big data and
cognitive computing

Review

A Review of Social Media Data Utilization for the Prediction of Disease Outbreaks and Understanding Public Perception

Alice Wang [1], Rozita Dara [1,*], Samira Yousefinaghani [1], Emily Maier [2] and Shayan Sharif [2]

[1] Data Management and Privacy Governance Lab, School of Computer Science, University of Guelph, Guelph, ON N1G 2W1, Canada
[2] Department of Pathobiology, Ontario Veterinary College, University of Guelph, Guelph, ON N1G 2W1, Canada
* Correspondence: drozita@uoguelph.ca

Abstract: Infectious diseases take a large toll on the global population, not only through risks of illness but also through economic burdens and lifestyle changes. With both emerging and re-emerging infectious diseases increasing in number, mitigating the consequences of these diseases is a growing concern. The following review discusses how social media data, with a focus on textual Twitter data, can be collected and processed to perform disease surveillance and understand the public's attitude toward policies around the control of emerging infectious diseases. In this paper, we review machine learning tools and approaches that were used to determine the correlation between social media activity in disease trends within regions, understand the public's opinion, or public health leaders' approaches to disease presentation. While recent models migrated toward popular deep learning methods, neural networks and algorithms that optimized existing models were also explored as new standards for social media data analysis in disease prediction and monitoring. As adherence to public health policies can be improved by understanding and responding to major concerns identified by sentiment analyses, the advancements and challenges in understanding text sentiment are also discussed. Recent sentiment classifiers include more complex classifications and can even recognize epidemiological considerations that affect the spread of outbreaks. The comprehensive integration of locational and epidemiological considerations with advanced modeling capabilities and sentiment analysis will produce robust models and more precision for both disease monitoring and prediction. Accurate real-time disease outbreak prediction models will provide health organizations with the capability to address public concerns and to initiate outbreak responses proactively rather than reactively.

Keywords: social media; monitoring; outbreaks; public health; sentiment analysis

check for updates

Citation: Wang, A.; Dara, R.; Yousefinaghani, S.; Maier, E.; Sharif, S. A Review of Social Media Data Utilization for the Prediction of Disease Outbreaks and Understanding Public Perception. *Big Data Cogn. Comput.* **2023**, *7*, 72. https://doi.org/10.3390/bdcc7020072

Academic Editors: Carson K. Leung, Fei Hao, Giancarlo Fortino and Xiaokang Zhou

Received: 5 February 2023
Revised: 10 April 2023
Accepted: 10 April 2023
Published: 12 April 2023

1. Introduction

Data from social media platforms, including Facebook, Twitter, and Sina Weibo, are used for trend prediction in a variety of applications, such as forecasting stock market share values [1]. Predictive models that use social media data are desirable because real-time data availability enables stakeholders to initiate an informed response earlier than when using traditional data collection methods [1]. In the context of public health, social media data can be used to identify infectious disease outbreaks earlier than previously possible [2]. Traditional disease surveillance methods require strict protocols, active surveillance, and a rigorous compilation process prior to publication, which results in a substantial delay in data availability. In contrast, collecting self-reported health information from social media posts requires much less time, expense, and manual labor [3,4]. Influenza is a common infectious disease that is monitored using data collected from internet applications [5], but other disease outbreaks, such as the Zika virus, were detected and forecasted using social media data as well [2]. Studies also focused on the use of Google Flu Trends (GFT) to predict influenza rates by analyzing the frequencies of certain undisclosed search terms. However, the GFT service was discontinued in 2015 [6] following overwhelming reports providing evidence that GFT tended to overestimate influenza infection rates, making it less reliable than surveillance using Twitter data [7–10]. Due to internet censorship in China, many websites, such as Twitter, are not accessible to the country's general population. Data from a Chinese social media platform called Sina Weibo, which is akin to Twitter, was used to detect and predict influenza and COVID-19 rates in China with high accuracy [11–13]. Alternative platforms leveraged for infectious disease monitoring include Wikipedia, Instagram, Tumblr, Reddit, and Daum [14–16].

Quantifying illness prevalence can inform a public health response in anticipation of the increased burden on the healthcare system, enabling the appropriate allocation of resources. In addition to outbreak detection and prediction, social media data can be used to identify common concerns and trends in public opinion regarding disease outbreaks via sentiment analysis [12]. For example, a 2022 study used Twitter and Reddit data to track public sentiment toward COVID-19 vaccines during development, testing, and deployment [17]. Public concern may influence compliance with public health measures, such as stay-at-home orders and vaccine mandates. Furthermore, leveraging nuances in social media posts mentioning diseases or symptoms creates the potential to detect novel illnesses that may not have been identified through surveillance programs in a timely manner. Ultimately, capitalizing on available social media data provides several valuable insights that can be used to inform decisions regarding public health preparation and responses to infectious diseases. Although the use of social media in disease outbreak monitoring is known to be effective, it was recommended only as a supplementary source of information to be used in conjunction with traditional methods [3].

The COVID-19 pandemic spurred more research on the use of social media data to quantify infectious disease outbreaks and evaluate public response, providing a new body of literature on the topic. Recent review articles summarized various aspects of such work. For example, Mohammed and Abbas (2021) systematically identified the best Twitter API for several research objectives related to infectious disease outbreak detection [18]. Alamoodi et al. (2021) assessed the ability to determine attitudes toward public health measures using machine learning models across many studies [19], while Sooknanan and Mays (2021) went a step further and summarized different methods for incorporating this information into mathematical models of disease forecasting [20]. Public health measures implemented to control the spread of this infectious disease became a polarizing topic [21]. Therefore, regional public health measures should reflect local disease prevalence and public opinion of the implemented measures. However, to the best of our knowledge, there are no post-pandemic review articles highlighting the applications or use cases of extracting precise locational data from social media posts. Furthermore, there is a limited number of recent review articles that examined the trends and commonly used practices in using machine learning models on social media data, with or without the addition of public health surveillance data, to detect and predict disease outbreaks. This paper aimed to bridge the gap in understanding the use of social media data analysis for infectious disease prediction and understanding public perception.

We used keywords and key phrases, including "disease surveillance and social media", "disease outbreak and sentiment analysis", "disease outbreak prediction and machine learning", and many other combinations of different terminologies to find relevant papers in search engines for scientific literature, such as IEEE and PubMed. We used the most relevant publications to categorize different works into topics that are presented in this paper, including disease surveillance, sentiment analysis, and measuring disease activity. We also elaborated on the challenges of using social media for disease modeling.

This review highlights predominant applications and the use cases for processing textual and locational data from social media posts, as well as emerging trends in monitoring disease outbreaks using social-media-inclusive models and understanding related concerns among defined populations from these data. Although an in-depth evaluation of methods and algorithms used in the studies cited in this paper was out of the scope of this review, we made an attempt to elaborate on their approaches, algorithms, and outcomes.

