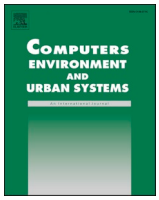
Computers, Environment and Urban Systems 100 (2023) 101922

Contents lists available at ScienceDirect

Computers, Environment and Urban Systems

journal homepage: www.elsevier.com/locate/ceus



# Plant pest invasions, as seen through news and social media

Laura G. Tateosian *, Ariel Saffer, Chelsey Walden-Schreiner, Makiko Shukunobe

*College of Natural Resources, Biltmore Hall 4008M — Campus Box 8004, North Carolina State University, 2800 Faucette Dr. Raleigh, NC 27695 USA*

ARTICLE INFO

*Keywords:*
Text mining
Invasive pests
Twitter
Online news
GDELT
Geospatial–temporal

ABSTRACT

Invasion by exotic pests into new geographic areas can cause major disturbances in forest and agricultural systems. Early response can greatly improve containment efforts, underscoring the importance of collecting up-to-date information about the locations where pest species are being observed. However, existing invasive species databases have limitations in both extent and rapidity. The spatial extent is limited by costs and there are delays between species establishment, official recording, and consolidation. Local online news outlets have the potential to provide supplemental spatial coverage worldwide and social media has the potential to provide direct observations and denser historical data for modeling. Gathering data from these online sources presents its own challenges and their potential contribution to historical tracking of pest invasions has not previously been tested. To this end, we examine the practical considerations for using three online aggregators, the Global Database of Events, Language and Tone (GDELT), Google News, and a commercial media listening platform, Brandwatch, to support pest biosurveillance. Using these tools, we investigate the presence and nature of cogent mentions of invasive species in these sources by conducting case studies of online news and Twitter excerpts regarding two invasive plant pests, Spotted Lanternfly and Tuta absoluta. Our results using past data demonstrate that online news and social media may provide valuable data streams to supplement official sources describing pest invasions.

## 1. Introduction

Biological invasions are instances where a species has been introduced, intentionally or unintentionally, to a new area where it successfully establishes and spreads. Invasive plant pests and pathogens, referred to collectively in this paper as pests, have devastating impacts on biodiversity, ecosystems, and human health, and cost billions in damages and control efforts (Pyšek et al., 2020; Seebens et al., Feb. 2017; Diagne et al., Apr. 2021; Savary et al., Mar. 2019). Although many pests are under regulatory control to prevent introductions and mitigate spread, the success of control and eradication efforts often hinges on early detection and rapid response (Martinez et al., Jan. 2020).

Rapid decision-making requires understanding how, where, and when pests species will arrive, establish, and spread, meaning that rich geospatial data and spatially explicit, temporally dynamic models are needed (Jones et al., June 2033). To provide the highest quality forecast possible, these models require both historical data about past pest invasions and up-to-date information about where pest species are being observed.

Pest observation and distribution data for these forecasting models commonly come from field observations, official reports, and genetic records consolidated through published literature and databases, collectively referred to in this paper as official records (Meentemeyer, Walden-Schreiner, Saffer, & Jones, 2021). While vital, these data can suffer from latency and spatial and temporal sparseness, due to the cost of collecting and collating these data at scale. Delays exist between species establishment and reporting in official datasets or peer-reviewed literature (Seebens et al., Feb. 2017; Seebens et al., 2015) and consolidation at the global scale by Plant Protection Organizations and scientific institutions introduces additional delays or incomplete data (Latombe et al., Sept. 2017; Latombe et al., 2019). These time lags have implications for the predictive capabilities of models and the success of control and eradication programs.

At the same time, there is growing evidence that Web media, such as online news and Twitter, could be a valuable source of supplementary data to boost coverage and tap into up-to-date information for modeling and predicting pest encroachments. Scientific approaches have been applied to analyze complex trends and make predictions from various social media platforms and other Internet publications to study spatio-temporal phenomena in human (Keller et al., May 2009; Lyon, Nunn,

* Corresponding author.
*E-mail address:* lgtateos@ncsu.edu (L.G. Tateosian).

https://doi.org/10.1016/j.compenvurbsys.2022.101922
Received 29 July 2022; Received in revised form 14 November 2022; Accepted 22 November 2022
Available online 28 December 2022
0198-9715/

Grossel, & Burgman, 2012; Lee, Kim, & Jang, 2018; Zhang, Ibaraki, & Schwartz, 2020) and animal disease epidemiology (Rabatel, Arsevska, & Roche,. 2019; Valentin, Lancelot, & Roche, 2021) and the social, natural, and physical sciences (Hawelka et al., May 2014; Sakaki, Okazaki, & Matsuo, 2010; Roxburgh et al., Jan. 2019). In fact, ecological researchers have codified the use of Web media as a source of indirect citizen science-like observations as iEcology. This approach formally extends the use of diverse, passively generated online data, like Tweets and news, to explore ecological questions (Jarić et al., July 2020). Early studies have found that Tweets can be an effective proxy for observations, especially in monitoring well-known invasive species (Hart, Carpenter, Hlustik-Smith, Reed, & Goodenough, 2018; Daume & Galaz, Mar. 2016). Similarly, Mammola et al., 2022 found that the timing and locations of news articles mentioning selected spiders matched the studied species' seasonal patterns of emergence and movement (Mammola et al., Dec. 2022). While less precise than direct measurement, timing and volume of Tweets and news explicitly mentioning a pest may provide a basis for establishing probabilistic timelines of presence in locations less represented in official records.

Given the proliferation of news aggregators constantly collecting worldwide data, online news offers potential for broad spatial coverage of both recent historical and breaking events. The uniform and information-dense format of Twitter posts ("Tweet"), proximity to firsthand observations, and the service's Application Programming Interface (API) access to raw data, also make it a likely data source for understanding both global and local trends. Though pursuing these data sources appears promising, the abundance, spatiotemporal coverage, and unstructured format of this data present unique challenges and opportunities. We explore these challenges and opportunities through two pest case studies (described in Figs. 1 and 2). We document our process along the way to detail specific hurdles and present a cost-benefit analysis describing the mechanics and potential of three online media aggregators for pest detection. We evaluate the results to address four research questions:

**Fig. 1. Local-scale case study pest.** The Spotted Lanternfly (Lycorma delicatula) is an insect pest native to China and Southeast Asia first found in the United States (US) in 2014 (cabi, 2022; Barringer et al., June 2015). In the years following its appearance in Berks County in southeastern Pennsylvania in September of 2014, Spotted Lanternfly has spread through much of Pennsylvania and into neighboring states. This pest travels predominantly as a "hitchhiker" on human movement of vehicles, stones, wood, and other items. Several sightings of the pest in states far from the original outbreak also suggest long-distance dispersal events (USDA APHIS 2022). The Spotted Lanternfly causes damage to several economically important crops (e.g., grapes, apples, cherries, and hops) and wood-producing trees (e.g., maples, oaks, poplars, walnuts, and willows). As a result, extensive management efforts have been enacted to prevent further spread and reduce existing populations of Spotted Lanternfly in the US (Jones et al., June 2033). Photo ©Rhododendrites - CC BY-SA 4.0. Inset: Spotted Lanternfly icon.

**Fig. 2. Global-scale case study pest.** Tuta absoluta (recently renamed *Phthorimaea absoluta*) is a moth and destructive tomato pest native to South America. The pest feeds on leaves and infests tomato fruits, dramatically reducing crop yields (Desneux et al., Aug. 2010). Upon arriving to Europe (first observed in Spain in 2006), Tuta absoluta spread throughout the Afro-Eurasian landmass and was reported in Morocco in 2008, South Africa in 2016, and India in 2017 (Biondi et al., 2018; Desneux et al., Aug. 2010; Mansour et al., Dec. 2018). Recent reports place Tuta absoluta in Eastern Europe and China (eppo, 2020). Tuta absoluta is now considered a major threat to tomato production at risk of reaching global distribution as it continues to spread through both trade and natural dispersal. Photo ©Patrick Clement/via Wikipedia - CC BY 2.0. Inset: Tuta absoluta icon.

1. How do results from different news aggregators (GDELT, Google News, Brandwatch) compare in volume and article selection?
2. How do results from different sources (news, Twitter) compare in volume and timing?
3. To what extent do the timing and volume of media posts describe the real-world movement of emerging invasive pests at the local (i.e., county to county) and global scales (i.e., country to country)?
4. What type of content useful for documenting historic and ongoing pest invasions appears in posts, and how does it differ across sources and pests?

Our goal is twofold: (1) to describe the potential for multiple Web media sources to provide content valuable to documenting pest spread, and (2) to help others who seek to track pests through Web media overcome the methodological hurdles associated with using text data from news and social media. Surveillance is key to tracking and controlling the devastating effects of pest invasions, making tapping into these novel data sources an important priority.

## 2. Method

While several API services provide searchable real-time news and media, past data (i.e., data more than a week old) typically have limited query options or are accessible only through fee-for-service platforms. We tested two low cost aggregators (GDELT and Google News) and one paid subscription service (Brandwatch) for collecting historical online news and Twitter data. Our methods are structured to address our research questions, following the workflow in Fig. 3. With the resulting posts, we compared the volume and timing across sources and aggregators, mapped locational data against official records, and characterized their text content.

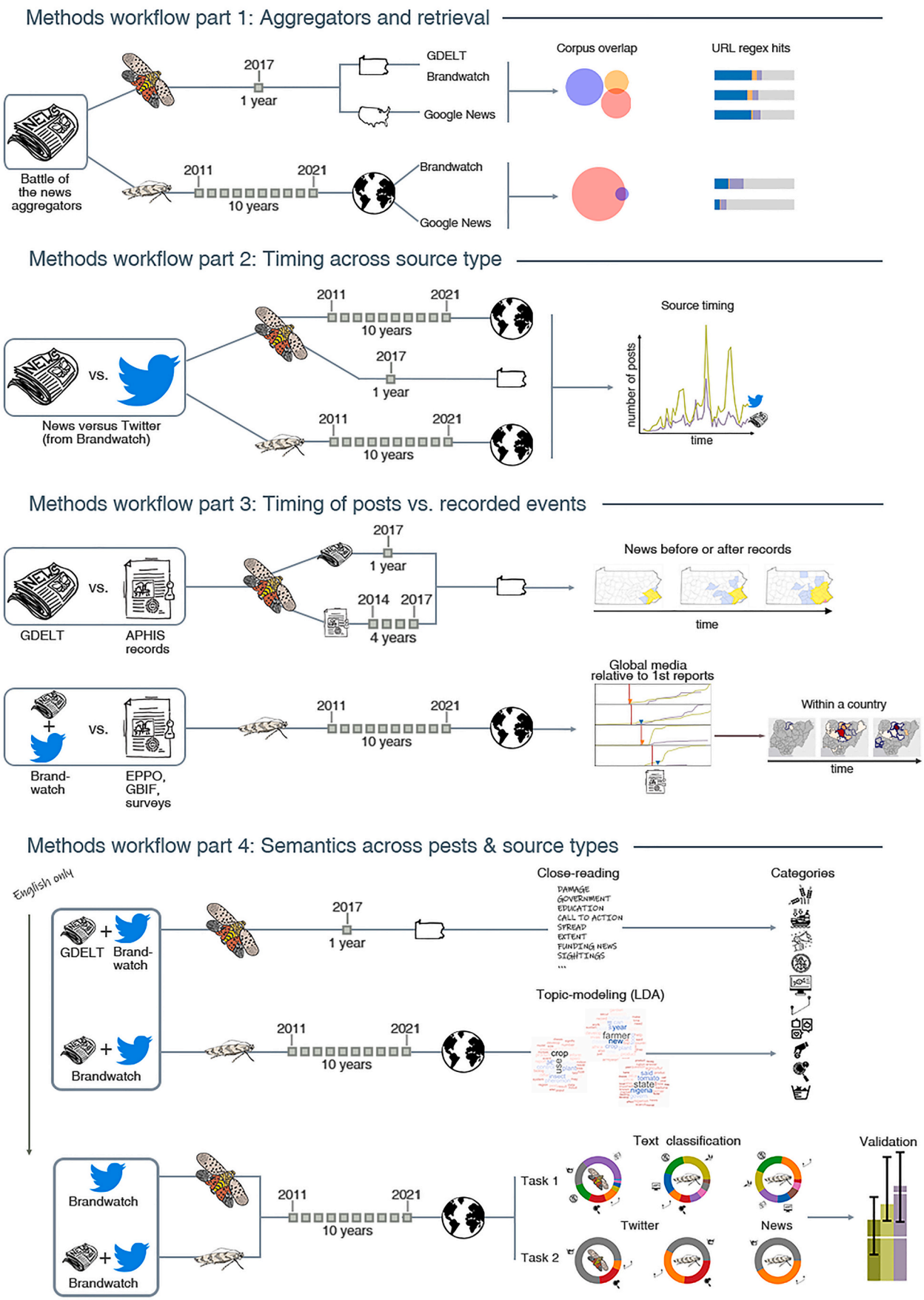



**Fig. 3.** Overview of our study design.

### 2.1. Aggregators and retrieval

We used GDELT, Google News, and Brandwatch to collect one year of Pennsylvania news about the Spotted Lanternfly. We also used Brandwatch to collect news and Twitter about Spotted Lanternfly from anywhere in the world, for 10 years, the maximum timespan available under our license. For Tuta absoluta, we used Google News to collect 10 years of news, and Brandwatch to collect 10 years of news and Twitter posts, both globally. Supplementary Table 1 summarizes our collection.

Within each source, we used Boolean queries to search for posts containing the pest's full scientific name and multi-language common names, provided by the Mediterranean Plant Protection Organization Global Database (EPPO-GD). This included 10 terms in English and French for Spotted Lanternfly and 27 terms in English, Dutch, Danish, German, Italian, Norwegian, Portuguese, Russian, Spanish, and Swedish for Tuta absoluta (see Supplementary Table 2 for the full list of terms). For each result, we recorded the URL, title, and additional data, as described below.

