza*. <https://www.cdc.gov/bird-flu/hcp/unpasteurized-raw-milk/index.html>
6. CDC. (2025). *H5 Bird Flu: Current Situation*. <https://www.cdc.gov/bird-flu/situation-summary/index.html>
7. Cerqueira, V., Torgo, L., & Mozetič, I. (2020). Evaluating time series forecasting models: An empirical study on performance estimation methods. *Machine Learning*, 109(11), 1997–2028. <https://doi.org/10.1007/s10994-020-05910-7>
8. Chen, Y., Dean, K. J., Fenske, G. J., Murphy, S. I., Gavelek, A., Pouillot, R., Van Doren, J. M., & Dennis, S. (2025). Rapid risk assessment to address emerging concerns of HPAI in raw and pasteurized milk. *PLOS One*, 20(6), e0322948. <https://doi.org/10.1371/journal.pone.0322948>
9. Cooper, L. N., Diaz, M. I., Hanna, J. J., Most, Z. M., Lehmann, C. U., & Medford, R. J. (2025). Birds of a feather? Mis- and dis-information on the social media platform X related to avian influenza. *Antimicrobial Stewardship & Healthcare Epidemiology*, 5(1), e4. doi:10.1017/ash.2024.471
10. Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7, 1–30. <https://dl.acm.org/doi/10.5555/1248547.1248548>
11. Duong, T. (2024). *Distribution of Reddit.com traffic 2024, by country*. Statista. <https://www.statista.com/statistics/325144/reddit-global-active-user-distribution/>
12. U.S. Food and Drug Administration. (2024). *The dangers of raw milk: Unpasteurized milk can pose a serious health risk*. <https://www.fda.gov/food/buy-store-serve-safe-food/dangers-raw-milk-unpasteurized-milk-can-pose-serious->