2. Disease Surveillance

Data retrieved from social media platforms were used for risk assessment, sentiment analyses, monitoring, and prediction [22]. In an epidemiolocal context, profiling disease risk and conducting sentiment analysis for the detection of misinformation and understanding public opinion are current use cases of social media data. Both detecting outbreaks promptly and being privy to public concerns during a crisis are important for an effective public health response.

The use of social media data containing real-time self-reported accounts of illness in disease surveillance is superior to traditional public health surveillance programs for data collection with respect to saving resources. The abundance of data produced by social media users can include textual information on disease states, symptoms, location, and profile engagement information. Many algorithms were used to correlate these data with disease trends detected through surveillance programs so that future outbreaks can be detected and predicted from social media data.

Due to its features, Twitter is the most studied social media platform in the context of public health surveillance [14]. Its short, text-based posts broadcasted to followers are often more accessible than posts shared amongst mutual connections. On a platform such as Facebook, privacy configurations for posts shared to mutual connections often have more restrictions, which limit the amount of data accessible to analysts [23]. Additionally, the nature of Twitter posts allows for easier analysis than photo-based platforms, such as Instagram. Textual posts are currently more easily and accurately interpreted using natural language processing (NLP) in comparison to images, resulting in an overwhelming focus on Twitter in social media studies. However, the first study that utilized Instagram to predict disease levels in Finland was published in 2018 with a promising implementation of convoluted neural networks to forecast influenza [15]. The format of Twitter posts may still pose challenges due to the frequent use of multiple-word hashtags, abbreviations, and sarcasm, which is often poorly detected if considered at all.

Many steps are required to establish a correlation between social media data and disease prevalence. Data retrieval, pre-processing, filtering, and further data mining were the target of efforts to produce higher-quality data for the purpose of disease modeling. Both statistical models and machine learning algorithms are useful for identifying relevant data. Constant improvements have helped to reduce errors in filtering and pre-processing to ensure that posts pertaining only to a user's current state of health are used in disease modeling.

2.1. Information Retrieval and Pre-Processing

Twitter data that is mined for the purpose of assessing illness prevalence and sentiment analysis are typically retrieved through an application programming interface (API), which can support primary filtering based on prescribed criteria. These criteria can include keywords, follower count thresholds, and user location. Different coding languages, primarily Java and Python, are used to collect and export relevant tweets and prepare them for analyses. Streaming libraries, such as Twitter4J and Tweepy [24], are used to integrate coding applications with APIs. However, there are limitations to the use of APIs. Search APIs collect more data than streaming APIs [25,26], but streaming APIs allow for real-time data collection. To overcome the limitations of each type of API, making continuous API calls is recommended to collect comprehensive data [15,22,25,27,28].

Topic modeling methods, such as bag of words (BoW), collect relevant keywords from a document of text. The number of times particular terms appear in a document is counted to determine keywords. A common metric used to determine the importance of a key term or phrase, called an n-gram, in social media posts is the term-frequency inverse-document frequency (TF-IDF). TF-IDF measures the relevance of the n-gram by analyzing its frequency across several posts [29]. The TF-IDF can also recognize syncategorematic words to discard insignificant terms. Other algorithms such as latent Dirichlet allocation (LDA) and guided LDA are also employed in topic modeling to extract topics from textual data [4]. LDA discovers topics that exist in a collection of documents and the corresponding keywords [30]. Guided LDA is a version of LDA that uses initial annotated keywords to guide the topic search.

From the collection of raw data, pre-processing steps are taken to clean the data before analysis takes place. Common steps include removing URLs, non-ASCII characters, punctuation, and stop words (e.g., an, the, of), as well as tokenizing letters, words, or key phrases [31–34]. Many tweets and Instagram posts include hashtags, which increase the discoverability of posts. However, hashtags cannot contain spaces, which can make the intended meaning of the hashtag difficult to interpret. One method used to split long hashtags into a string of reasonable words is the Aho-Corsick word-splitting algorithm [35]. After tweets with keywords are processed, machine learning techniques can be applied to determine whether the tweet is relevant in the context of disease surveillance.

2.2. Ensuring and Assessing Post Relevancy

Classification models can be employed not only for classifying the content of the posts but also for predicting the type of account that uploaded the post. By analyzing account information, such as how long the account has existed, the number of tweets and replies associated with the account, and the ratio of messages containing URLs, personal accounts can be distinguished from government, business, and bot accounts. Accounts such as those belonging to news stations may report on public health statistics and recommendations but they do not indicate the health status of a specific person. For this reason, these accounts were not used for disease prediction modeling in some models [36,37]. Similarly, other studies only considered posts from users with fewer than 2000 followers to filter out celebrities, news agencies, and similar broadcasting accounts [36]. Some studies only considered social media posts that indicated the account user themselves was ill. These methods discarded “indirect tweets” that indicated the illness of an individual other than the account user. Although indirect tweets were discarded by some studies, other researchers, such as Wakamiya et al. (2018), used indirect tweets to increase the geographical range of their disease modeling [38]. Tweets and active Twitter users tend to congregate around urban areas; therefore, disease modeling performance often suffers in rural locations with fewer or less active social media users. Indirect tweets bolster information on such locations [38]. In 2013, Broniatowski et al. further improved the classification of tweets related to influenza by focusing on increasing the specificity of the model [9]. This was accomplished by constructing filters to differentiate between posts that indicated infection and tweets that only mentioned the topic of influenza, such as in a tweet encouraging annual flu shots. These filters considered other linguistic information, such as semantics, syntax, and writing style, to distinguish between relevant and irrelevant tweets [9]. Most social-media-based analyses also removed duplicate posts (e.g., retweets on Twitter) to improve accurate modeling of disease prevalence [22].

The relevance of social media posts to the topic of investigation is directly related to the quality of the keywords and key phrases used to extract the data. Several factors must be considered when selecting keywords to maximize relevance. First, keywords must be updated over time as language evolves to reflect the ever-changing infectious disease landscape [39,40]. A potential problem for tracking novel diseases on social media was revealed in studies conducted during the initial COVID-19 outbreaks in China. The number of keywords used to refer to COVID-19 on social media drastically increased over time and popular keywords often changed, which made COVID-19-related discussion harder to track [13,41]. One solution to this issue is the use of dynamic topic models (DTMs), which consider the natural evolution of developing keywords that most frameworks do not account for [12]. Another important factor is ensuring that relevant keywords for all languages analyzed by the model are included. Notably, the meaning of a word is not always conveyed in another language when direct translation is used [27]. One recent program capable of utilizing both English and Arabic Twitter posts suggested that combining data from both languages improved the accuracy of the model [24]. Thus, keywords can be language-specific and must be carefully determined for each language that will be included in the model.