#### 2.1.1. Collecting news with GDELT

The Global Database of Events, Language, and Tone Project (GDELT) provides low-cost coverage of Web news from nearly every country in the world, collected every 15 min. GDELT also codes a set of events as a Global Knowledge Graph (GKG) and monitors broadcast news and online news images in addition to the text in news articles. Aspects of this data are curated in several databases and can be retrieved in different ways, depending on the target timespan. Current articles (from the past 15 min) can be downloaded directly from the GDELT website, a free API can be used to download recent data (from the last 7 days), and Google BigQuery can be used to filter past articles temporally, geographically, and based on attributes specified in the GKG Codebook (GKG, 2022). GDELT's original focus on geopolitical events meant that we were only able to take advantage of spatial and temporal filtering at the time of extraction. We used Google BigQuery to extract our dataset from the GDELT 2.0 Event database.

The Spotted Lanternfly infestation was first reported in 2014 in Berks County, Pennsylvania, and by September 2018 had spread to 13 other Pennsylvania counties and had begun to make its way into neighboring states, New Jersey and Virginia. Since the epicenter of the Spotted Lanternfly outbreak was in Pennsylvania during 2017, we used a SQL query in Google BigQuery (See Supplementary Table 2) to return URLs for all the 2017 news articles coded to Pennsylvania. We scraped the title, content, and publication date of each of these articles from the web using a Scala language implementation of the Goose article extractor and a key term search to select articles that mention the Spotted Lanternfly.

Querying URLs rather than article full text would reduce time and costs, as it avoids acquiring and processing the article content prior to determining if it is relevant. However, this approach risks missing articles that do not include relevant terms in the URL. To determine the hit rate with each approach, we tested pattern matching based solely on the URLs (from GDELT and the other two sources, to increase the sample), using regular expressions composed of the species names (listed in Supplementary Table 2), as well as general terms ("pests" and "invasive species") translated to the languages included in the original query.

#### 2.1.2. Collecting news with Google's News search

Our GDELT approach was suitable for a geographically and temporally limited pest, but not for the multi-continental spread of Tuta absoluta. We therefore explored Google keyword searches as a cost-effective alternative to quickly survey articles worldwide. Though Google search offers a paid API service, we opted for a free semi-automated approach. We searched Google for each query term, selected "News", set a custom 10-year date range, manually downloaded the results, and used Python to scrape the pages for the URL, publisher, title in the original language, and a short description of each article (~150 characters). We used the GoogleTrans Python package to identify language and translate the title to English. We assigned a country using the publisher's location based on the country code domain in the URL, where possible. To prototype an environment for human-in-the-loop classification of emerging pest events from news, we built an interactive tool with HTML, Javascript, Leaflet, and MapBox to display the count by country overlaid with per country GeoJSON choropleth of crop acreages from the United Nations Food and Agriculture Organization (FAO). We also queried Google for Spotted Lanternfly news in the United States in 2017 to compare the results obtained from different news aggregators.

#### 2.1.3. Collecting news and social media with Brandwatch

To gather more data to study our questions and compare low-cost news aggregators to a proprietary alternative, we used Brandwatch, a Web-based subscription service providing access to live and historical data from social media and news outlets. Keyword or semantic queries can be used to search the databases in selected languages within a time range and geographic extent. Brandwatch makes query results available for manual or API download, displays the results via configurable visualizations, and offers built-in AI analytic tools. For Twitter, Brandwatch returns Tweet content, the URL, and other metadata. For News, Brandwatch returns the URL, a relevant snippet of the article (~350 characters long), plus other metadata. For both content types, metadata may include a publication location determined by Brandwatch based on publisher geo-coordinates, profile location, timezone, domain, or geo-IP (Brandwatch, 2022). We tested Brandwatch's "Custom classifier" tool to classify text to user-defined categories (discussed in Section 2.4). To analyze post volume, all languages were included. For classification, posts were filtered to English-language only.

### 2.2. Timing across source type

Though Twitter has received much attention in the academic community, how pest invasions are communicated in informal social media posts relative to more formal news media is unknown. The volume of pest-related posts from both sources is a potentially valuable metric that may indicate concurrent real-world pest events. We first compared news and Twitter post volumes over time at multiple geographic scales using visual inspection of time-series and Spearman's rank correlation ($r_s$).

### 2.3. Timing of posts versus recorded events

We compared the timing and locations of posts with real-world events using official phytosanitary data describing local (Spotted Lanternfly) and global (Tuta absoluta) pest spread. For Spotted Lanternfly news from GDELT, we manually extracted place names where pest presence or quarantine was described and automated mapping using Pennsylvania municipality and county gazetteers. We compared these locations to point data of reported observations and field survey results from USDA APHIS from 2014–2017, aggregated to the county level (i.e., at least one confirmed Spotted Lanternfly observation within a county signified presence).

For the global Tuta absoluta case study, we evaluated an automated approach using publicly available phytosanitary data. The EPPO Global Database (EPPO-GD) was the most comprehensive source of Tuta absoluta reports between several global databases considered (CABI Invasive Species Compendium, FAO International Plant Protection Convention reports, EPPO Reporting Service) (cabi, 2022; fao, 2022; eppo, 2020). From the EPPO-GD Distribution pages, we scraped the earliest date that Tuta absoluta was reported present in each country ("First report date"). We also scraped the publication date of the earliest reference ("Reference date") as an indicator of when the report was made available in the database. We compared the timing of reports and references with posts from each country location (i.e., the Tweet location if provided or Twitter user's location; the news article or publisher's

location).

We further explored the state-level places mentioned in posts for Nigeria, the country with the most Twitter and news posts and a prominent outbreak in 2016. We downloaded all "Present" points for "Tuta absoluta (Meyrick, 1917)" from the Global Biodiversity Information Facility (GBIF) (gbif, 2022). We also extracted the timing of presence by state and year from surveys cited in the EPPO-GD. We compared both data sources with yearly state-level post locations and place names mentioned in posts.

### 2.4. Semantics across pests and source types

In addition to the volume of posts, we examined conversation content about pests and the variability of the discussions across platforms and pests. Machine learning-based text classification is an attractive approach for topic labeling to estimate the distribution of content categories (Ikonomakis, Kotsiantis, & Tampakas, 2005). Brandwatch's built-in text classification tool, Brightview (Firat, 2017), provides a convenient means of implementing this approach. Brightview has been used for comparable matters such as classifying agricultural concerns and monitoring conjunctivitis discussions (Ofori & El-Gayar, June 2021; Deiner et al., Feb. 2018).

#### 2.4.1. Category definition

Supervised text classification requires that topics be determined apriori. We identified a set of topics relevant to the invasive pest subject matter and present in the post content through a combination of close- (detailed manual review to extract relevant themes) and distant-reading (generative topic modeling to assign probabilistic topics from text).

- *Close-reading: Human-identified topics.* Two authors performed close-reading of the Spotted Lanternfly GDELT-extracted news articles and identified one to five themes for each article. During a first pass, each reader defined and assigned themes. A second pass was used to streamline the theme assignments once all the topics had been proposed. A similar process was performed on a small random selection of the Twitter posts from Brandwatch.
- *Distant-reading: Generative topic modeling.* We used Latent Dirichlet Allocation (LDA) to generate a set of topics from the Brandwatch-extracted Tuta absoluta news and Twitter posts. LDA uses Bayesian hierarchical modeling to generate probable topics from word frequency (described by stems, or word roots) within each document, but ignores word order and the semantic role of words within phrases (Blei, Ng, & Jordan, 2003). We ran separate LDA models on news and Twitter, limited to posts in English. We modeled 20 topics for each source, using a Dirichlet prior alpha parameter of 0.2 to encourage the concentration of probability mass in fewer topics per document. We then grouped the resulting topics thematically by their relationship to concepts relevant to the study of invasive pests.

Through discussion between two reviewers, we aligned the themes generated with close- and distant-reading into "Cross-pest themes" relevant to understanding conversation around the local and global spread of invasive pests (Table 1).

#### 2.4.2. Text classification

We considered each Cross-pest theme as a category for supervised classification of the larger news snippet and Tweet datasets. To facilitate consistent coding between reviewers, each category was defined and keywords identified (Supplementary Table 3).

To classify data from our two sources and two species, we trained and applied separate supervised Brightview models to each: Tuta absoluta news, Tuta absoluta Tweets, and Spotted Lanternfly Tweets. We located relevant training snippets with keywords. We avoided selecting duplicate or overly similar training snippets to sustain diversity and prevent overfitting. Only single-topic snippets were accepted for training. We developed a training set for each corpus and applied the Brightview classification algorithm to the full population of posts for two classification tasks:

- Task 1: Classify content across all thematic categories.
  - –Per Brandwatch's guidelines, a single reviewer coded 10–30 training snippets to each category (Spread/extent, Direct sighting, Management/control, Damage/costs, Government action/biosecurity, Public awareness, Research, Funding, Reactions to control, Other). If sufficient training snippets were not found for a given category, it was excluded for that case study.
- Task 2: Classify location-information relevant categories
  - –To further focus the algorithm on potential pest observations and distribution data, we trained the classifier for each species with only three categories: Spread/extent, Direct sighting, and Other. To represent the diverse content that could appear in "Other", we sampled evenly from each remaining category from Task 1. Three reviewers cross-validated the training coding to ensure that category definitions were sufficiently delineated. For each category, we trained the classifier with 30 snippets with agreement across all three reviewers.

#### 2.4.3. Validation and cross-verification

To evaluate the algorithm, two reviewers cross-validated a sample of each corpus. Before sampling the data, we removed Retweets and duplicated content (after stripping URLs). As the Spotted Lanternfly corpus was much larger than the Tuta absoluta corpora, we sampled 1% of the Spotted Lanternfly posts and 5% of Tuta absoluta posts (307 Spotted Lanternfly Tweets, 163 Tuta absoluta Tweets, and 190 news articles, for each task). Because posts can contain multiple topics, each reviewer was allowed to assign up to 2 categories to a post (the Brightview classifier, however, assigns only one category per post). We calculated inter-reviewer agreement (Krippendorf's alpha, $\alpha_{\text{INTER-REVIEWER}}$) as agreement on either of the two categories selected between reviewers (i.e., 2 out of up to 4 reviewer-selected categories agreed) for a given post, and algorithm classification accuracy (with both $\alpha_{\text{CLASSIFIER}}$ and Macro F1 scores). We chose to represent classification accuracy with the Macro F1-score as it gives equal importance to all classes, and therefore better captures performance on rare classes. We considered agreement to be a positive match with any of up to four reviewer-selected categories. We summarized these metrics for each category, task, and case study.

## 3. Results

### 3.1. Aggregators and retrieval

#### 3.1.1. Articles and Twitter posts retrieved from aggregators

Twitter provided the most results, returning 76,764 Tweets mentioning Spotted Lanternfly and 17,604 Tweets mentioning Tuta absoluta, both over 10 years. Across comparable news queries, Brandwatch provided the greatest number of articles, most notably over the larger time period and geographic scope of the Tuta absoluta case study (14,601 vs. 477 from Google News for Tuta absoluta and 52 vs. 32 from GDELT for Spotted Lanternfly in Pennsylvania in 2017). Brandwatch also captured content in more languages (57 vs. 20 for Google News) in the global case study. Nearly all (93.6%) news articles returned by Brandwatch included a publication location at least at the country scale, but few (4.6% for Spotted Lanternfly, 0.4% for Tuta absoluta) included a subnational location (e.g., city or state). Country information was less complete for posts from Twitter (69.9% for Spotted Lanternfly, 63.8% for Tuta absoluta), but included more subnational location data than news (41.4% for Spotted Lanternfly, 40.4% for Tuta absoluta). The Brandwatch global queries for Tuta absoluta returned news results in 57 languages from 142 countries and Twitter results in 42 languages from 131 countries. These results are summarized in Supplementary Table 4.