13. Gagneur, A., Gutnick, D., Berthiaume, P., Diana, A., Rollnick, S., & Saha, P. (2024). From vaccine hesitancy to vaccine motivation: A motivational interviewing based approach to vaccine counselling. *Human Vaccines & Immunotherapeutics*, 20(1). <https://doi.org/10.1080/21645515.2024.2391625>
14. Garg, S., Reinhart, K., Couture, A., Kniss, K., Davis, C. T., Kirby, M. K., Murray, E. L., Zhu, S., Kraushaar, V., Wadford, D. A., Drehoff, C., Kohnen, A., Owen, M., Morse, J., Eckel, S., Goswitz, J., Turabelidze, G., Krager, S., Unutzer, A., ... Olsen, S. J. (2025). Highly pathogenic avian influenza A(H5N1) virus infections in humans. *New England Journal of Medicine*, 392(9), 843–854. <https://doi.org/10.1056/NEJMoa2414610>
15. Hanu, L., & Unitary team. (2020). Detoxify [Computer software]. GitHub. <https://github.com/unitaryai/detoxify>
16. Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., & Zhao, L. (2019). Latent Dirichlet Allocation (LDA) and topic modeling: models, applications, a survey. *Multimedia Tools and Applications*, 78(11), 15169–15211. <https://doi.org/10.1007/s11042-018-6894-4>
17. Jigsaw. (2017). Perspective API. Google Jigsaw. <https://perspectiveapi.com/>
18. Koebel, K. J., Bulut, E., Alcaine, S. D., Trmčić, A., Nooruzzaman, M., Covaleda, L. M., Diel, D. G., & Ivanek, R. (2026). Quantitative microbial risk assessment of human H5N1 infection from consumption of fluid cow's milk. *Risk Analysis*, 46(7). <https://doi.org/10.1111/risa.70264>
19. Koski, L., Kisselburgh, H., Landsman, L., Hulkower, R., Howard-Williams, M., Salah, Z., Kim, S., Bruce, B. B., Bazaco, M. C., Batz, M. B., Nichols, M., Basler, C., Hall, A. J., Barlow, K., Boyle, M., Cooley, L. A., ... Gould, L. H. (2022). Foodborne illness outbreaks linked to unpasteurised milk and relationship to changes in state laws – United States, 1998–2018. *Epidemiology & Infection*, 150, e183. <https://doi.org/10.1017/S0950268822001649>
20. Lamsal, R., Harwood, A., & Rodriguez Read, M. (2022). Twitter conversations predict the daily confirmed COVID-19 cases. *Applied Soft Computing*, 129, 109603. <https://doi.org/10.1016/j.asoc.2022.109603>
21. Mallapaty, S. (2024, April 27). Bird flu virus has been spreading among US cows for months, RNA reveals. *Nature*. <https://www.nature.com/articles/d41586-024-01256-5>
22. Motta, M., Lunz Trujillo, K., Stecula, D., Callaghan, T., Ophir, Y., & Walter, D. (2026). Public risk perceptions and policy attitudes toward highly pathogenic avian influenza (HPAI). *Journal of Health Politics, Policy and Law*. <https://doi.org/10.1215/03616878-12317977>
23. Ou, Y., de Bruijn, G.-J., & Schulz, P. J. (2025). Social media as an emotional barometer: Bidirectional encoder representations from transformers–Long short-term memory sentiment analysis on the evolution of public sentiments during influenza A on Sina Weibo. *Journal of Medical Internet Research*, 27, e68205. <https://doi.org/10.2196/68205>
24. Prato-Previde, E., Basso Ricci, E., & Colombo, E. S. (2022). The complexity of the human–animal bond: Empathy, attachment and anthropomorphism in human–animal relationships and animal hoarding. *Animals*, 12(20), 2835. <https://doi.org/10.3390/ani12202835>
25. Py_Team. (2020). *text2emotion* (Version 0.0.5) [Computer software]. PyPI. <https://pypi.org/project/text2emotion/>
26. Reddit, Inc. (2026). *Reddit reports second quarter 2026 results* [Press release]. https://s203.q4cdn.com/380862485/files/doc_financials/2026/q2/Q2-26-Earnings-Press-Release.pdf
27. Reynolds, B., & Seeger, M. W. (2005). Crisis and emergency risk communication as an integrative model. *Journal of Health Communication*, 10(1), 43–55. <https://doi.org/10.1080/10810730590904571>
28. Shanahan, E. A., Han, X., Salam, S., Washburn, S., & Slipher, S. K. (2026). One Health in the public mind: Narrative constructions of zoonotic spillovers. *One Health*, 22, 101411. <https://doi.org/10.1016/j.onehlt.2026.101411>
29. Sigalo, N., Awasthi, N., Abrar, S. M., & Frias-Martinez, V. (2023). Using COVID-19 Vaccine Attitudes on Twitter to Improve Vaccine Uptake Forecast Models in the United States: Infodemiology Study of Tweets. *JMIR*