Even when keywords are selected methodically, not all posts containing these words will be relevant [28]. For instance, an API may collect tweets containing the keyword "fever", but a tweet referring to "Bieber fever" is not relevant or indicative of disease. Therefore, the algorithm employed to identify posts related to the selected topics must also be carefully selected to facilitate the collection of relevant data. For instance, the alignment topic aspect model (ATAM) [42] was used to identify Twitter users who were sick by recognizing symptoms and illness treatments mentioned in their tweets. The classification power of the ATAM was improved, producing the ATAM+, which can recognize and distinguish between more illnesses and conditions, even when they are similar [43]. In India, V. K. Jain and S. Kumar (2015) analyzed several machine learning classification techniques, such as support vector machine (SVM), naïve Bayes (NB), random forest, and decision tree, and determined that SVM and NB were the most accurate at determining whether Tweets were referring to influenza infection [4]. The SVM classifier was used to classify posts as relevant or not relevant [4,8,9,25,44,45]. The SVM classifier applies kernel functions to numerical data to create an N-dimensional space with a hyperplane that can most distinctly classify all data points into discrete groups. Most recently, a majority-voting ensemble deep learning (MVEDL) model was shown to distinguish between informative and uninformative COVID-19 tweets with 90.75 percent accuracy [46].

Recent strides in text classification include improved prediction accuracy with the application of deep neural networks (DNNs). DNNs learn hidden patterns in data without the need for the feature engineering seen in traditional machine learning models. During the development of the integrated model program SENTINEL, the effectiveness of SVM and NB disease models were compared with two DNN models: convolutional neural network (CNN) and a long short-term memory network (LSTM). The DNNs were found to be more accurate [35]. Alessa et al. (2019) used FastText word embeddings for text classification in comparison to several conventional machine learning algorithms and concurred that using deep learning models were more accurate at classifying Twitter data [47].

2.3. Collecting and Using Location Data

To track disease outbreaks by region using social media posts, the location of the posting user must be known. Often, infected user counts are normalized using census data to account for differences in population density [16,34,36]. Geographic information system (GIS) methods were used for location data collection and normalization [25]. Some profiles and posts [25] are tagged with a location that references geographical coordinates, such as check-in data associated with Sina Weibo posts [48]. For other social media platforms, strategies can be used to identify where a user is located if this information is not available due to the security settings of the account and other factors. For example, the Python library Carmen [9,16] works to determine the location of users by utilizing information from the user's profile and by resolving aliases for any locations listed, allowing for more posts to be utilized. Alternatively, recent work by Essam et al. (2021) used an enhanced dialect identification model to predict user location based on the region-specific dialect of Arabic used in a set of COVID-19-related tweets with more than 97% accuracy [49]. Distinguishing between locations, such as different cities, is important due to regional variabilities of posting patterns [34] that result in variations in the effectiveness of algorithms that detect disease levels [3,50]. These variations in posting patterns may reflect the different perceived thresholds for concern in each major population center during an outbreak [34]. Thus, the interpretation of disease prediction outcomes must be adjusted depending on location. Furthermore, detection and prediction models suffer when and where social media activity is less frequent, suggesting that some models may only be appropriate for populous urban locations.

Determining user locations is also useful for creating visuals, such HealthMap or BioCaster [51,52], to show outbreak distribution on a risk map and to create diagrams that show the path of transmission [44,53]. The density-based spatial clustering of applications with noise (DBSCAN) algorithm was used to differentiate clusters of active users who are located in proximity, such as users located in two neighboring cities [54]. DBSCAN collects the geographical coordinates of users and creates clusters that reflect the separation of two areas without prior knowledge of geographical regions [54]. The ability to distinguish clusters of users enabled improved accuracy in modeling disease prevalence and spread [54].

2.4. Recognizing States of Health

The ability to identify user health states as susceptible, infected, or recovered was also improved in some models for more accurate disease modeling. In a program called SimNest, social media users' health states were inferred from posts that mentioned symptoms. From these symptoms and health states, the disease progression of infected individuals can be modeled from infection to recovery using probabilities for recovery periods. This concept is also used in the hidden flu-state tweet model (HFSTM), where four states of health, namely, healthy, exposed, infected, and recovered, were detected and modeled. These states of health allow for realistic modeling of transitions into recovery based on time series and recovery probabilities [55]. Despite advancements in the ability to learn states of health from social media posts, the ability to detect and differentiate between health states was not incorporated into most disease-tracking and prediction models.

3. Sentiment Analysis

As previously mentioned, sentiment analysis provides insight into public opinion on outbreak-response-related topics [56], including attitudes toward stay-at-home orders, mask mandates, physical distancing, and vaccination. Understanding public perception is advantageous because it identifies groups of people who are dissatisfied or distrusting of the existing public health response and, therefore, provides an opportunity for targeted public education campaigns and outreach from public health or government officials to build trust [57]. Moreover, sentiment analysis facilitates more accurate disease forecasting through the proper representation of noncompliant groups of people in predictive models [9,58].

Machine learning and statistical models can be used as sentiment analysis tools to determine whether user posts have positive, negative, or neutral tones [12,58]. For example, statistical models were used to analyze the sentiment of Tweets [59]. To accomplish this, n-gram models constructed using samples of positive and negative tweets were used to classify new tweets based on the language they contained [59]. Although this study did not use infectious-disease-related tweets, the same method could be used for the sentiment analysis of disease outbreaks. Recently, the use of LSTM with global vectors for word representation (GloVe) by H. Menguri et al. in 2020 advanced the capabilities of sentiment analysis by adding the ability to interpret emojis and adding more emotional classifications based on Plutchik's wheel of emotion [32]. Common sentiment analysis tools include CoreNLP; SentiStrength; SentiCircles; Textblob; and recently, Valence Aware Dictionary and Sentiment Reasoner (VADER) [12]. In 2016, Stanford's CoreNLP was recognized as the most accurate sentiment analysis tool [52]. However, advanced neural network classifiers, such as bidirectional encoder representations from transformers (BERT) and recurrent neural networks (RNNs) were used in COVID-19-related studies [12,60,61]. These neural network classifiers are superior to traditional classifier models because of their accuracy and classification ability. Despite these advantages, the learning process for these classifiers is time-consuming and requires a large amount of data for training [60].

One way that sentiment analysis of social media data was used as a tool during the SARS-CoV-2 outbreak in China was to learn about public concerns by determining topics with negative sentiments toward COVID-19 [12]. From this study, four major themes of negative sentiment were identified in the hopes that governing bodies could address these concerns [12]. Similarly, in 2021, Yin et al. performed sentiment analysis on COVID-19-related tweets to assess public needs requiring urgent government action, such as regional food shortages and job insecurity [12]. One notable finding from the work of Dubey et al. (2020) was that people from different countries expressed different attitudes toward the same topics on Twitter, suggesting that a country's public health and welfare support response should correspond to the specific needs of its people.