**Table 1**
Supervised topics constructed through close- and distant-reading approaches.

| Distant reading<br>LDA-generated topics (Tuta absoluta) | Cross-pest themes | Close reading<br>Human-identified topics (Spotted lanternfly) |
|---|---|---|
| Twitter:<br>tomato destroy scarciti farm return (7.5%)<br>caus diseas scarciti absoluta (6.5%) | **Damage/cost**<br>(Twitter LDA: 14.0%,<br>News LDA: 23.1%) | CROP DAMAGE direct discussion of pest impacts on agricultural production (Twitter) |
| News:<br>state kaduna nigeria farm price (7.1%)<br>read nigerian juli devast farmer (3.5%)<br>known scarciti tomato govern nigeria (9.5%)<br>year tomato price product farm (7.2%)<br>crop loss caus estim billion (3%)<br>loss huge farmer pest market (2.3%) | | DAMAGE specific descriptions of types of damage caused by the pest (News) |
| Twitter:<br>expert feder govern said (5.4%)<br>tomato import ban due effect (4.6%) | **Government action/biosecurity**<br>(Twitter LDA: 10.0%,<br>News LDA: 9.5%) | GOVERNMENT government actions, rulings, etc. (News) |
| News:<br>known scarciti tomato govern nigeria (9.5%) | | |
| Twitter:<br>can crop farm will need (5.3%)<br>pest plant crop armyworm african (4.9%)<br>life tomato day crop agricultur (3.5%)<br>tomato almost nigeria harvest someon (3.2%) | **Public awareness**<br>(Twitter LDA: 16.9%,<br>News LDA: 22.4%) | POTENTIAL NEW LOCATIONS preemptive alerts of places the pest may be at risk of appearing (Twitter) |
| | | INFORMATIVE general information about the pest (Twitter) |
| News:<br>diseas fall pest armyworm food (7.5%)<br>tomato south import american product (4.5%)<br>agricultur food outbreak said import (3%)<br>farmer fall pest armyworm diseas (2.5%)<br>tomato farmer can per last (4.9%) | | AG farmer's meeting or news (News) |
| | | EDUCATION opportunity/project to educate the community about the pest or pests (News) |
| Twitter:<br>control new use biolog ipm (5.8%)<br>manag control ipm sustain koppertkenya (4.5%)<br>control icip loss caus well (4%) | **Management/Control**<br>(Twitter LDA: 14.3%,<br>News LDA: 22.0%) | CALL TO ACTION direct instructions to act on to control or report the pest (Twitter) |
| | | CONTROL control measures that readers can take or that are being taken by authorities or farmers (News) |
| News:<br>crop includ field mani develop (3.7%)<br>manag develop sustain program relat (2.7%)<br>use control pest product crop (7.4%)<br>control use insect lepidoptera meyrick (8.2%) | | SPECIMEN how to collect and submit pest, if found (News) |
| Twitter:<br>eurekamag develop temperatur scrobipalpuloid influenc (5.2%)<br>lepidoptera gelechiida control time anabagail (5.1%)<br>diseas solut expert proffer work (5.0%)<br>eurekamag effect bacillus thuringiensi berl (4.6%)<br>develop research call remedi abu (4.1%)<br>nigeria solut grown home (4.0%) | **Research**<br>(Twitter LDA: 28.0%,<br>News LDA: 7.2%) | RESEARCH research initiatives or stories about researchers (News) |
| News:<br>crop attack veget pheromon pea (3.3%)<br>nigeria institut said research import (3.9%) | | |
| Twitter:<br>year last outbreak agric cchukudebelu (5.8%)<br>south africa first report outbreak (5.7%)<br>invas manag food spread help (5.4%) | **Spread/extent**<br>(Twitter LDA: 16.9%,<br>News LDA: 15.8%) | SPREAD describing the presence of a pest in a new |
| News:<br>south american america spread africa (7.9%)<br>state plant said kano farm (5.7%)<br>nigeria state say kaduna (2.2%) | | EXTENT lists geographic locations which have been quarantined or where the pest has been sighted (News and Twitter) |
| | **Reactions to control** | COLLATERAL DAMAGE unintentional secondary effects and people's reactions to pest control measures (Twitter) |
| | **Funding** | FUNDING NEWS money set aside to control pest (News and Twitter) |
| | **Direct sighting** | SIGHTING direct, first-person observations of the pest (Twitter) |
| | **Other** | GENERAL the pest of interest is mentioned but the content is off-topic, not primarily about pests or pest of interest only one of many discussed (News and Twitter) |

We compare the coverage of content across aggregators in Fig. 4, which shows the overlap in the URLs, article titles, and web domains resulting from each query. In the SLF case study, the results from each aggregator were largely mutually exclusive. There was no overlap between the URLs returned from GDELT and Google News, despite both being news aggregators managed by Google, while GDELT and Brandwatch provided 5 of the same URLs (out of 79 unique URLs, 6.33%). We also compared article titles, as news outlets may repeat articles with distinct URLs or republish same-titled articles from other sources. There were fewer distinct titles than URLs in both case studies (96.23% for SLF, 71.88% for Tuta absoluta). While URLs point to specific articles, domains refer to the article's outlet or publisher, describing the population of sites curated by each aggregator. For Spotted Lanternfly, the domains were still predominantly (86.75%) mutually exclusive. For Tuta absoluta, Brandwatch captured the majority (79.3%) of domains included by Google News, along with 15.4 times as many additional domains not captured in Google News.

### 3.1.2. Filtering by URL and by content

The full queries for the three aggregators relied on searching the article contents for regular expression matches. When the same queries were applied to only the URLs and not the article contents, fewer articles were matched, though coverage varied between the two case studies and across the aggregators.

Results are shown in Fig. 5. For Spotted Lanternfly, species name terms were matched in 48.08% of Brandwatch URLs (25 out of 52), 47.5% of Google News URLs (38 out of 80), and 53.12% of GDELT URLs (17 out of 32). With general terms ("pest" and "invasive species" in multiple languages), an additional 9.62% of Brandwatch (5 out of 52), 7.5% of Google News (6 out of 80) and 6.25% of GDELT (2 out of 32) URLs were matched.

For Tuta absoluta, URL filtering captured significantly fewer articles. Species name terms matched 7.47% of Brandwatch URLs (1,090 out of 14,601), and 18.87% of Google News URLs (90 out of 477). General terms matched an additional 7.64% of Brandwatch (1,116 of 14,601) and 17.99% of Google News (85 of 477) URLs.

### 3.1.3. Pest news explorer

To engage stakeholders and enable experts to explore articles, we created a working prototype for browsing a collection of pest articles once they have been identified. The user selects a pest name from the drop-down menu to view related articles and an overview of hits on the map. Fig. 6 displays Tuta absoluta articles extracted from Google News. In the left column, each listing shows the article title and other metadata. The user can click on a title to open the full article on the Web. The bubble map shows article counts by country, with bubble size proportional to count. The country colors can be used to display pertinent information. In this example, the choropleth shows harvest acreages for tomatoes, a crop vulnerable to Tuta absoluta. The viewer can zoom and pan the map.

## 3.2. Timing across source type

Time series graphs and high positive $r_s$ values demonstrate correlation over time between the volume of news and Twitter posts extracted with Brandwatch. We examined this relationship at a global scale over 10 years and local scale (state-level, Pennsylvania) during 2017 for Spotted Lanternfly (Fig. 7a), and at the country scale over 10 years for Tuta absoluta (Fig. 7b). Twitter volume about Spotted Lanternfly was consistently higher than news, amplifying similar peaks in the data at both scales (global SLF $r_s = 0.93$ and PA SLF $r_s = 0.8$). The relationship between Tuta absoluta post timing from the two sources was more variable. Correlation was high for countries with the highest overall post volume, but neither news nor Twitter post volume was consistently higher between countries (Fig. 7b). For Tuta absoluta, we further explored the geographic distribution and relationship with post volume and known pest presence in Fig. 8. Countries with high volumes of posts had high correlation between the timing of Twitter and News posts, however total post volume and correlation did not depend on whether the country had reported the pest as present.

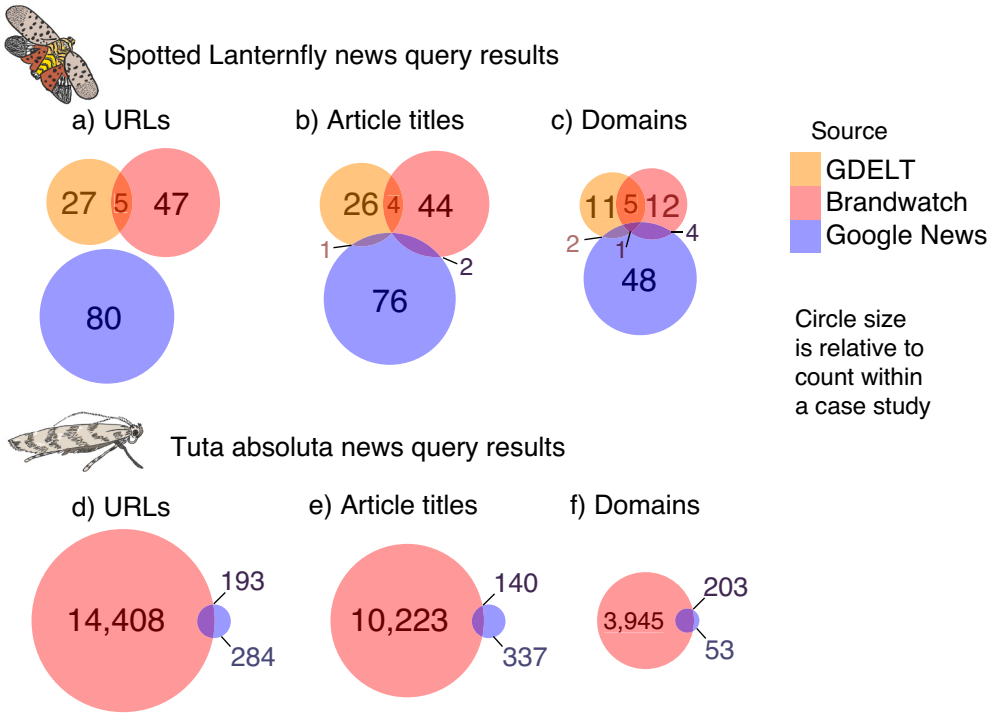



**Fig. 4.** Overlap of (a, d) article URLs, (b, e) titles, and (c, f) domains extracted from GDELT, Google News, and Brandwatch for Spotted Lanternfly (a–c) and Tuta absoluta (d–f) using the queries described in Supplementary Tables 2 and 3. Though comparable queries were used, the returned articles and their domains were mostly unique to each aggregator, suggesting differing coverage of both global and local news sources.

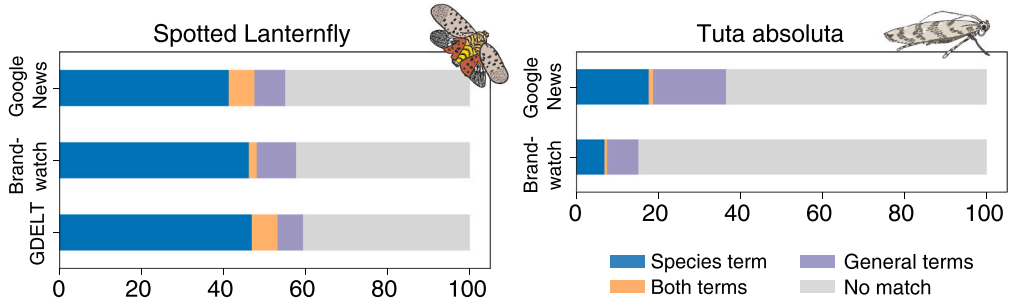



**Fig. 5.** Comparing the article URL regular expression matches to article content regular expression matches. Color bar length indicates the percentage of articles matched by querying only the article URL for species-specific and general pest terms, for each species and aggregator considered. The full bar (i.e., 100%) represents articles matched using similar queries of the full article content.

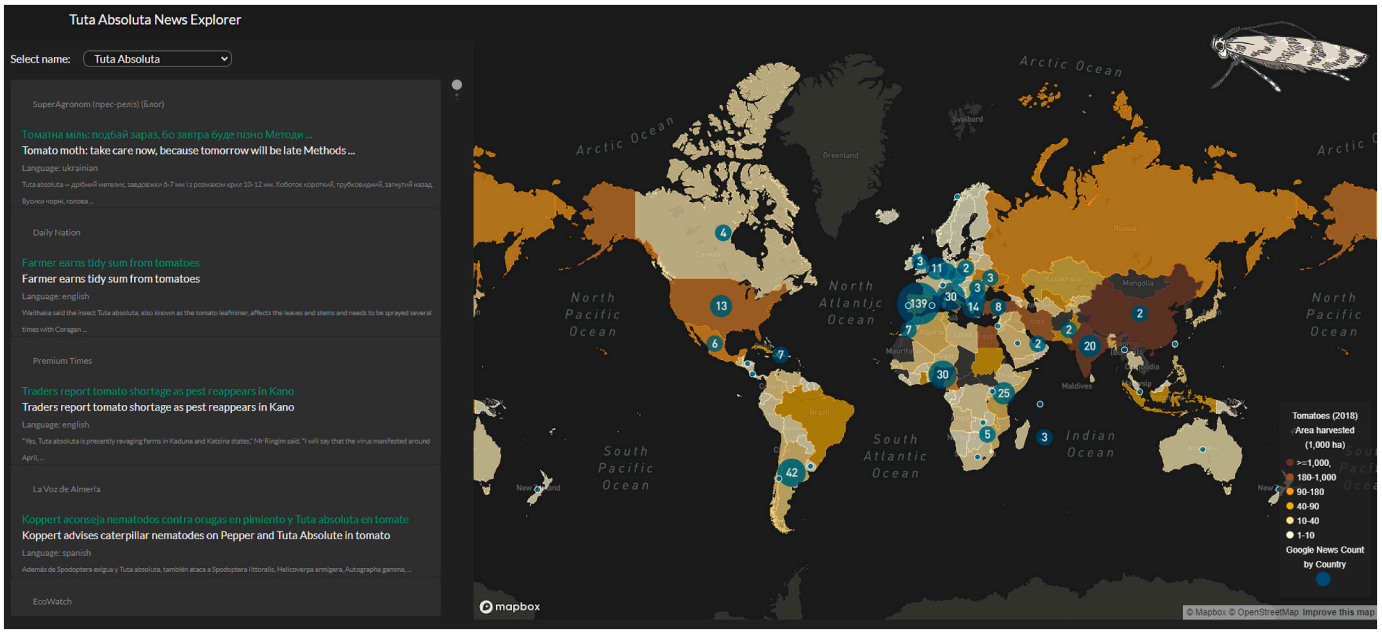



**Fig. 6.** Dashboard for pest-related news. Articles are listed on the left. The map overlays article counts on countries colored by tomato crop harvest acreage.

### 3.3. Timing of posts versus recorded events

#### 3.3.1. Spotted Lanternfly

Of the 32 articles returned by the GDELT 2017 Pennsylvania Spotted Lanternfly news query, 18 mentioned locations of Spotted Lanternfly events, such as quarantine orders and damage. Location names were specified on three administrative levels, state, county, and municipality (cities, towns, townships, or boroughs). We mapped these by article, to see the progression of mentions, and by month to compare them to official records.