infodemiology, 3, e43703. <https://doi.org/10.2196/43703>

30. Stokes, D. C., Andy, A., Guntuku, S. C., Ungar, L. H., & Merchant, R. M. (2020). Public priorities and concerns regarding COVID-19 in an online discussion forum: longitudinal topic modeling. *Journal of General Internal Medicine*, 35(7), 2244–2247. <https://pmc.ncbi.nlm.nih.gov/articles/PMC7217615/>
31. U.S. Food and Drug Administration. (2024). The dangers of raw milk: Unpasteurized milk can pose a serious health risk. <https://www.fda.gov/food/buy-store-serve-safe-food/dangers-raw-milk-unpasteurized-milk-can-pose-serious-health-risk>
32. U.S. Department of Agriculture Animal and Plant Health Inspection Service (USDA APHIS). (2024). *Federal Order: Highly Pathogenic Avian Influenza (HPAI) in Livestock*. <https://www.aphis.usda.gov/sites/default/files/20241205-federal-order-final.pdf>
33. U.S. Department of Agriculture Animal and Plant Health Inspection Service (USDA APHIS). (2026a). H5N1 Influenza. <https://www.aphis.usda.gov/h5n1-hpai>
34. U.S. Department of Agriculture Animal and Plant Health Inspection Service (USDA APHIS). (2026b). *HPAI Confirmed Cases in Livestock*. <https://www.aphis.usda.gov/livestock-poultry-disease/avian/avian-influenza/hpai-detections/hpai-confirmed-cases-livestock>
35. Vaiknoras, K. (2026). *Meat Price spreads*. U.S. Department of Agriculture. <https://www.ers.usda.gov/data-products/meat-price-spreads>
36. Van Bavel, J. J., Baicker, K., Boggio, P. S., Capraro, V., Cichocka, A., Cikara, M., Crockett, M. J., Crum, A. J., Douglas, K. M., Druckman, J. N., Drury, J., Dube, O., Ellemers, N., Finkel, E. J., Fowler, J. H., Gelfand, M., Han, S., Haslam, S. A., Jetten, J., ... Willer, R. (2020). Using social and behavioural science to support COVID-19 pandemic response. *Nature Human Behaviour*, 4(5), 460–471. <https://doi.org/10.1038/s41562-020-0884-z>
37. Vonk, J., & Bouma, E. M. C. (2024). Attachment as the catalyst for the attribution of complex cognition and emotion to companion cats. *Animals*, 14(14), 2123. <https://doi.org/10.3390/ani14142123>
38. Wenchang, X., Xuanhui, Y., Qun, Z., & Xiao, C. (2026). Health information anxiety in social media users during public health emergencies: A qualitative comparative analysis using attribution theory. *PLOS ONE*, 21(1), e0340674. <https://doi.org/10.1371/journal.pone.0340674>
39. Wilson, A. E., Lehmann, C. U., Saleh, S. N., Hanna, J., & Medford, R. J. (2021). Social media: A new tool for outbreak surveillance. *Antimicrobial Stewardship & Healthcare Epidemiology*, 1(1), e50. doi:10.1017/ash.2021.225
40. Xue, J., Chen, J., Hu, R., Chen, C., Zheng, C., Su, Y., & Zhu, T. (2020). Twitter discussions and emotions about the COVID-19 pandemic: Machine learning approach. *Journal of Medical Internet Research*, 22(11), e20550. <https://doi.org/10.2196/20550>
41. Yu, S., Eisenman, D., & Han, Z. (2021). Temporal dynamics of public emotions during the COVID-19 pandemic at the epicenter of the outbreak: Sentiment analysis of Weibo posts from Wuhan. *Journal of Medical Internet Research*, 23(3), e27078. <https://doi.org/10.2196/27078>