Tracking changes in sentiment polarities can be useful for public health organizations to quickly identify and mitigate the spread of misinformation. A recent study found that spreading accurate information about COVID-19, thereby correcting misinformation, was associated with a flatter disease curve [62]. Sentiment analysis of influenza and measles found that a growing anti-vaccine community could influence the sentiment of others' Twitter posts through the spread of misinformation [63]. Twitter was identified as the most used social media platform by the anti-vaccine community [64]. This is especially concerning because automated bots were posting misinformation at an alarmingly high rate, thereby amplifying unrest during the COVID-19 pandemic by triggering anger transmission [65]. To address this issue, public health organizations must first recognize the problem. Timely recognition of this issue is crucial so that measures can be taken to combat the spread of misinformation and mitigate resulting noncompliance. Official organizations and public health units may also take advantage of social media to broadcast public announcements. Historically, official announcements were effective at decreasing the spread of misinformation and reducing negative sentiment toward topics associated with disease outbreaks, such as travel and vaccine concerns [44].

For public health organizations and policymakers, knowing when and how to communicate and enforce legislation can increase public cooperation and decrease disease spread [66]. Studies conducted on Sina Weibo identified a trend in public attention toward COVID-19-related topics, which peaked and then decayed exponentially [13,41]. This trend indicates that official announcements will garner the most attention and, therefore, success if made at the peak of the trend. Similar studies should be conducted on Twitter to understand and track attention patterns in different regions to improve the effectiveness of the messaging.

4. Measuring Disease Activity

Social media has a high potential for identifying disease outbreak patterns. When Twitter data was added to a linear autoregressive forecasting model, the nowcasting error decreased by 16.8 to 29.6 percent from the standard model, which used only historic data from the Centers for Disease Control and Prevention's (CDC's) influenza-like illness (ILI)Net to forecast disease [10]. Furthermore, influenza-related tweets were used to accurately estimate ILI incidence one to two weeks faster than official estimates made via case reporting [44]. Earlier identification of a spike in incidence facilitates an accelerated public health response, which has the potential to prevent an epidemic or pandemic.

4.1. Detecting and Predicting Influenza Rates

Recent social-media-based machine learning models were found to measure influenza incidence with greater performance than previous models. In 2022, a BERT base multilingual NLP model called Deepluenza was able to detect influenza cases from Twitter with 99% accuracy [48]. This model outperformed older k-nearest neighbor, decision tree, neural network, SVM, bidirectional LSTM, and CNN models [48].

Forecasting infectious disease transmission is possible using models with time series inputs, such as the Box–Jenkins model, which can make predictions based on historical trends. In 2013, Broniatowski et al. found that ILI-related tweets correlated with CDC-reported

ILI rates with a lag of one week using a Box–Jenkins model, even finding significance in correlations over seasonal trends [9]. In accordance, a 2018 review of the use of big data in predicting infectious diseases found that time series lags of seven days had the highest correlation for all infectious diseases [67].

Some infectious diseases, such as influenza, are endemic; therefore, machine learning models must be capable of detecting the incidence above the usual seasonal level. The C3 method of the Early Aberration Reporting System (EARS) is one such model, which detects unusual increases in relevant Twitter posts to identify disease outbreaks above usual levels [34]. Improvements to data prediction can be produced by more perceptive methods, such as Bayesian change point analysis, due to its ability to detect subtle changes in time series data more precisely than other aberration detection methods [68].

Statical models, including auto-regressive integrated moving average (ARIMA) models, were also used for influenza surveillance [69], but a study performed in Korea determined that DNN and LSTM models performed better at predicting infectious disease trends. The comparison of these two models was important because the DNN was identified to be a better model for predicting the minimum cases of disease occurrence, while the LSTM was better at predicting scenarios where maximum infection occurred [67]. Both cases are important to be aware of so that a swift public health response can be initiated.

A regional ILI prediction model, namely, the improved artificial-tree-optimizing back-propagation neural network (IAT-BPNN), integrates CDC and Twitter data [6]. This model uses an improved artificial tree algorithm to optimize the parameters of a neural network, allowing the model to be dynamically trained and calibrated [6]. Based on the IAT-BPNN, the improved particle swarm optimization algorithm with SVM regression (IPSO-SVR) also used an algorithm to optimize regression parameters. Additionally, IPSO-SVR improved disease-modeling capabilities by using inter-regional ILI data for better prediction of ILI transmission. In designing this model, a genetic algorithm that adopted cross-validation was employed. In comparison to other algorithms, the IPSO-SVR model had the fastest calculations and the least error when predicting regional ILI in the USA [70]. The use of inter-regional ILI data enabled IPSO-SVR to achieve higher precision and accuracy compared with the IAT-BPNN model [70]. Due to the success of using inter-regional data, future models should consider incorporating parameters of disease transmission and the influence of transportation regionally, nationally, and internationally. Current mathematical, statistical, and network-based models simulating disease transmission via transportation networks [71] could potentially be incorporated into disease-forecasting models to further improve performance.

Twitter data was also used in models that integrated data from multiple sources, including other social media platforms and hospital admittance records [7,27,45,72]. Recent evidence suggests that integrating data from multiple sources produces high-quality prediction models [45,72]. The least absolute shrinkage and selection operator (Lasso) algorithm is particularly useful for integrating data because of its ability to reduce the influence of redundant information in the resulting model [27,72]. Alternatively, a multilayer perceptron (MLP) model with backpropagation that used Twitter posts, ILI-related physician visits, and historical CDC data was able to forecast influenza two to three weeks ahead of the CDC surveillance system [7]. The software system DEFENDER also integrates data from multiple social media platforms, webpages, and news outlets to detect and forecast disease [45]. DEFENDER was improved using subset selection for choosing regression terms. This approach was found to be superior to Lasso [45]. SENTINEL similarly uses multiple data sources for surveillance [35]. However, compared with DEFENDER, news data was employed as a secondary source to calibrate confidence intervals for social media data [35]. Data integration can also be accomplished via bagging or ensemble learning. In 2021, a multi-view ensemble learning method was used to generate a more robust Twitter-based dataset on which sentiment analysis was performed [73]. This approach could be applied to infectious-disease-related Tweets for the same purpose. Ensemble learning was also used to forecast zoonotic diseases with long-term accuracy that is superior to non-ensemble approaches [74].

4.2. Potential for Detecting Novel Diseases and Strains

Novel diseases or strains may have similar symptoms to known diseases. This offers an opportunity for internet searches to serve as a sentinel for novel pathogens when an increase in searches related to a known disease is incongruent with expected disease prevalence. In such a scenario, the discrepancy indicates that there may be another pathogen causing similar symptoms that has yet to be identified. For example, there was an anomalous surge of searches including the keywords "SARS" or "pneumonia" on the Chinese search engine Baidu prior to the announcement that a novel respiratory pathogen was circulating [11]. This finding suggests that anomaly detection algorithms could provide early warning of novel diseases or strains. Furthermore, during the initial COVID-19 outbreaks, increased smell-related searches in areas across the globe correlated to infection epicenters [75], suggesting that an unusual increase in searches for specific symptoms could also indicate the emergence of a novel pathogen and provide insight into new illness characteristics [76]. Furthermore, Lim et al. (2017) used the uncertainty of symptoms of infectious diseases not yet formally recognized by public health organizations to their advantage. They proposed an unsupervised machine learning model to discover uncharacterized diseases using social media data by tracking users' symptoms and sentiment while clustering temporal information [33]. These recent advancements suggest that Twitter data has a high potential for not only detecting and predicting known disease rates but also for detecting unidentified diseases. The use of a bottom-up approach for disease detection without prior information, such as names and associated symptoms, also provides a groundwork for future models that can detect early signs of novel diseases.