Maps of the 18 articles and their publication dates are shown in Fig. 9. The first map is based on the following content in a January 23, 2017 article: "Union and Robeson townships, as well as the borough of Birdsboro, are among the newest municipalities added to Pennsylvania's spotted lanternfly quarantine list. The spotted lanternfly […] has been located in 17 Berks County municipalities […], as well as in a number of Chester and Montgomery county municipalities." Newly added municipalities are shown in dark pink (Union, Robenson, and Birdsboro in the 01/23 map). Municipalities where presence or quarantine is repeated in the current article or already mentioned in earlier articles are royal blue (Union, Robenson, and Birdsboro in the 02/01 map). Counties are marked in pink for new SLF event mentions in that county, as in Berks county in the 01/23 map or Bucks in the 02/01 map. Chester county is also pink in the 02/01 map because the article newly mentions additional municipalities therein in connection with SLF. Counties with previously accumulated events and no newly added municipalities are pale blue. Strictly blue maps represent articles that do not add new information to the accumulated mentions. Fig. 9 includes five identical winter maps from November and December articles (11/05, 11/08, 11/13, 11/14, and 12/26) in which no municipalities or counties were newly mentioned. All mentions were in Pennsylvania, except the final article (12/29) mentioned the state of New York and a county in Delaware. At the county and state levels, the mapped mentions are contiguous over time, but not at the municipality level.

Fig. 10 shows a cumulative monthly county level comparison of USDA APHIS Spotted Lanternfly Pennsylvania county presence records to the 2017 GDELT news articles. Month 1 shows the APHIS presence records accumulated from 2014 through the first month of 2017 and any January 2017 news article mentions. At this point, Bucks, Lehigh, and Northampton counties have APHIS records but no news mentions, while

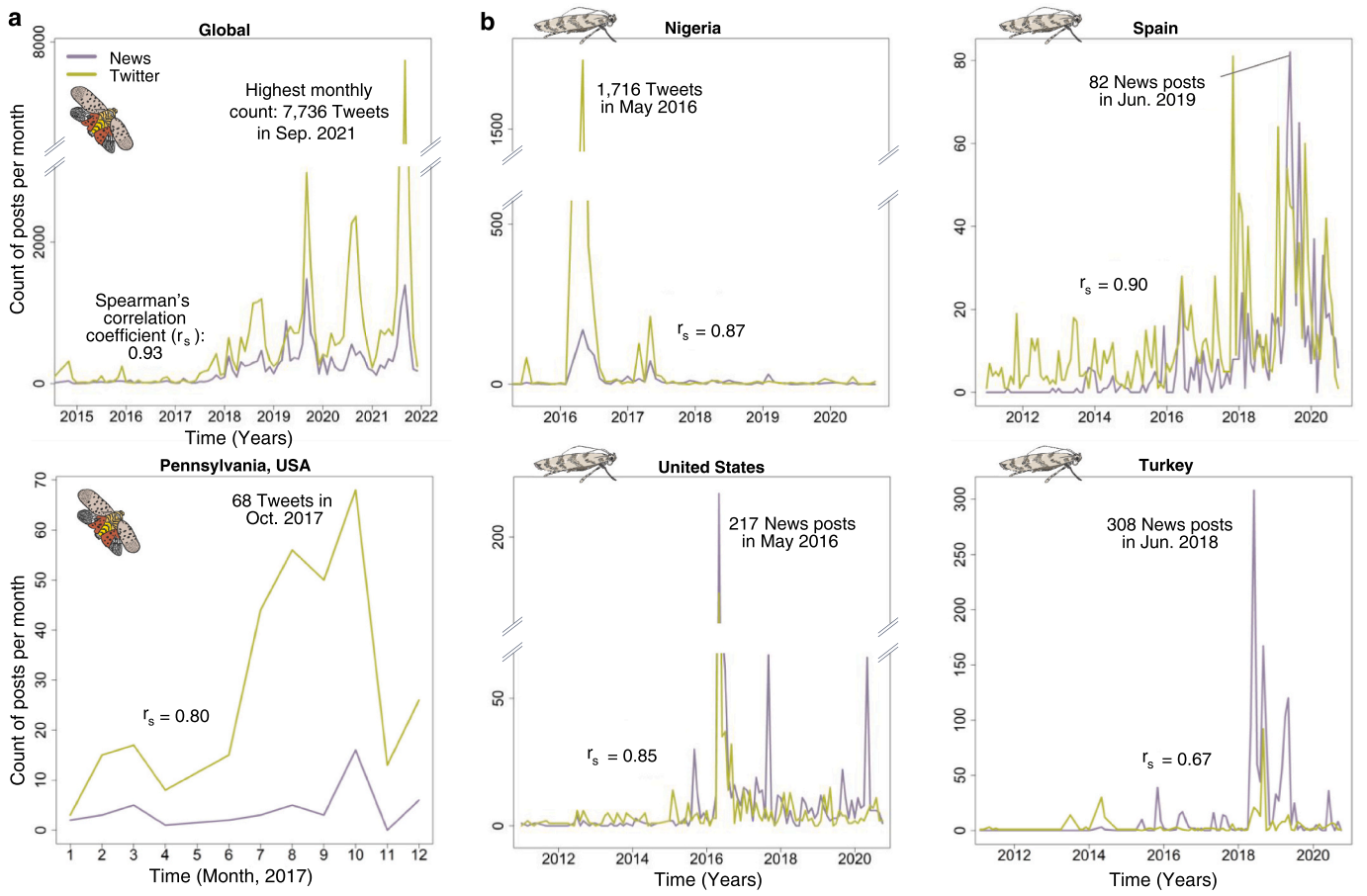



**Fig. 7.** Timing of news and Twitter posts acquired with Brandwatch for (a) Spotted Lanternfly and (b) Tuta absoluta. In (a), plots are shown for the global multi-year query (top) and for Pennsylvania, USA in 2017 (bottom). In (b), plots are shown for the four countries with the highest overall post volume.

Berks, Chester, and Montgomery counties appear in both sources. Initially, only adjacent county records are added and Luzerne County is the first non-adjacent county with a report. At the county level, news reports of quarantines generally capture a subset of the APHIS records. The only pink county appears in month 10, when the SLF quarantine in Delaware County was mentioned in an October 23 article published in 'The Intelligencer', a newspaper from Doylestown, in Montgomery County Pennsylvania. In the Jan. 2016–Dec. 2017 APHIS records, the first Delaware County record was in Nov. 2017.

### 3.3.2. *Tuta absoluta*

We extracted 93 dated "First reports" of Tuta absoluta from the EPPO-GD, and evaluated the timing of these reports and their references, in comparison with news and social media. The EPPO consolidates these global reports from multiple primary and secondary sources. Of these reports, the earliest referenced source was most often a journal publication (59), followed by other Plant Protection Organizations (PPOs, 26), and 9 from Internet sources including Facebook groups, pest-specific web pages, government websites, and news providers. The type of source varied geographically (Fig. 11a). Most first reports (54 of 93) from all three sources included a dated reference within the same year (e.g. "First found in May 2016 on tomato plants", reference dated 2016; Fig. 11b). First reference timing ranged from 0–1 years for PPOs, 0–2 years for Internet sources, and 0–6 years for Journal Publications.

In many countries (e.g., Senegal, Tanzania, Nigeria), news and Twitter posts about Tuta absoluta began shortly after the First report and First reference dates in EPPO (Fig. 12a). In other cases (e.g., Nepal, India), posts began near the time of the First report but before the First dated reference, or before the First report (e.g., Ghana, Mauritius, Armenia). We further considered the places mentioned in posts. For Nigeria, pattern matching for country and state names returned 1,176 news and 2,180 Twitter posts, with timing similar to the posts originating from the country. The main country origins of these posts were Nigeria, the US, UK, and South Africa (Fig. 12c), though varied by source.

Using place names mentioned in posts and Twitter post locations, we further evaluated sub-national temporal data. GBIF included observations of Tuta absoluta in 31 of the 93 countries reported in EPPO-GD, 58.7% with geo-location. Additionally, periodic EPPO reports provide temporal and subnational information (there are 77 reports for the 93 countries listed). However, neither source contained additional information about the sub-national spread of Tuta absoluta in Nigeria. We therefore extracted observations from the publication cited by EPPO-GD: state-level presence data from a 2017–2018 survey (Aigbedion-Atalor et al., Dec. 2019) and a 2016 survey (Borisade, Kolawole, Adebo, & Uwaidem, 2017) cited in the previous paper. Both referred to Katsina State as the location of the first observed occurrence of Tuta absoluta in 2015.

Mentions in news and Twitter (Fig. 13a and b) were more spatially and temporally consistent with survey locations with Tuta absoluta detections than the source of the post (Fig. 13c). State mentions in news (a) captured all detected states, which by volume included 81% and 89% of state mentions in this source in 2016 and 2017–2018, respectively. An additional 16 states and 7 states were mentioned in news in 2016 and

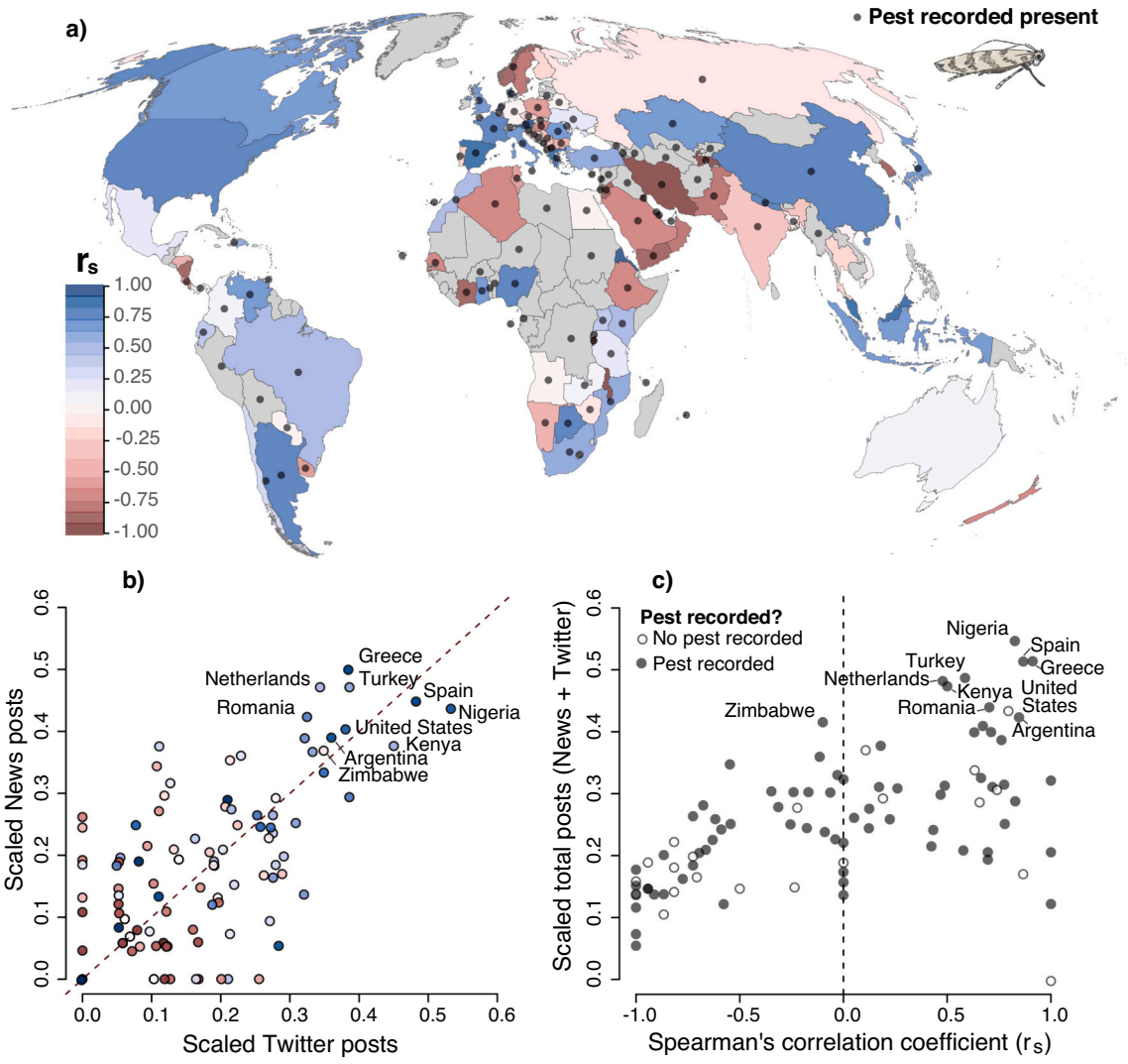



**Fig. 8.** Correlation between the time-series of news and Twitter post volume by country. (a) and (b) show $r_s$ values as fill color (blues indicate high correlation). Black dots in (a) and (c) indicate countries where the pest was reported present. In (b) and (c), the log of post volume is scaled by the log of the number of internet users per country, according to the World Bank Open Data in 2020. Correlation was not calculated for countries with posts from only one of the two sources; these countries are colored gray in (a).

2017–2018, respectively. Twitter mentions (b) captured 6 of 7 detected states in 2016 and 5 of 9 in 2017–2018 (93% and 78% of state mentions by volume, respectively) and mentioned an additional 8 and 3 states. Twitter posts (c) in 2016 had sources in 6 of 7 detected states, but also from an additional 26 states, with only 8% of posts coming from states with detections. Twitter post sources in 2017–2018 included 8 of 9 detected states, which included 68% of the post volume, with posts from 9 additional states. None of the three means (a, b, c) captured the first observed location in 2015, although 3 additional states were mentioned in news, 1 in Twitter, and Tweets came from 9 states during that year.

### 3.4. Semantics across pests and source type

#### 3.4.1. Category definition

Close-reading of Spotted Lanternfly news and Twitter posts resulted in 17 topics. The number of topics identified within an article by the two reviewers ranged from 1 to 6 (median = 3 topics per article). The LDA-generated topics (40 across both sources for Tuta absoluta) were manually grouped based on thematic similarities. Both sets of topics were then aligned and labeled, producing 10 themes spanning both case studies ("Cross-pest themes" in Table 1).