Figures

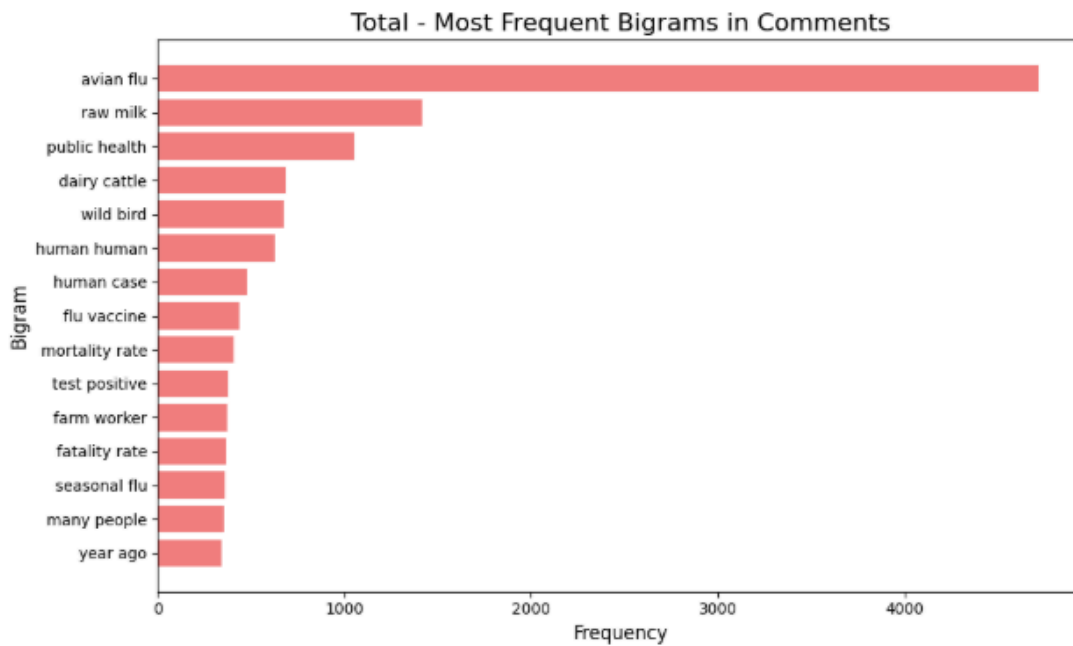


Figure 1

Horizontal bar chart of the 15 most frequent two-word sequences (bigrams) in the preprocessed Reddit comment corpus across the study window (January 2024 – December 2025; $n = 42,957$ comments).

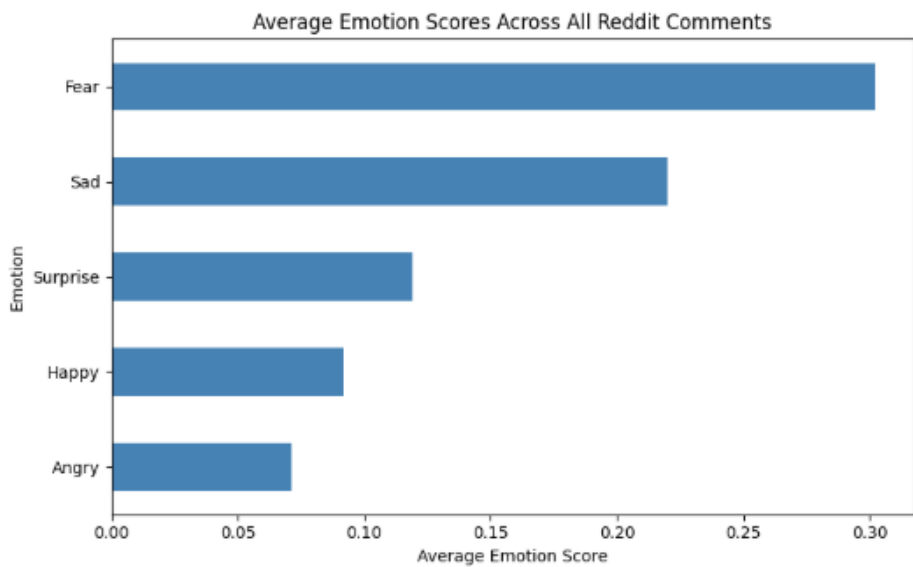


Figure 2

Average Text2Emotion intensity scores across all Reddit comments. Bars show the average Text2Emotion intensity score for each of the five emotion categories – fear, sadness, surprise, happiness, and anger – across the 42,957 Reddit comments in the study window (January 2024 – December 2025).