There is also a need to maintain the efficacy of existing surveillance systems in light of a novel disease outbreak. Recently, CALI-Net was created by adapting a previous forecasting model to address a major concern over the efficacy of influenza surveillance following the introduction of COVID-19 [77]. This neural transfer learning tool was adapted to model influenza disease using social media data while accounting for COVID-19, which presents with similar symptoms [77]. CALI-Net's performance is comparable to regular influenza surveillance models.

Table 1 summarizes each primary application of social-media-based infectious disease modeling, as well as associated sub-categories and relevant literature for each.

Table 1. Summary of relevant literature.

Topic	Sub-Topic	Relevant Literature
Disease surveillance	Disease surveillance	[14,15,22,23]
	Information retrieval and pre-processing	[15,22,24–35]
	Determining post relevancy and their applications	[4,8,9,12,13,22,24,25,27,28,35–47]
	Collecting and using location data	[3,9,16,25,34,36,44,48–54]
	Recognizing states of health	[55]
Sentiment analysis	Sentiment analysis	[9,12,13,32,41,44,52,56–66]
Measuring disease activity	Measuring disease activity	[10,44]
	Detecting and predicting influenza rates	[6,7,9,27,35,45,48,67–74]
	Potential for detecting novel diseases/strains	[11,33,75–77]

5. Challenges of Using Social Media for Implementing Disease Models

Although Twitter is the most widely used social media platform in epidemiological studies, the demographics of Twitter users, whose data are being used to inform disease prevalence, creates a source of bias. Twitter is not meant to be used by children and has more young adult users than people over the age of 65; therefore, the results of studies that use Twitter as a data source likely overrepresent young adults and underrepresent children and older adults [35]. The cohort of individuals older than 65 years of age is of particular interest in the context of infectious disease surveillance due to the increased risk of severe illness caused by emerging infectious diseases, such as COVID-19. However, people in this cohort are less likely than younger adults to share their health status publicly, making social-media-based research difficult [14]. This bias also affects other social media platforms but may be mitigated over time as the amount of available data increases exponentially.

There may also be generational differences in the connotations of certain phrases and emojis, and thus, the intended meaning of a social media post may be dependent on the age of the writer. Deep learning methods may be a solution to understanding the connotations of certain phrases or emojis depending on the age group [32]. Deep learning methods may be useful in detecting sarcasm as well. Currently, there is no reliable and consistent approach to handling sarcasm; text-mining models either cannot detect sarcasm at all or can only poorly detect it. In one study, sarcasm detection relied on the use of quotation marks [35], but in another study, phrases in quotation marks were removed before data analysis because they were considered to be reported speech and not direct speech from the users themselves [61].

Another challenge is the disproportionate representations of some special interest groups through the use of automated bots to amplify their social media presence. For instance, the group of people who were opposed to COVID-19 vaccination, known as the anti-vax community, was relatively small. However, the use of automated bots created the appearance of a much larger and more active community on Twitter, which could bias sentiment analyses [64,65]. One study identified trends among bot accounts on Twitter, which can potentially be used to screen accounts to prevent the inclusion of posts from bot accounts during analysis. These trends include new accounts containing numbers in their usernames, high volumes of retweets, and a high ratio of accounts followed to following accounts [66,78].

Data analysis suffers when data are sparse, which poses a challenge in rural locations where internet access is limited or in regions where censorship prevents access to the social media platform from which data are being collected. Even in regions where Twitter is frequently used, there are fluctuations in posting patterns, which can affect infection and recovery modeling [44]. Furthermore, the limited number of geo-located posts relies on the use of secondary tools to accurately determine location. Across larger geographical distances, such as between countries, the main challenge for international disease modeling is that few existing models accommodate multiple languages with a high degree of accuracy. Lastly, heightened privacy settings on social media accounts may continue to limit the data that is available for analysis.

6. Conclusions

Social media data are valuable for tracking and understanding disease transmission, evaluating public health concerns, and ascertaining public attitudes toward disease outbreaks and control policies. Twitter posts in particular were incorporated into many statistical and machine learning models for monitoring disease spread, which are more efficient and inexpensive than traditional methods. Here, we critically assessed the existing body of literature on infectious disease modeling using social media data and identified three primary applications: disease surveillance, sentiment analysis, and measuring disease activity. For each, we specified subtopics where applicable and discussed popular and promising approaches. Finally, we identified challenges associated with the use of social media data to model infectious disease outbreaks and the resulting consequences.

Improving the analysis of social media: As machine learning and deep learning techniques become more sophisticated, the interpretations of social media data will become more accurate and efficient. There is the potential to use approaches such as transfer learning and semi-supervised learning, which are advantageous when limited labeled data are available.

Use of social media data by public health units: improvements in contextual analyses of sentiments may provide public health units with valuable data to make more informed and directed decisions about control strategies, public health announcements, and plans to mitigate misinformation.

Use of social media resources other than Twitter: recent research primarily focused on algorithms that apply Twitter data to predict disease trends, but other social media platforms, such as Instagram, are now being investigated as well [14].

Development of decision support systems: Recent advancements in disease modeling are already beginning to consider the temporal progression of disease infection to recovery and external factors contributing to disease spread. These new models collect data from different sources and combine and analyze them. This has created more robust models of disease forecasting. Future works should aim to create a model that comprehensively incorporates additional risk factors, such as major transportation routes, health data, vaccination rates, and evolving keywords in multiple languages to detect and predict disease outbreaks and understand public sentiment more accurately.

Expanding the use of social media data and using it in decision-making: The ability to obtain insight from social media data saves time, resources, and lives. Earlier implementation of an infectious disease outbreak response plan can reduce the destruction caused by an outbreak. Thus, promoting the incorporation of social media data into infectious disease modeling facilitates optimal response times and can enhance public cooperation during infectious disease outbreaks will be beneficial to both the public and the healthcare system.

Author Contributions: Conceptualization, A.W. and R.D.; methodology, A.W. and R.D.; validation, S.Y. and S.S.; writing—original draft preparation, A.W.; writing—review and editing, R.D., E.M., S.Y. and S.S.; supervision, R.D. and S.S.; funding acquisition, R.D. and S.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported in part by the University of Guelph's Food from Thought initiative and the Ontario Ministry of Agriculture and Rural Affairs' Alliance Tier I Research Program.

Conflicts of Interest: The authors declare that there is no conflict of interest.