#### 3.4.2. Text classification and validation

The detailed classification results by category for Task 1 (T1) and Task 2 (T2) are shown in Supplementary Fig. 2. Categories with insufficient training snippets available were excluded (T1: "Government action/biosecurity" for Spotted Lanternfly Twitter, "Funding" for both Tuta absoluta sources, "Other" for Tuta absoluta news; T2: "Direct sighting" for Tuta absoluta news).

Discussion of crop loss (Damage/cost) was classified as more prevalent in conversations around Tuta absoluta than Spotted Lanternfly, while Public awareness was a dominant theme for Spotted Lanternfly. More Twitter posts were classified as "Other" for Spotted Lanternfly than for Tuta absoluta (17.7% vs. 5.6%). Classification volume was low (less than 10%) for "Funding", "Government action/biosecurity", and "Reactions to control".

In T2, "Other" included the majority of posts, previously captured across other thematic categories. The algorithm assigned more posts to locational categories ("Spread/extent" and "Direct sighting") for Tuta absoluta (42.3–57.4%) than Spotted Lanternfly (23.8%).

To validate the algorithm, we calculated agreement and accuracy statistics $\alpha_{\text{INTER-REVIEWER}}$, $\alpha_{\text{CLASSIFIER}}$, and Macro $F1_{\text{DIFF}}$ (see Table 2). For $\alpha_{\text{INTER-REVIEWER}}$ and $\alpha_{\text{CLASSIFIER}}$, a value of one indicates perfect

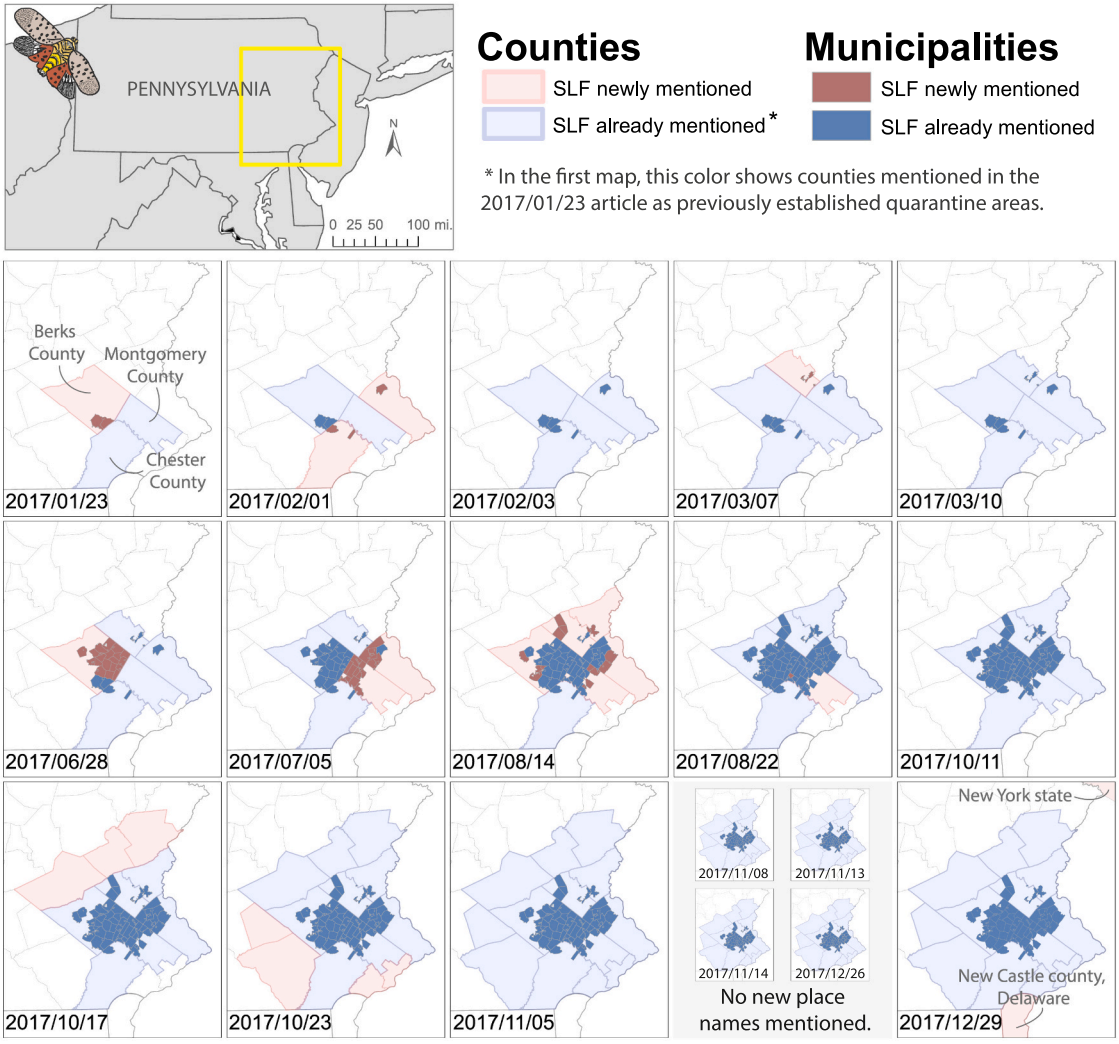



**Fig. 9.** Counties and municipalities in southeastern Pennsylvania mapped by mentions in the 2017 Spotted Lanternfly (SLF) news article collection from GDELT. Each map pertains to one article published on the date shown. Counties turn pink if they are newly mentioned or if a contained municipality is newly mentioned. Counties are blue if they are previously mentioned and contain no newly mentioned municipalities. Municipalities are dark pink if they are newly mentioned in this corpus and then dark blue in subsequent maps.

reliability, zero indicates absence of reliability, and values less than zero indicate systematic disagreement. Since the number of classes varied from 2 to 9, we calculated Macro $F1_{DIFF}$ as the difference between the actual validation Macro F1 and the Macro F1 score expected at random, given the number of classes ($F1_{ACT}$ - $F1_{RAND}$ = Macro $F1_{DIFF}$).

While not directly comparable across tasks, we provide the summary of classification metrics in Table 2. Classification accuracy derived from the validation sample was low across both tasks and case studies, with some variability by task and metric ($F1_{ACT}$ = 0.49–0.78, Macro $F1_{DIFF}$ = 0.26–0.38, $\alpha_{CLASSIFIER}$ = 0.41–0.56). In Fig. 14, the F1 scores describe the performance on each category, and the error bars indicate category over- and under-prediction by the algorithm, according to the reviewers. For instance, in T2 many Tuta absoluta Twitter posts classified as "Direct sighting" were coded as "Other" by reviewers, while for Spotted Lanternfly, many posts classified as "Other" were coded "Direct sighting" or "Spread/extent" by reviewers.

## 4. Discussion

The goal of this study is to examine news and Twitter for their pest-monitoring potential and to provide practical guidance about gathering this type of online data. Our results show that both news and Twitter contain large volumes of pest-related data and our samples contain timely useful information about Spotted Lanternfly at the local extent and Tuta absoluta at the global extent. The discussion includes lessons about data retrieval and evaluation of our research questions.

### 4.1. Aggregators and retrieval

In our case studies, all three aggregators (GDELT, Google News, Brandwatch) proved viable for accessing historical news and in the case of Brandwatch, for accessing historical Twitter posts. Sources differed in cost and effort to acquire data and in the volume of results returned. We share our general observations about working with each aggregator in Table 3.

For news, Brandwatch produced more results for comparable queries, and required less data retrieval and post-processing effort. The service also provides useful attributes like location and language that must otherwise be derived. This ease-of-use is a trade-off with the cost to access data and the limited historical time period of data provided under most licenses. Google News was simpler to search and download than GDELT, but our process included some manual effort to conduct the searches and prepare the returned results for extraction. GDELT data extraction effort could be reduced substantially by querying URLs rather

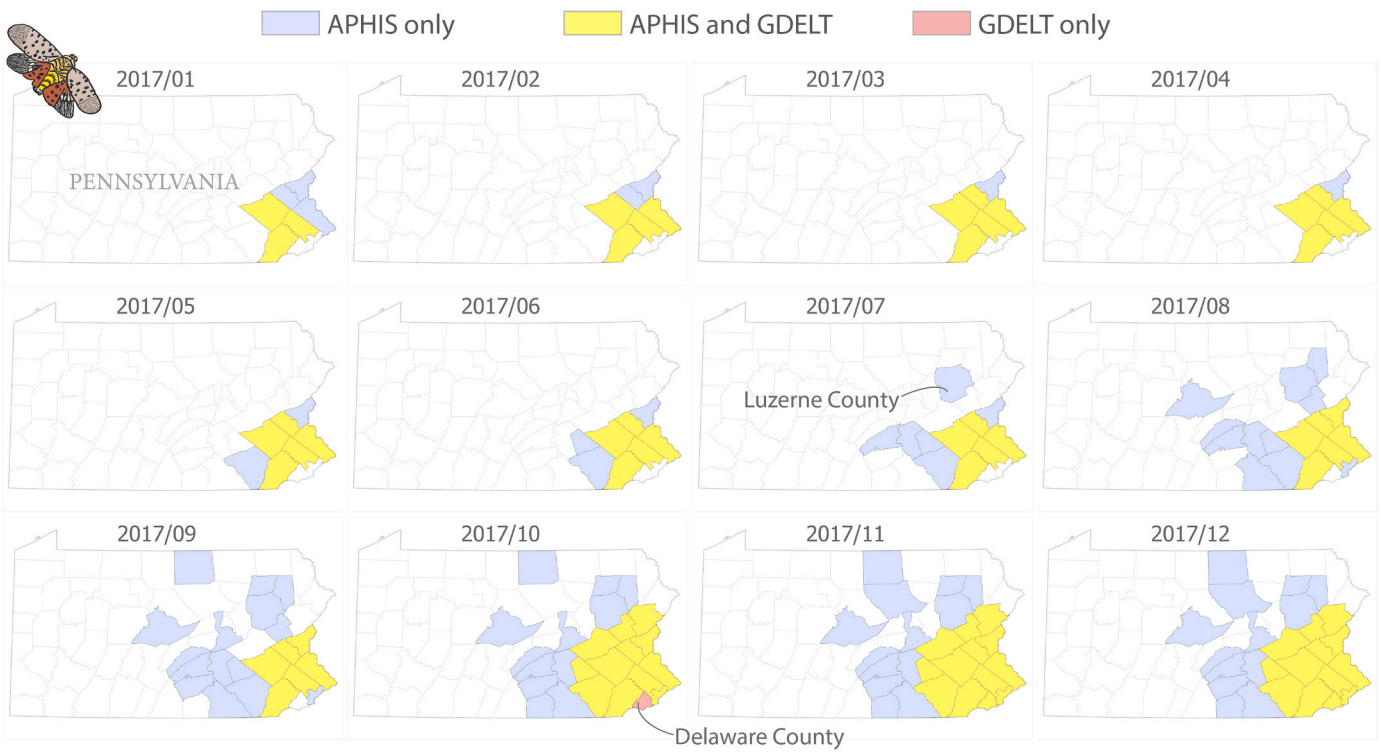



**Fig. 10.** Cumulative monthly Spotted Lanternfly maps of 2014–2017 official records and 2017 GDELT news mentions with data aggregated to the county level. Blue counties have accumulated at least one record by USDA APHIS only. Yellow counties have accumulated both APHIS records and GDELT article mentions, and pink counties have only GDELT article mentions.

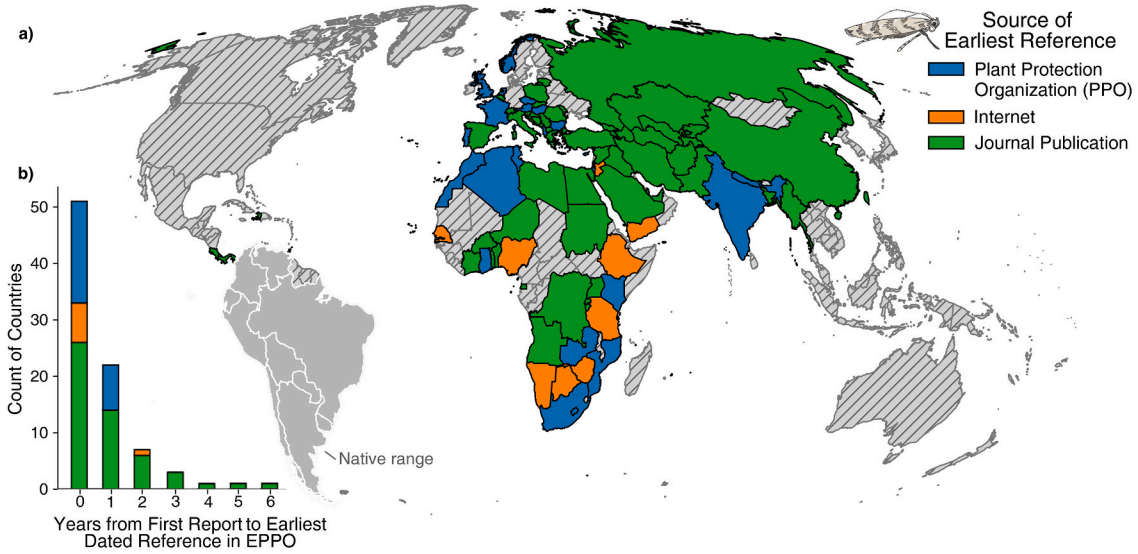



**Fig. 11.** Source of Earliest Reference for countries with Tuta absoluta reports documented in the EPPO-GD. The map (a) shows the source type of the earliest reference provided for that country. If the reference was dated, (b) shows the difference in timing between the First report date and the publication date of the earliest reference. Supplementary Fig. 1 displays this timing geographically.

than full article text, but with an expected 41–85% reduction in hits. Google News does not provide geographic filtering at the level of US states, so we were unable to directly compare the US Google News article count to the GDELT article count (restricted to Pennsylvania).