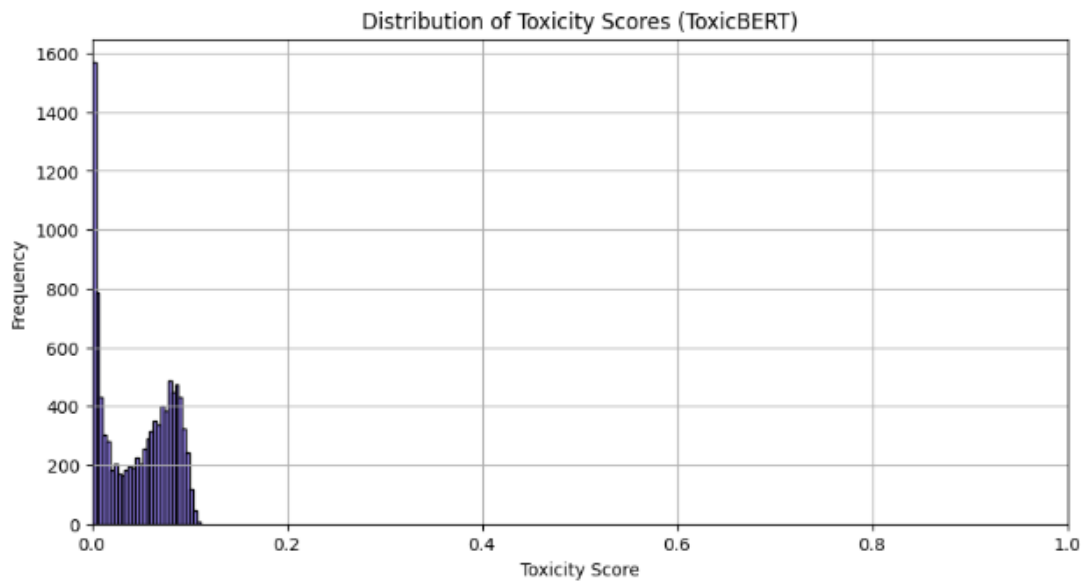


Figure 3

Distribution of ToxicBERT toxicity scores across all Reddit comments. Histogram of per-comment ToxicBERT toxicity scores across the 42,957 comments in the study window.

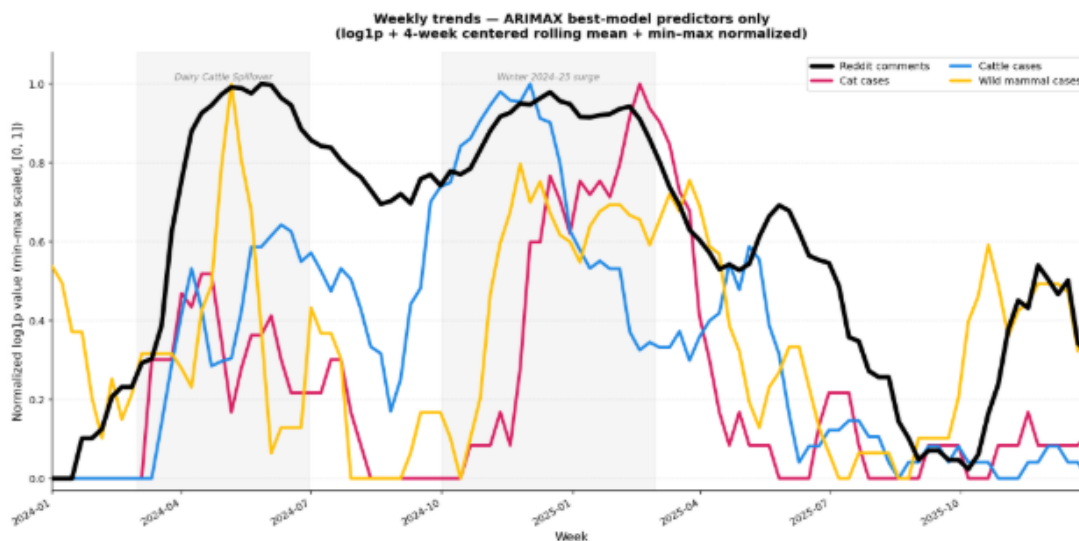


Figure 4

Weekly trends in Reddit comment volume (response variable) and the three ARIMAX exogenous predictors (cat, cattle, and wild mammal cases), January 2024 – December 2025. Each series was log1p-transformed, 4-week smoothed, and min-max scaled to [0, 1]. Shaded gray regions mark the two waves of Reddit engagement.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [LuizzaDallaPriaSupplementalMaterials.docx](#)