References

1. Bollen, J.; Mao, H.; Zeng, X. Twitter mood predicts the stock market. *J. Comput. Sci.* **2011**, *2*, 1–8. [CrossRef]
2. McGough, S.F.; Brownstein, J.S.; Hawkins, J.B.; Santillana, M. Forecasting Zika Incidence in the 2016 Latin America Outbreak Combining Traditional Disease Surveillance with Search, Social Media, and News Report Data. *PLoS Negl. Trop. Dis.* **2017**, *11*, e0005295. [CrossRef] [PubMed]
3. Aslam, A.A.; Tsou, M.-H.; Spitzberg, B.H.; An, L.; Gawron, J.M.; Gupta, D.K.; Peddecord, K.M.; Nagel, A.C.; Allen, C.; Yang, J.-A.; et al. The reliability of tweets as a supplementary method of seasonal influenza surveillance. *J. Med. Internet Res.* **2014**, *16*, e250. [CrossRef]
4. Jain, V.K.; Kumar, S. An Effective Approach to Track Levels of Influenza-A (H1N1) Pandemic in India Using Twitter. *Procedia Comput. Sci.* **2015**, *70*, 801–807. [CrossRef]
5. Aiello, A.E.; Renson, A.; Zivich, P.N. Social media–and internet-based disease surveillance for public health. *Annu. Rev. Public Health* **2020**, *41*, 101–118. [CrossRef]
6. Hu, H.; Wang, H.; Wang, F.; Langley, D.; Avram, A.; Liu, M. Prediction of influenza-like illness based on the improved artificial tree algorithm and artificial neural network. *Sci. Rep.* **2018**, *8*, 4895. [CrossRef]
7. Lee, K.; Agrawal, A.; Choudhary, A. Forecasting Influenza Levels Using Real-Time Social Media Streams. In Proceedings of the 2017 IEEE International Conference on Healthcare Informatics (ICHI), Park City, UT, USA, 23–26 August 2017. [CrossRef]
8. Aramaki, E.; Maskawa, S.; Morita, M. Twitter catches the flu: Detecting influenza epidemics using Twitter. In Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, Edinburgh, Scotland, 27–31 July 2011; pp. 1568–1576.
9. Broniatowski, D.A.; Paul, M.J.; Dredze, M. National and local influenza surveillance through twitter: An analysis of the 2012-2013 influenza epidemic. *PLoS ONE* **2013**, *8*, e83672. [CrossRef]
10. Paul, M.J.; Dredze, M.; Broniatowski, D. Twitter Improves Influenza Forecasting. *PLoS Curr.* **2014**, *6*. [CrossRef]
11. Dai, Y.; Wang, J. Identifying the outbreak signal of covid-19 before the response of the traditional disease monitoring system. *PLoS Negl. Trop. Dis.* **2020**, *14*, e0008758. [CrossRef]
12. Wang, T.; Lu, K.; Chow, K.P.; Zhu, Q. COVID-19 Sensing: Negative Sentiment Analysis on Social Media in China via BERT Model. *IEEE Access* **2020**, *8*, 138162–138169. [CrossRef]
13. Zhao, Y.; Cheng, S.; Yu, X.; Xu, H. Chinese public's attention to the COVID-19 epidemic on social media: Observational descriptive study. *J. Med. Internet Res.* **2020**, *22*, e18825. [CrossRef]

14. Gupta, A.; Katarya, R. Social media based surveillance systems for healthcare using machine learning: A systematic review. *J. Biomed. Inform.* **2020**, *108*, 103500. [CrossRef] [PubMed]
15. Gencoglu, O.; Ermes, M. Predicting the Flu from Instagram. *arXiv* **2018**. [CrossRef]
16. Sharpe, J.D.; Hopkins, R.S.; Cook, R.L.; Striley, C.W. Evaluating Google, Twitter, and Wikipedia as tools for influenza surveillance using Bayesian change point analysis: A comparative analysis. *JMIR Public Health Surveill.* **2016**, *2*, e161. [CrossRef]
17. Melton, C.A.; White, B.M.; Davis, R.L.; Bednarczyk, R.A.; Shaban-Nejad, A. Fine-tuned Sentiment Analysis of COVID-19 Vaccine–Related Social Media Data: Comparative Study. *J. Med. Internet Res.* **2022**, *24*, e40408. [CrossRef]
18. Mohammed, I.A.B.; Abbas, A.S. Twitter APIs for Collecting Data of Influenza Viruses, A Systematic Review. In Proceedings of the 2021 International Conference on Communication & Information Technology (ICICT), Barash, Iraq, 5–6 June 2021.
19. Alamoodi, A.H.; Zaidan, B.B.; Zaidan, A.A.; Albahri, O.S.; Mohammed, K.I.; Malik, R.Q.; Almadi, E.M.; Chyad, M.A.; Tareq, Z.; Albahri, A.S.; et al. Sentiment analysis and its applications in fighting COVID-19 and infectious diseases: A systematic review. *Expert Syst. Appl.* **2021**, *167*, 114155. [CrossRef] [PubMed]
20. Sooknanan, J.; Mays, N. Harnessing Social Media in the Modelling of Pandemics—Challenges and Opportunities. *Bull. Math. Biol.* **2021**, *83*, 57. [CrossRef] [PubMed]
21. Findling, M.G.; Blendon, R.J.; Benson, J.M. Polarized Public Opinion About Public Health During the COVID-19 Pandemic: Political Divides and Future Implications. *JAMA Health Forum* **2022**, *3*, e220016. [CrossRef]
22. Yousefinaghani, S.; Dara, R.A.; Poljak, Z.; Sharif, S. A decision support framework for prediction of avian influenza. *Sci. Rep.* **2020**, *10*, 19011. [CrossRef]
23. Batrinca, B.; Treleaven, P.C. Social media analytics: A survey of techniques, tools and platforms. *AI Soc.* **2015**, *30*, 89–116. [CrossRef]
24. Alkouz, B.; Aghbari, Z.A.; Abawajy, J.H. Tweetluenza: Predicting flu trends from twitter data. *Big Data Min. Anal.* **2019**, *2*, 273–287. [CrossRef]
25. Allen, C.; Tsou, M.H.; Aslam, A.; Nagel, A.; Gawron, J.M. Applying GIS and machine learning methods to twitter data for multiscale surveillance of influenza. *PLoS ONE* **2016**, *11*, e0157734. [CrossRef] [PubMed]
26. Joseph, K.; Landwehr, P.M.; Carley, K.M. Two 1%s Don't Make a Whole: Comparing Simultaneous Samples from Twitter's Streaming API. In *Social Computing, Behavioral-Cultural Modeling and Prediction. SBP 2014. Lecture Notes in Computer Science*; Kennedy, W.G., Agarwal, N., Yang, S.J., Eds.; Springer: Cham, Switzerland, 2014; Volume 8393.
27. Woo, H.; Cho, Y.; Shim, E.; Lee, J.-K.; Lee, C.-G.; Kim, S.H. Estimating influenza outbreaks using both search engine query data and social media data in South Korea. *J. Med. Internet Res.* **2016**, *18*, e177. [CrossRef]
28. Yousefinaghani, S.; Dara, R.; Poljak, Z.; Bernardo, T.M.; Sharif, S. The assessment of Twitter's potential for outbreak detection: Avian influenza case study. *Sci. Rep.* **2019**, *9*, 18147. [CrossRef] [PubMed]
29. Havrlant, L.; Kreinovich, V. A simple probabilistic explanation of term frequency-inverse document frequency (tf-idf) heuristic (and variations motivated by this explanation). *Int. J. Gen. Syst.* **2017**, *46*, 27–36. [CrossRef]
30. Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent Dirichlet allocation. *J. Mach. Learn. Res.* **2003**, *3*, 339–1022.
31. Dubey, A.D. Twitter Sentiment Analysis during COVID19 Outbreak. *SSRN Electron. J.* **2020**. [CrossRef]
32. Imran, A.S.; Daudpota, S.M.; Kastrati, Z.; Batra, R. Cross-cultural polarity and emotion detection using sentiment analysis and deep learning on covid-19 related tweets. *IEEE Access* **2020**, *8*, 181074–181090. [CrossRef]
33. Lim, S.; Tucker, C.S.; Kumara, S. An unsupervised machine learning model for discovering latent infectious diseases using social media data. *J. Biomed. Inform.* **2017**, *66*, 82–94. [CrossRef]
34. Cuomo, R.E.; Purushothaman, V.; Li, J.; Cai, M.; Mackey, T.K. Sub-national longitudinal and geospatial analysis of COVID-19 tweets. *PLoS ONE* **2020**, *15*, e0241330. [CrossRef]
35. Șerban, O.; Thapen, N.; Maginnis, B.; Hankin, C.; Foot, V. Real-time processing of social media with SENTINEL: A syndromic surveillance system incorporating deep learning for health classification. *Inf. Process Manag.* **2019**, *56*, 1166–1184. [CrossRef]
36. Lopreite, M.; Panzarasa, P.; Puliga, M.; Riccaboni, M. Early warnings of COVID-19 outbreaks across Europe from social media. *Sci. Rep.* **2021**, *11*, 2147. [CrossRef] [PubMed]
37. De las Heras-Pedrosa, C.; Sánchez-Núñez, P.; Peláez, J.I. Sentiment analysis and emotion understanding during the COVID-19 pandemic in Spain and its impact on digital ecosystems. *Int. J. Environ. Res. Public Health* **2020**, *17*, 5542. [CrossRef]
38. Wakamiya, S.; Kawai, Y.; Aramaki, E. Twitter-based influenza detection after flu peak via tweets with indirect information: Text mining study. *JMIR Public Health Surveill.* **2018**, *4*, e65. [CrossRef] [PubMed]
39. Qin, L.; Sun, Q.; Wang, Y.; Wu, K.-F.; Chen, M.; Shia, B.-C.; Wu, S.-Y. Prediction of number of cases of 2019 novel coronavirus (COVID-19) using social media search index. *Int. J. Environ. Res. Public Health* **2020**, *17*, 2365. [CrossRef] [PubMed]
40. Seo, D.-W.; Jo, M.-W.; Sohn, C.H.; Shin, S.-Y.; Lee, J.; Yu, M.; Kim, W.Y.; Lim, K.S.; Lee, S.-I. Cumulative query method for influenza surveillance using search engine data. *J. Med. Internet Res.* **2014**, *16*, e289. [CrossRef]
41. Cui, H.; Kertész, J. Attention dynamics on the Chinese social media Sina Weibo during the COVID-19 pandemic. *EPJ Data Sci.* **2021**, *10*, 8. [CrossRef]
42. Paul, M.J.; Dredze, M. *A Model for Mining Public Health Topics from Twitter*; Technical Report; Johns Hopkins University: Baltimore, MD, USA, 2011.
43. Paul, M.J.; Dredze, M. You Are What You Tweet: Analyzing Twitter for Public Health. *Proc. Int. AAAI Conf. Web Soc. Media* **2021**, *5*, 265–272. [CrossRef]