There is little documentation of how each aggregator samples global historical news. Brandwatch does not disclose details about the sites included. GDELT and Google News, though both managed by Google, provided entirely distinct article results for Spotted Lanternfly, and sampled from just three common domains (out of 62 unique domains between the two aggregators). Though GDELT, Google News, and Brandwatch yielded largely distinct results, the quantity of articles across the aggregators was comparable for Spotted Lanternfly. The dramatic differences in the volume of posts returned from Brandwatch and Google News for the global 10-year Tuta absoluta query (14,601 from Brandwatch, 477 from Google News) suggest that Google News results are pre-filtered to reduce the total count (i.e., eliminating similar articles from the returned search results). While useful to a reader, result counts from Google News may therefore be less appropriate for

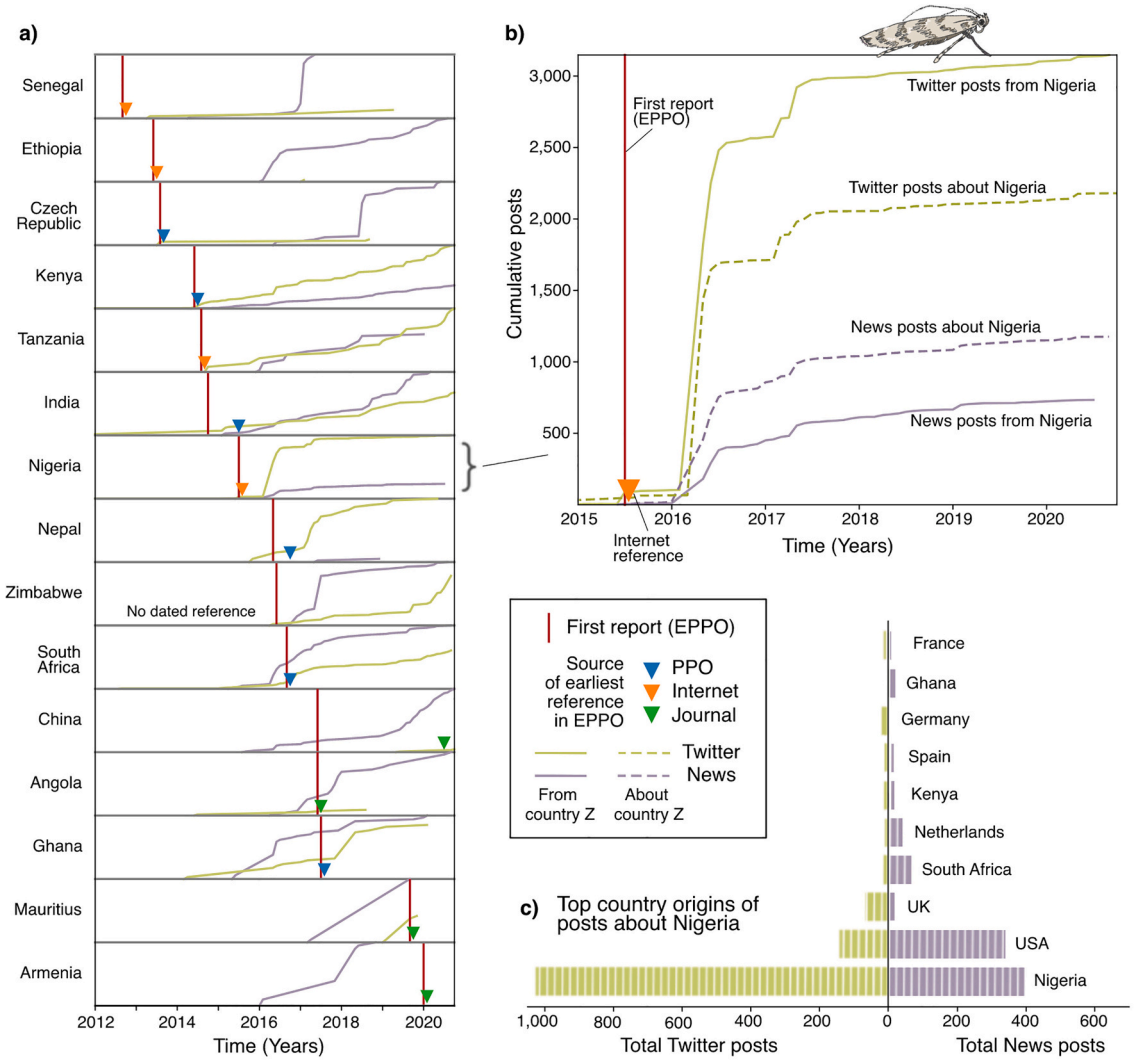



**Fig. 12.** Comparing the First report (red vertical lines) and First reference (caret; color = Source Type) with the timing of Twitter and news posts about Tuta absoluta (cumulative counts). (a) Shows countries with First reports during the study period (2011 – 2022). In (b) news and Twitter posts from Nigeria (post origin location in Nigeria) and posts about Nigeria (post mentions Nigeria or a state in Nigeria from all post origins) are shown. For posts about Nigeria, (c) presents the 10 most common country origins.

describing trends in space and time.

Some of the difference in articles returned may also be attributed to how location is defined within the query for each service. With GDELT, Pennsylvania was the "Actor1" (entity performing an action within a news event), whereas with Google News, the location used to filter (United States) likely describes the source of the post. Brandwatch similarly classifies location as the publisher's location. The approach taken by GDELT, however, may provide information that is more useful to answering questions about where a pest has been observed and describing its spread, with less post-processing of text.

To distinguish unique articles, we relied on URL and article title. However, two distinct URLs may lead to the same article and distinct titles may be applied to the same content by different news media sites (Lyon et al. 2012). Furthermore, additional articles do not necessarily provide new information. For example, when reading the 32 GDELT SLF articles, we noticed that a 02/01/2017 article from a Croatian online news outlet and a 02/03/2017 article from lancasterfarming.com contained some unique content and different wording. However, both articles mentioned the same three Pennsylvania townships being added to the quarantine (as shown by the blue-only 02/03 map in Fig. 9).

Twitter presents several advantages over news in terms of acquisition and processing. Overall, similar queries returned many more Tweets than news articles. The recent (November 2021) Academic Research Twitter API introduces a low cost alternative to subscription aggregators that provides broad coverage for researchers. Additionally, the short, content-dense text segments in Tweets can present a focused overview of "important" news. With less extraneous information, the topic and relevance of a post may be easier to assess automatically.

As a data source, Twitter also has specific limitations. The short post length may exclude context useful to classification and other downstream tasks. In attributing content location, a post geo-tag or the post author's location is typically used, though these were infrequently provided. Aggregation services like Brandwatch further infer location from user behavior and other context. Location may also need to be traced through a chain of posts. For example, a reply identifying an image as "Tuta absoluta" may come from a different location than the original post. Further, the Twitter platform is subject to variable regulation (several countries including China and Iran have long-term bans on Twitter, and others including Egypt, Turkey and Nigeria have had short-term bans) and future changes to data sharing policies could impact continuity.

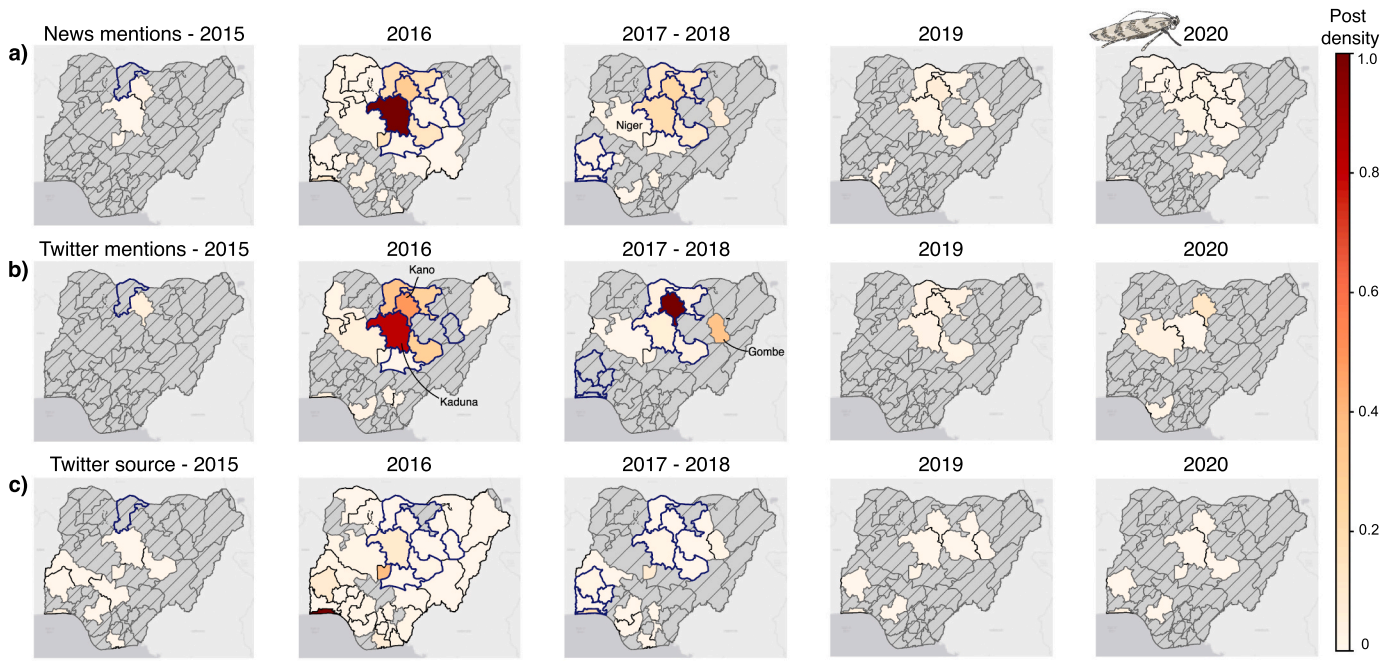



**Fig. 13.** Comparing subnational pest observations in Nigeria with news and Twitter posts, at the state and year scale. Blue outlines show where Tuta absoluta was observed in 2016 (Borisade et al., Aug. 2017) and 2017–2018 (Aigbedion-Atalor et al., Dec. 2019) surveys and the 2015 first observation. Color fill shows the density of states mentioned in news posts (a), Twitter posts (b) and the state origin of Twitter posts (c) for years 2015–2020, scaled across years. Labeled states highlight key events and the need for place-name disambiguation, referenced in the Discussion.

**Table 2**
Agreement and accuracy statistics calculated for the multi-category classification with *n* classes (9, 8, 3, or 2) for Tasks 1 and 2 for the two case studies.

| | | |
|---|---|---|
| *Spotted Lanternfly: Twitter* | | |
| | **Task 1 (n = 9)** | **Task 2 (n = 3)** |
| $\alpha_{\text{INTER-REVIEWER}}$ | 0.95 | 0.93 |
| $\alpha_{\text{CLASSIFIER}}$ | 0.41 | 0.43 |
| Macro $F1_{\text{DIFF}}$ | 0.49–0.11 = 0.38 | 0.62–0.33 = 0.29 |
| *Tuta absoluta: Twitter* | | |
| | **Task 1 (n = 9)** | **Task 2 (n = 3)** |
| $\alpha_{\text{INTER-REVIEWER}}$ | 0.99 | 0.92 |
| $\alpha_{\text{CLASSIFIER}}$ | 0.45 | 0.41 |
| Macro $F1_{\text{DIFF}}$ | 0.47–0.11 = 0.36 | 0.59–0.33 = 0.26 |
| *Tuta absoluta: News* | | |
| | **Task 1 (n = 8)** | **Task 2 (n = 2)** |
| $\alpha_{\text{INTER-REVIEWER}}$ | 0.90 | 0.93 |
| $\alpha_{\text{CLASSIFIER}}$ | 0.46 | 0.56 |
| Macro $F1_{\text{DIFF}}$ | 0.52–0.20 = 0.32 | 0.78–0.5 = 0.28 |

### 4.2. Timing across source type

In our case studies, Twitter post volumes largely mirrored news volumes in areas of high conversation (e.g., at the epicenters of major outbreaks, in locations like the United States with high overall post volume). For Spotted Lanternfly, this trend was consistent at both the global and local (state-level) scales. For Tuta absoluta, correlation was variable across countries and many were represented in only one of the two sources (27 countries with only news, 22 countries with only Tweets). Future studies should consider multiple sources and carefully select the type of media used based on known usage and relative volumes, as well as the information content of posts (topics discussed, places mentioned).

### 4.3. Timing of posts versus recorded events

Activity on both platforms paralleled real world events. Spotted Lanternfly activity followed the seasonal pest cycle, with the largest peaks of Twitter and news activity in the summer through fall months. Tuta absoluta posts coincided with global pest spread, according to the first reports consolidated by the EPPO. In cases where posts predate reports or their references, the potential for earlier access to global first reports should be further evaluated.

Posts provided valuable geographic information documenting pest spread in both case studies. By manually extracting the Spotted Lanternfly event location names mapped in Fig. 9, we illustrated the potential of news articles to map the progress of a pest encroachment and management efforts to control spread. The county appearing in the news prior to the APHIS data (Delaware County, pink in Fig. 10) demonstrates that if modelers are only using official records to model the spread, they would miss presence in this location. While information about current quarantine restrictions is readily available, dated historical quarantine records (e.g., when the quarantines started) may not be easily accessible to modelers. Maps like Fig. 9 generated from news articles can provide a finer record to examine the efficacy of past control actions.

Automated extraction can also offer continuous access to time-sensitive pest information for years and places where survey data is not available. For Tuta absoluta, news and Twitter aligned with observed pest locations before their publication in scientific sources, while the origin of Twitter posts provided more diffuse information. States mentioned in news matched all 2016 and 2017–2018 survey observations (information later published in 2017 and 2019, respectively). Other locations mentioned in news could describe continued pest presence between survey years (e.g., Gombe state from 2016 to 2017/2018 in Fig. 13) or places where scientific observations are not available or feasible. Further, place mention volumes may reflect the intensity of the invasion. For example, Kano and Kaduna were the most mentioned states in both news and Tweets in 2016 (Fig. 13b), matching the impactful outbreaks in these states (a state of emergency was declared in the tomato sector in Kaduna state and over 2 billion Nigerian

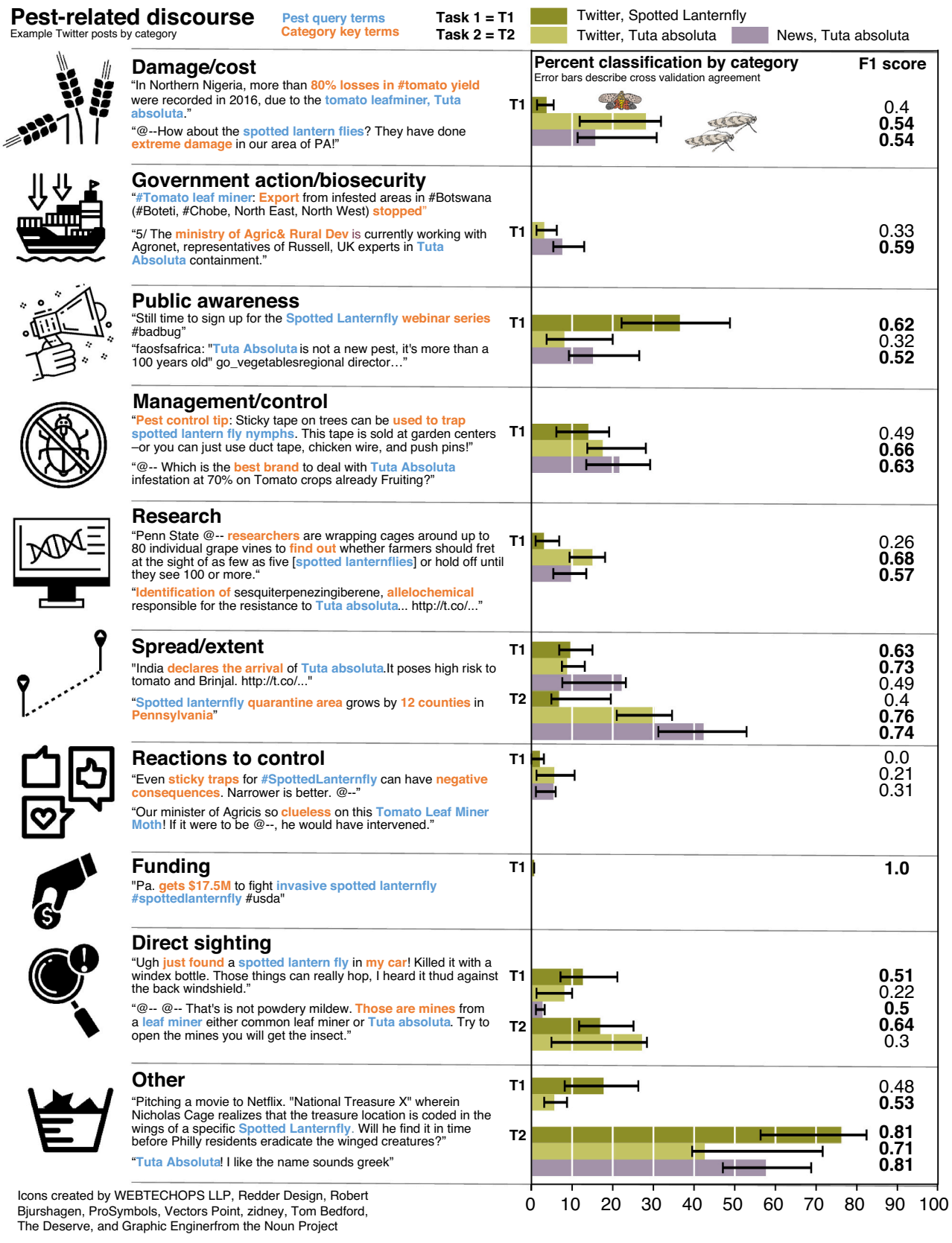



**Fig. 14.** Classification and validation of pest-related discourse in news and Twitter. Example posts are provided for the topic categories used in the two classification tasks.The error bars correspond with the over- and under-prediction by the algorithm for each class based on reviewer validation.

**Table 3**
Overview comparison of our use of news and Twitter aggregators.

| | GDELT with Google BigQuery | Google News | Brandwatch for News | Brandwatch for Twitter |
|---|---|---|---|---|
| **Query options** | Geographic regions (specified as ACTORS), timespan, and CODES. CODES limited to ontology indexed by developers. **No custom keyword queries.** | Keywords, timespan; **No direct geographic filtering.** Geographic region at the country level, can be specified with user settings. A lower administrative level can be specified as a keyword. | Keywords, timespan, geographic region, and language. | Keywords, timespan, geographic region, and language. |
| **Query output** | Structured data: URL, theme, actors (locations), additional metadata; **no title, snippet, or summary.** | Web pages within a browser. | Structured data: **snippet**, URL, location, language, and additional metadata (e.g., source type). | Structured data: **full post**, URL, location, language, additional metadata (e.g., user, is Retweet). |
| **Post-processing** | Scrape with Scala; Search article contents with Python to further filter. **Many downloaded articles discarded.** | Manually download search results; Extract basic article information from HTML with Python, *assuming use of basic information as proxy for full article.* | N/A, *assuming use of snippet as proxy for full article.* | Remove Retweets, URLs, and duplicates. |
| **Post-processing output** | Title, **full article**, date of publication (plus query output). | URL, publisher, **title**, language, publisher, date, and **short description** of each article. | N/A. | Cleaned query output. |
| **Cost** | Google BigQuery is a paid service ($); Code development and time-intensive download for full-article scrape. | Free manual access; Code development for extracting article metadata from download. | **Paid subscription ($$ $).** | Paid subscription ($$$). |
| **Additional considerations** | Fewer results obtained than for comparable Brandwatch queries. Option to filter by URL keywords for efficiency, but fewer results. **Old full articles may be unavailable.** | Fewer results obtained than for comparable Brandwatch queries. No limits to timespan, but **pre-filtering is opaque** (e.g., to reduce redundancy, remove inactive links, or select articles from a region). **Searcher location and language settings impact results.** | Historical timespan depends on license (e.g., 10 years) **Snippets warehoused by Brandwatch (full articles may be unavailable).** | Historical timespan depends on license. |

Naira was lost to the pest in 2016, reported by the BBC and Punch Newspaper).

Challenges exist in comparing records with finite and imperfectly aligned timespans. Fig. 10 uses multiple years of official Spotted Lanternfly records to more accurately reflect the official historical knowledge of pest presence, but only one year of news. When previous official records are not used, the first four months are pink and yellow, as 2017 APHIS records had Berks County for Jan.-Apr., and Lehigh in March. Given only the 2017 frame of reference, GDELT appears to be ahead of APHIS during the first four months, likely due to the seasonal nature of the pest. 2016 news articles (not studied) may have reported presence in the blue (APHIS only) counties.

For automated approaches, disambiguation is needed to ensure that place name matches do not correspond to other words or proper nouns and place names are frequently not unique to a single geographic location (e.g., references tagged to Niger state in Fig. 13a may refer to the country of Niger). Incorporating Named Entity Recognition (NER) and geoparsing can improve on this. Further, geoparsing can use context to extract hierarchical locational information at various spatial scales. For example, the 2017 news post "The tomato pest, Tuta Absoluta, which has been ravaging tomato farms in many states since last year has resurfaced in *Gombe State*, resulting in huge losses to the farmers. The affected areas are *Boltungo and Zambuk in Yamaltu Deba* and *Garin Faruku in Akko* local government areas respectively" contains detailed locational information at the city, locality, and state scales (italic).

### 4.4. Semantics across pests and source types

The major pest topics appeared across sources, though refined classification is needed to determine their true distribution and reliably separate themes. The categories, keyword terms, and examples identified support future efforts to extract this information from text.


**Fig. 15.** Illustrative word clouds of 90 Spotted Lanternfly "Direct sightings" Tweets. 32 training and 58 verification Tweets were stripped of Spotted Lanternfly terms and stopwords and faceted by parts of speech (from left to right: verbs, adverbs, adjectives, nouns, and proper nouns).

Additionally, high-value content documenting historical pest spread and new immediate observations of pests was found for both case studies across the different media sources.

Comparing semantics from our two case studies, the focus of conversation (Fig. 14), and the language and tone used to describe similar concepts varied by pest species, and location. Researchers can learn new keywords and local vernacular through a sub-sample of posts from different locations. Fig. 15 illustrates the verbs, adverbs, adjectives, nouns, and proper nouns (barring the words for Spotted Lanternflies) in 90 Spotted Lanternfly "Direct sightings" Tweets. We found many Spotted Lanternfly direct sightings using emphatic verbs, such as "mashed", "stomped", etc., which may reflect popular outreach campaigns to manage the pest in northeastern US, while Tuta absoluta was popularly referred to as "tomato Ebola" in several countries in Africa.

Differences between pests have implications for classifying "Direct sightings", possibly due to probability of detection and semantic variability. This category was more frequently coded in Tweets by reviewers for Spotted Lanternfly (19%) than Tuta absoluta (4%). More people may recognize Spotted Lanternfly by its distinctive red spotted wing-patterns than Tuta absoluta, a small brown moth. Another consideration is a pest's presence in agricultural versus residential areas. For example, Spotted Lanternfly had frequent sightings in high density urban areas (hence key terms like "my car", "my house", "my apartment", "my porch", and "hitching a ride" and nouns like windshield, patio, and sidewalk appear the Fig. 15 Tweets), while Tuta absoluta is likely primarily observed directly by vegetable growers, who may be fewer and potentially less technologically connected. Furthermore, Spotted Lanternfly Tweets predominantly came from a single country (91% US, with 22% from the city of Philadelphia, Pennsylvania) versus multiple countries for Tuta absoluta (Nigeria had the highest representation, 28% of total Tweet volume). This likely contributes to more variable ways each pest is discussed, even within English-language content.

For news, we saw that single posts may contain useful content on multiple topics that may be better classified at a sub-post level. A snippet of news discussing agricultural matters may describe spread and damage caused by the pest, as well as people's reactions to provide context about the severity of the situation. The following news snippet from our corpus includes multiple topics noted in brackets: "A seller, Mr. Y— I— said that between April and May, most tomato farmers from Kano and Zaria had been complaining of the tuta absoluta pest attacks [Spread/extent]. I— said that the same scarcity of fresh tomatoes [Damage/cost] occurred within the same period in 2016, and appealed to government intervene to save the situation [Reactions to control]." To achieve agreement or capture the distribution of topics for complex content where human interpretation may be difficult, we suggest involving multiple reviewers and using text classifications algorithms that provide a multi-topic result.

The low F1 scores in Fig. 14 may be due, in part, to the multi-topic issue. But also, recognized limitations of Brightview mean that classification accuracy may not be sufficient for uses beyond content exploration (Hayes et al., Jan. 2021). This tool presents an ease-of-use tradeoff with limited options for fine tuning. Further, though our training data sizes followed Brightview guidelines, there is a relationship between training dataset size and performance that suggests that additional labeled training data could improve classification (Barberá, Boydstun, Linn, McMahon, & Nagler,. 2021). Natural Language Processing (NLP) approaches for gathering more fine-grained pest event data from news and Twitter (such as regular expression matching, NER, geoparsing, and translation), can be used in tandem with or in lieu of machine learning classification, as researchers have done with human and animal disease spread (Valentin et al., Dec. 2021; Rabatel et al., Feb. 2019; Freifeld, Mandl, Reis, & Brownstein, 2008).

## 5. Conclusions

The invasion of new insect species has devastating agricultural and economic consequences. Low-latency information is vital in the fight against invasive pests. Our local- and global-scale case studies showed the presence of key information about both species in Tweets and news articles. We laid bare the trade-offs of three retrieval mechanisms for online news for tracking pests. **Subscription costs, query flexibility, and computer programming proficiency requirements are the most prominent differences**. For applications that need abundant context or specific metadata, the returned elements (e.g., snippet versus full article, post location) will also influence the data retrieval workflow. For long historical time periods and large spatial scales, commercial services like Brandwatch may become more cost effective than comparable Google News and GDELT workflows. Since both news and Twitter contained information relevant to pest tracking and post volumes were often temporally correlated, use cases should leverage the complementary content captured in each. The immediacy of Twitter gives it an edge for real-time monitoring applications, but news may more systematically track pest events.

In both case studies, our results indicated that news and Twitter activity and content parallel real-world events. We found that **online news and Tweets included valuable spatial information, documenting both direct sightings and the geographic extent of pests at spatial and temporal scales not otherwise publicly available**. Past media posts with artifacts of pest presence and movement can play a vital role in filling gaps where official records are spatially or temporally sparse. Incorporating a media volume parameter could be a rapid low-cost first step for injecting this data into predictive models. After this, text classification and event detection can further refine data extraction. **The diverse textual content of posts could be used to assess the availability and use of management strategies, crop loss, or to gauge the success of public awareness campaigns**. Moving from reactive to proactive real-time monitoring, media may provide early indication of emerging pest outbreaks in new locations for human-in-the-loop listening systems (e.g., Fig. 6). To battle the accelerating threat posed by invasive species, timely extraction and communication of pest events from news and social media can provide rapid access to data in support of global pest surveillance.