44. Signorini, A.; Segre, A.M.; Polgreen, P.M. The use of Twitter to track levels of disease activity and public concern in the U.S. during the influenza A H1N1 pandemic. *PLoS ONE* **2011**, *6*, e19467. [CrossRef]
45. Thapen, N.; Simmie, D.; Hankin, C.; Gillard, J. DEFENDER: Detecting and forecasting epidemics using novel data-analytics for enhanced response. *PLoS ONE* **2016**, *11*, e0155417. [CrossRef]
46. Malla, S.; Alphonse, P.J.A. COVID-19 outbreak: An ensemble pre-trained deep learning model for detecting informative tweets. *Appl. Soft Comput.* **2021**, *107*, 107495. [CrossRef] [PubMed]
47. Alessa, A.; Faezipour, M. Preliminary flu outbreak prediction using twitter posts classification and linear regression with historical centers for disease control and prevention reports: Prediction framework study. *JMIR Public Health Surveill.* **2019**, *5*, e12383. [CrossRef]
48. Yuan, M.; Liu, T.; Yang, C. Exploring the Relationship among Human Activities, COVID-19 Morbidity, and At-Risk Areas Using Location-Based Social Media Data: Knowledge about the Early Pandemic Stage in Wuhan. *Int. J. Environ. Res. Public Health* **2022**, *19*, 6523. [CrossRef] [PubMed]
49. Essam, N.; Moussa, A.M.; Elsayed, K.M.; Abdou, S.; Rashwan, M.; Khatoon, S.; Hasan, M.M.; Asif, A.; Alshamari, M.A. Location Analysis for Arabic COVID-19 Twitter Data Using Enhanced Dialect Identification Models. *Appl. Sci.* **2021**, *11*, 11328. [CrossRef]
50. Abd-Alrazaq, A.; Alhuwail, D.; Househ, M.; Hai, M.; Shah, Z. Top concerns of tweeters during the COVID-19 pandemic: A surveillance study. *J. Med. Internet Res.* **2020**, *22*, e19016. [CrossRef] [PubMed]
51. Collier, N.; Doan, S.; Kawazoe, A.; Goodwin, R.M.; Conway, M.; Tateno, Y.; Ngo, Q.-H.; Dien, D.; Kawtrakul, A.; Takeuchi, K.; et al. BioCaster: Detecting public health rumors with a Web-based text mining system. *Bioinformatics* **2008**, *24*, 2940–2941. [CrossRef]
52. Byrd, K.; Mansurov, A.; Baysal, O. Mining twitter data for influenza detection and surveillance. In Proceedings of the 2016 IEEE/ACM International Workshop on Software Engineering in Healthcare Systems (SEHS), Austin, TX, USA, 14–15 May 2016. [CrossRef]
53. Brownstein, J.S.; Freifeld, C.C.; Reis, B.Y.; Mandl, K.D. Surveillance sans frontières: Internet-based emerging infectious disease intelligence and the HealthMap project. *PLoS Med.* **2008**, *5*, e151. [CrossRef]
54. Thapen, N.; Simmie, D.; Hankin, C. The early bird catches the term: Combining twitter and news data for event detection and situational awareness. *J. Biomed. Semant.* **2016**, *7*, 61. [CrossRef] [PubMed]
55. Chen, L.; Hossain, K.S.M.T.; Butler, P.; Ramakrishnan, N.; Prakash, B.A. Flu Gone Viral: Syndromic Surveillance of Flu on Twitter Using Temporal Topic Models. In Proceedings of the 2014 IEEE International Conference on Data Mining, Shenzhen, China, 14–17 December 2014.
56. Jain, P.K.; Pamula, R.; Srivastava, G. A systematic literature review on machine learning applications for consumer sentiment analysis using online reviews. *Comput. Sci. Rev.* **2021**, *41*, 100413. [CrossRef]
57. Jain, P.K.; Srivastava, G.; Lin, J.C.-W.; Pamula, R. Unscrambling Customer Recommendations: A Novel LSTM Ensemble Approach in Airline Recommendation Prediction Using Online Reviews. *IEEE Trans. Comput. Soc. Syst.* **2022**, *9*, 1777–1784. [CrossRef]
58. Bhat, M.; Qadri, M.; Beg, N.-u.-A.; Kundroo, M.; Ahanger, N.; Agarwal, B. Sentiment analysis of social media response on the Covid19 outbreak. *Brain Behav. Immun.* **2020**, *87*, 136–137. [CrossRef] [PubMed]
59. Saeed, K.; Homenda, W.; Chaki, R. *Towards the Exploitation of Statistical Language Models for Sentiment Analysis of Twitter Posts*; Springer International Publishing: Cham, Switzerland, 2017; Volume 10244, pp. 253–263.
60. Nemes, L.; Kiss, A. Social media sentiment analysis based on COVID-19. *J. Inf. Telecommun.* **2021**, *5*, 1–15. [CrossRef]
61. Klein, A.Z.; Magge, A.; O'Connor, K.; Amaro, J.I.F.; Weissenbacher, D.; Hernandez, G.G. Toward Using Twitter for Tracking COVID-19: A Natural Language Processing Pipeline and Exploratory Data Set. *J. Med. Internet Res.* **2021**, *23*, e25314. [CrossRef] [PubMed]
62. Wątroba, P.; Bródka, P. Influence of Information Blocking on the Spread of Virus in Multilayer Networks. *Entropy* **2023**, *25*, 231. [CrossRef] [PubMed]
63. Santamaría, L.P.; Tuñas, J.M.; Peces-Barba, D.F.; Jaramillo, A.; Cotarelo, M.; Menasalvas, E.; Fernández, A.C.; Arce, A.; de Miguel, A.G.; González, A.R. Influenza and Measles-MMR: Two case study of the trend and impact of vaccine-related Twitter posts in Spanish during 2015–2018. *Hum. Vaccines Immunother.* **2022**, *18*, 1–16. [CrossRef]
64. Ortiz-Sánchez, E.; Velando-Soriano, A.; Pradas-Hernández, L.; Vargas-Román, K.; Gómez-Urquiza, J.L.; la Fuente, G.A.C.-D.; Albendín-García, L. Analysis of the anti-vaccine movement in social networks: A systematic review. *Int. J. Environ. Res. Public Health* **2020**, *17*, 5394. [CrossRef]
65. Shi, W.; Liu, D.; Yang, J.; Zhang, J.; Wen, S.; Su, J. Social bots' sentiment engagement in health emergencies: A topic-based analysis of the covid-19 pandemic discussions on twitter. *Int. J. Environ. Res. Public Health* **2020**, *17*, 8701. [CrossRef]
66. Yousefinaghani, S.; Dara, R.; Mubareka, S.; Papadopoulos, A.; Sharif, S. An analysis of COVID-19 vaccine sentiments and opinions on Twitter. *Int. J. Infect. Dis.* **2021**, *108*, 256–262. [CrossRef]
67. Chae, S.; Kwon, S.; Lee, D. Predicting infectious disease using deep learning and big data. *Int. J. Environ. Res. Public Health* **2018**, *15*, 1596. [CrossRef]
68. Kass-Hout, T.A.; Xu, Z.; McMurray, P.; Park, S.; Buckeridge, D.L.; Brownstein, J.S.; Finelli, L.; Groseclose, S.L. Application of change point analysis to daily influenza-like illness emergency department visits. *J. Am. Med. Inform. Assoc.* **2012**, *19*, 1075–1081. [CrossRef]
69. Paul, S.; Mgbere, O.; Arafat, R.; Yang, B.; Santos, E. Modeling and Forecasting Influenza-like Illness (ILI) in Houston, Texas Using Three Surveillance Data Capture Mechanisms. *Online J. Public Health Inform.* **2017**, *9*, e187. [CrossRef] [PubMed]