## CRediT authorship contribution statement

**Laura G. Tateosian:** Conceptualization, Data-curation, Methodology, Software, Writing-original-draft, Writing-review-editing, Visualization. **Ariel Saffer:** Conceptualization, Data-curation, Formal-analysis, Methodology, Software, Writing-original-draft, Writing-review-editing, Visualization. **Chelsey Walden-Schreiner:** Data-curation, Methodology, Writing-original-draft, Writing-review-editing. **Makiko Shukunobe:** Data-curation, Software, Visualization.

## Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgements

We thank Chris Jones, Research Scholar at North Carolina State University, for providing and explaining the Spotted Lanternfly survey data. This material was made possible, in part, by Cooperative Agreements from the United States Department of Agriculture's Animal and Plant Health Inspection Service (APHIS). It may not necessarily express APHIS' views.

## Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.compenvurbsys.2022.101922.

## References


Aigbedion-Atalor, P. O., Oke, A. O., Oladigbolu, A. A., Layade, A. A., Igbinosa, I. B., & Mohamed, S. A. (Dec. 2019). Tuta absoluta (Lepidoptera: Gelechiidae) invasion in Nigeria: First report of its distribution. *Journal of Plant Diseases and Protection, 126*(6), 603–606. https://doi.org/10.1007/s41348-019-00255-3

Barberá, P., Boydstun, A. E., Linn, S., McMahon, R., & Nagler, J. (Jan. 2021). Automated Text Classification of News Articles: A Practical Guide. *Political Analysis, 29*(1), 19–42. https://doi.org/10.1017/pan.2020.8. Publisher: Cambridge University Press.

Barringer, L. E., Donovall, L. R., Spichiger, S.-E., Lynch, D., & Henry, D. (June 2015). The First New World Record of Lycorma delicatula (Insecta: Hemiptera: Fulgoridae). *Entomological News, 125*(1), 20–23. https://doi.org/10.3157/021.125.0105

Biondi, A., Guedes, R. N. C., Wan, F.-H., & Desneux, N. (2018). Ecology, Worldwide Spread, and Management of the Invasive South American Tomato Pinworm, Tuta absoluta: Past, Present, and Future. *Annual Review of Entomology, 63*(1), 239–258. https://doi.org/10.1146/annurev-ento-031616-034933

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research, 3*(Jan), 993–1022.

Borisade, O. A., Kolawole, A. O., Adebo, G. M., & Uwaidem, Y. I. (Aug. 2017). The tomato leafminer (Tuta absoluta) (Lepidoptera: Gelechiidae) attack in Nigeria: Effect of climate change on over-sighted pest or agro-bioterrorism? *Journal of Agricultural Extension and Rural Development, 9*(8), 163–171. https://doi.org/10.5897/JAERD2017.0856. Publisher: Academic Journals.

Brandwatch (2022). FAQ: How does Brandwatch classify location?

Centre for Agriculture and Bioscience International Invasive Species Compendium.https://www.cabi.org/ISC. Accessed: 2022-01-31.

Daume, S., & Galaz, V. (Mar. 2016). "Anyone Know What Species This Is?" – Twitter Conversations as Embryonic Citizen Science Communities. *PLOS ONE, 11*(3), Article e0151387. https://doi.org/10.1371/journal.pone.0151387

Deiner, M. S., McLeod, S. D., Chodosh, J., Oldenburg, C. E., Fathy, C. A., Lietman, T. M., et al. (Feb. 2018). Clinical Age-Specific Seasonal Conjunctivitis Patterns and Their Online Detection in Twitter, Blog, Forum, and Comment Social Media Posts. *Investigative Ophthalmology & Visual Science, 59*(2), 910–920. https://doi.org/10.1167/iovs.17-22818

Desneux, N., Wajnberg, E., Wyckhuys, K. A. G., Burgio, G., Arpaia, S., Narváez-Vasquez, C. A., et al. (Aug. 2010). Urbaneja, Biological invasion of European tomato crops by Tuta absoluta: Ecology, geographic expansion and prospects for biological control. *Journal of Pest Science, 83*(3), 197–215. https://doi.org/10.1007/s10340-010-0321-6

Diagne, C., Leroy, B., Vaissière, A.-C., Gozlan, R. E., Roiz, D., Jarić, I., et al. (Apr. 2021). High and rising economic costs of biological invasions worldwide. *Nature, 592*(7855), 571–576. https://doi.org/10.1038/s41586-021-03405-6. Number: 7855 Publisher: Nature Publishing Group.

European and Mediterranean Plant Protection Organization Global Database.https://gd.eppo.int/. Accessed: 2020-11-23.

Food and Agriculture Organization of the United Nations official pest report.https://www.ippc.int/en/countries/all/pestreport/. Accessed: 2022-11-12.

Firat, A. (2017). *(71) Applicant*. Boston: CRIMSON HEXAGON, INC.. https://patentimages.storage.googleapis.com/06/4d/b4/4cedcb68a877d9/US20170046630A1.pdf. pp. 13.

Freifeld, C. C., Mandl, K. D., Reis, B. Y., & Brownstein, J. S. (2008). HealthMap: Global Infectious Disease Monitoring through Automated Classification and Visualization of Internet Media Reports. *Journal of the American Medical Informatics Association: JAMIA, 15*(2), 150–157. https://doi.org/10.1197/jamia.M2544

Global Biodiversity Information Facility Occurrence Downlaod.https://www.gbif.org/occurrence/download/0005826-220831081235567. Accessed: 2022-11-12.

The GDELT global knowledge graph (gkg) data format codebook v2.1.http://data.gdeltproject.org/documentation/GDELT-Global_Knowledge_Graph_Codebook-V2.1.pdf. Accessed: 2022-07-07.

Hart, A. G., Carpenter, W. S., Hlustik-Smith, E., Reed, M., & Goodenough, A. E. (2018). Testing the potential of Twitter mining methods for data acquisition: Evaluating novel opportunities for ecological research in multiple taxa. *Methods in Ecology and Evolution, 9*(11), 2194–2205. https://doi.org/10.1111/2041-210X.13063

Hawelka, B., Sitko, I., Beinat, E., Sobolevsky, S., Kazakopoulos, P., & Ratti, C. (May 2014). Geo-located Twitter as proxy for global mobility patterns. *Cartography and Geographic Information Science, 41*(3), 260–271. https://doi.org/10.1080/15230406.2014.890072

Hayes, J. L., Britt, B. C., Evans, W., Rush, S. W., Towery, N. A., & Adamson, A. C. (Jan. 2021). Can Social Media Listening Platforms' Artificial Intelligence Be Trusted? Examining the Accuracy of Crimson Hexagon's (Now Brandwatch Consumer Research's) AI-Driven Analyses. *Journal of Advertising, 50*(1), 81–91. https://doi.org/10.1080/00913367.2020.1809576. Publisher: Taylor & Francis Ltd.

Ikonomakis, M., Kotsiantis, S., & Tampakas, V. (2005). Text Classification Using Machine Learning Techniques.

Jarić, I., Correia, R. A., Brook, B. W., Buettel, J. C., Courchamp, F., Di Minin, E., et al. (July 2020). iEcology: Harnessing Large Online Resources to Generate Ecological Insights. *Trends in Ecology & Evolution, 35*(7), 630–639. https://doi.org/10.1016/j.tree.2020.03.003

Jones, C., Skrip, M. M., Seliger, B. J., Jones, S., Wakie, T., Takeuchi, Y., et al. (June 2022). Spotted lanternfly predicted to establish in California by 2033 without preventative management. *Communications Biology, 5*(1), 1–9. https://doi.org/10.1038/s42003-022-03447-0. Number: 1 Publisher: Nature Publishing Group.

Keller, M., Blench, M., Tolentino, H., Freifeld, C. C., Mandl, K. D., Mawudeku, A., et al. (May 2009). Use of Unstructured Event-Based Reports for Global Infectious Disease Surveillance. *Emerging Infectious Diseases, 15*(5), 689–695. https://doi.org/10.3201/eid1505.081114

Latombe, G., Canavan, S., Hirsch, H., Hui, C., Kumschick, S., Nsikani, M. M., et al. (2019). A four-component classification of uncertainties in biological invasions: Implications for management. *Ecosphere, 10*(4), Article e02669. https://doi.org/10.1002/ecs2.2669

Latombe, G., Pyšek, P., Jeschke, J. M., Blackburn, T. M., Bacher, S., Capinha, C., et al. (Sept. 2017). A vision for global monitoring of biological invasions. *Biological Conservation, 213*, 295–308. https://doi.org/10.1016/j.biocon.2016.06.013

Lee, M., Kim, J. W., & Jang, B. (2018). DOVE: An Infectious Disease Outbreak Statistics Visualization System. *IEEE Access, 6*, 47206–47216. https://doi.org/10.1109/ACCESS.2018.2867030. Conference Name: IEEE Access.

Lyon, A., Nunn, M., Grossel, G., & Burgman, M. (2012). Comparison of Web-Based Biosecurity Intelligence Systems: BioCaster, EpiSPIDER and HealthMap. *Transboundary and Emerging Diseases, 59*(3), 223–232. https://doi.org/10.1111/j.1865-1682.2011.01258.x

Mammola, S., Malumbres-Olarte, J., Arabesky, V., Barrales-Alcalá, D. A., Barrion-Dupo, A. L., Benamú, M. A., et al. (Dec. 2022). An expert-curated global database of online newspaper articles on spiders and spider bites. *Scientific Data, 9*(1), 109. https://doi.org/10.1038/s41597-022-01197-6

Mansour, R., Brévault, T., Chailleux, A., Cherif, A., Grissa-Lebdi, K., Haddi, K., et al. (Dec. 2018). Occurrence, biology, natural enemies and management of Tuta absoluta in Africa. *Entomologia Generalis, 38*(2), 83–112. https://doi.org/10.1127/entomologia/2018/0749

Martinez, B., Reaser, J. K., Dehgan, A., Zamft, B., Baisch, D., McCormick, C., et al. (Jan. 2020). Technology innovation: advancing capacities for the early detection of and rapid response to invasive species. *Biological Invasions, 22*(1), 75–100. https://doi.org/10.1007/s10530-019-02146-y

Meentemeyer, R., Walden-Schreiner, C., Saffer, A., & Jones, C. (2021). Invasive Species. In *International Encyclopedia of Geography* (pp. 1–9). John Wiley & Sons Ltd.. https://doi.org/10.1002/9781118786352.wbieg2004

Ofori, M., & El-Gayar, O. (June 2021). Drivers and challenges of precision agriculture: A social media perspective. *Precision Agriculture, 22*(3), 1019–1044. https://doi.org/10.1007/s11119-020-09760-0

Pyšek, P., Hulme, P. E., Simberloff, D., Bacher, S., Blackburn, T. M., Carlton, J. T., et al. (2020). Scientists' warning on invasive alien species. *Biological Reviews, 95*(6), 1511–1534. https://doi.org/10.1111/brv.12627

Rabatel, J., Arsevska, E., & Roche, M. (Feb. 2019). PADI-web corpus: Labeled textual data in animal health domain. *Data in Brief, 22*, 643–646. https://doi.org/10.1016/j.dib.2018.12.063

Roxburgh, N., Guan, D., Shin, K. J., Rand, W., Managi, S., Lovelace, R., et al. (Jan. 2019). Characterising climate change discourse on social media during extreme weather events. *Global Environmental Change, 54*, 50–60. https://doi.org/10.1016/j.gloenvcha.2018.11.004

Sakaki, T., Okazaki, M., & Matsuo, Y. (2010). Earthquake shakes Twitter users: Real-time event detection by social sensors (pp. 10).

Savary, S., Laetitia, W., Pethybridge, S. J., Esker, P., McRoberts, N., & Nelson, A. (Mar. 2019). The global burden of pathogens and pests on major food crops. *Nature Ecology & Evolution, 3*(3), 430–439. https://doi.org/10.1038/s41559-018-0793-y. Num Pages: 430–439 Place: London, United States Publisher: Nature Publishing Group.

Seebens, H., Blackburn, T. M., Dyer, E. E., Genovesi, P., Hulme, P. E., Jeschke, J. M., et al. (Feb. 2017). No saturation in the accumulation of alien species worldwide. *Nature Communications, 8*(1), 14435. https://doi.org/10.1038/ncomms14435. Bandiera_abtest: a Cc_license_type: cc_by Cg_type: Nature Research Journals Number: 1 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Invasive species;Macroecology Subject_term_id: invasive-species;macroecology.

Seebens, H., Essl, F., Dawson, W., Fuentes, N., Moser, D., Pergl, J., et al. (2015). Global trade will accelerate plant invasions in emerging economies under climate change. *Global Change Biology, 21*(11), 4128–4140. https://doi.org/10.1111/gcb.13021

Valentin, S., Arsevska, E., Rabatel, J., Falala, S., Mercier, A., Lancelot, R., et al. (Dec. 2021). PADI-web 3.0: A new framework for extracting and disseminating fine-grained information from the news for animal disease surveillance. *One Health, 13*, 100357. https://doi.org/10.1016/j.onehlt.2021.100357

Valentin, S., Lancelot, R., & Roche, M. (Jan. 2021). Identifying associations between epidemiological entities in news data for animal disease surveillance. *Artificial Intelligence in Agriculture, 5*, 163–174. https://doi.org/10.1016/j.aiia.2021.07.003

Zhang, Y., Ibaraki, M., & Schwartz, F. W. (Feb. 2020). Disease surveillance using online news: Dengue and zika in tropical countries. *Journal of Biomedical Informatics, 102*, Article 103374. https://doi.org/10.1016/j.jbi.2020.103374