70. Xue, H.; Bai, Y.; Hu, H.; Liang, H. Regional level influenza study based on Twitter and machine learning method. *PLoS ONE* **2019**, *14*, e0215600. [CrossRef] [PubMed]
71. Li, J.; Xiang, T.; He, L. Modeling epidemic spread in transportation networks: A review. *J. Traffic Transp. Eng. Engl. Ed.* **2021**, *8*, 139–152. [CrossRef]
72. Santillana, M.; Nguyen, A.T.; Dredze, M.; Paul, M.J.; Nsoesie, E.O.; Brownstein, J.S. Combining Search, Social Media, and Traditional Data Sources to Improve Influenza Surveillance. *PLoS Comput. Biol.* **2015**, *11*, e1004513. [CrossRef] [PubMed]
73. Ye, X.; Dai, H.; Dong, L.; Wang, X. Multi-view ensemble learning method for microblog sentiment classification. *Expert Syst. Appl.* **2021**, *166*, 113987. [CrossRef]
74. Sharma, V.C.; Frankenfield, D.; Gupta, A.; Singh, R.K. Ensemble Approach for Zoonotic Disease Forecasting Using Machine Learning Techniques. *Int. J. Bus. Anal. Intell.* **2015**, *3*, 11–24. [CrossRef]
75. Walker, A.; Hopkins, C.; Surda, P. Use of Google Trends to investigate loss-of-smell-related searches during the COVID-19 outbreak. *Int. Forum Allergy. Rhinol.* **2020**, *10*, 839–847. [CrossRef]
76. Heymann, D.L.; Rodier, G. Global Surveillance, National Surveillance, and SARS. *Emerg. Infect. Dis.* **2004**, *10*, 173–175. [CrossRef]
77. Rodríguez, A.; Muralidhar, N.; Adhikari, B.; Tabassum, A.; Ramakrishnan, N.; Prakash, B.A. Steering a Historical Disease Forecasting Model Under a Pandemic: Case of Flu and COVID-19. *arXiv* **2020**. [CrossRef]
78. Tavoschi, L.; Quattrone, F.; D'Andrea, E.; Ducange, P.; Vabanesi, M.; Marcelloni, F.; Lopalco, P.L. Twitter as a sentinel tool to monitor public opinion on vaccination: An opinion mining analysis from September 2016 to August 2017 in Italy. *Hum. Vaccines Immunother.* **2020**, *16*, 1062–1069. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.