Digital Epidemiology in Action: A Cross-Platform Review of Social Media and Internet-Based Surveillance for Infectious Disease Outbreaks

Esha Madamalla[1]

(Student)[1]
[1]Thomas Jefferson High School for Science and Technology,
6560 Braddock Rd, Alexandria, VA, 22312, USA

Publication Date: 2025/07/23

Abstract: This review evaluates how social media and internet-based platforms can enhance infectious disease surveillance by supplementing traditional epidemiological methods. Drawing from 15 studies published between 2015 and 2023, the paper examines platforms such as Twitter, Facebook, Google Trends, Reddit, Wikipedia, Baidu, and Sina Weibo in the context of disease outbreaks like COVID-19, Influenza, Zika, and Ebola across countries including the U.S., China, Brazil, and Saudi Arabia. Findings show that spikes in user activity, such as tweets, search queries, and online discussions, often precede official case reporting by several days to weeks, offering valuable lead time for public health response. Twitter excelled in real-time detection, Google Trends in population-level awareness, and Reddit and Facebook in sentiment and misinformation tracking. Multi-platform AI models demonstrated improved accuracy over single-platform approaches. However, challenges such as demographic bias, language limitations, and misinformation remain. The study concludes that digital platforms are most effective when integrated into hybrid systems that combine social, clinical, and environmental data for more timely and adaptive disease monitoring.

Keywords: *Computational Biolgy and Bioinformatics; Computational Epidemiology; Digital Epidemiology; Social Media Surveillance; Infectious Disease Monitoring.*

How to Cite: Esha Madamalla (2025) Digital Epidemiology in Action: A Cross-Platform Review of Social Media and Internet-Based Surveillance for Infectious Disease Outbreaks. *International Journal of Innovative Science and Research Technology,* 10(7), 1768-1776. https://doi.org/10.38124/ijisrt/25jul887

I. INTRODUCTION

Public health surveillance has been at the forefront of identifying, containing, and mitigating infectious disease outbreaks. Traditionally, surveillance relied on structured epidemiological tools such as physician case reports, laboratory confirmations, hospitalization logs, and syndromic surveillance (e.g., tracking respiratory symptoms in ER visits). While these methods are robust, they often suffer from significant time delays, high resource demands, and limited flexibility in rapidly evolving situations (Fung et al. 2). The recent proliferation of big data, artificial intelligence, and ubiquitous digital platforms have introduced new opportunities to augment traditional methods. Researchers have increasingly explored the use of unstructured, real-time digital data, including search engine queries, user-generated social media content, and participatory reporting apps to provide faster, scalable, and often more granular disease insights (Aiello et al. 5; Wang et al. 12).

Key metrics of traditional outbreak tracking have included lab-confirmed cases, mortality rates, hospital admissions, and population-based surveys. However, these indicators frequently underrepresent true disease prevalence due to underreporting and delayed confirmations. The 2002 SARS outbreak, for instance, was marked by significant information suppression in early weeks, which delayed international alerts and response coordination (Fung et al. 3). Similarly, during the 2009 H1N1 pandemic, surveillance systems failed to keep pace with real-time transmission dynamics, especially in low-resource or rural areas, where lab infrastructure was limited (Wang et al. 11). Efforts to modernize surveillance led to innovations like Google Flu Trends (GFT), launched in 2008, which monitored search behavior to estimate influenza prevalence. While pioneering, GFT eventually failed due to overestimation of influenza-like illness (ILI) cases and a lack of adaptability to media-driven spikes and evolving search behaviors (Aiello et al. 5; Wang et al. 12). Initially hailed as a breakthrough, GFT claimed to accurately track influenza-like illness (ILI) in real-time based on keyword frequency. However, during the 2012–2013 flu season, GFT dramatically overestimated

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25jul887

ILI prevalence by up to 50% due to overfitting and its failure to account for media influence on search behavior. These shortcomings emphasized the need for more adaptive, hybrid systems combining structured and unstructured data as well as the limitations of relying solely on behavioral search data in the absence of ground-truth validation from clinical sources.

The rise of internet-connected platforms spurred a new generation of disease surveillance tools. Beyond GFT, participatory tracking systems like FluNearYou and Influenzanet enabled individuals to self-report symptoms, creating citizen-powered early warning systems (Aiello et al. 4). Around 2010, researchers began applying natural language processing (NLP) and machine learning to social media data, especially Twitter, to identify illness-related behavior, sentiment, and misinformation (Fung et al. 2; Al-Garadi et al. 3). Studies found strong correlations between influenza-related tweets and official influenza-like illness (ILI) trends (Broniatowski et al. 5; Al-Garadi et al. 3). Platforms like Wikipedia, Reddit, and Sina Weibo also entered the surveillance landscape. Sina Weibo was used in China to model H7N9 successfully and COVID-19 outbreaks (Wang et al. 11), while Wikipedia access logs helped flu trends in the U.S. (Aiello et al. 4). These systems collectively illustrated the viability of social media and internet data as complementary surveillance layers.

The COVID-19 pandemic catalyzed a global shift in how disease surveillance was conceptualized and executed. Platforms like Twitter and Reddit became critical tools not just for public communication but also for real-time epidemiological modeling. Research showed that tweet content embeddings when modeled with LSTM architectures could predict COVID-19 case trends up to 8 days in advance with high accuracy (R^2 > 0.7) (Kazijevs et al. 7). Additionally, another paper demonstrated the importance of dialect-sensitive NLP, applying deep learning to Arabic Twitter to classify tweets about COVID-19 and influenza, achieving F1 scores over 90%, a metric that balances precision and recalls to assess classification accuracy, and geolocation prediction accuracy of 54% (Alsudias & Rayson 13). This showcased social media's capacity for fine-grained, linguistic, and regional disease tracking. Public health agencies like The World Health Organization (WHO) also embraced social media usage Twitter to combat COVID-19 misinformation and issue safety alerts. The Centers for Disease Control and Prevention (CDC) monitored social media platforms to identify and counteract vaccine myths. Johns Hopkins University COVID-19 dashboard became a global data hub by integrating web, news, and social indicators (Wang et al. 11). Tools like Epicosm (Tanner et al. 10) enabled researchers to securely link longitudinal cohort data with Twitter timelines, providing mental health insights. Platforms such as DEFENDER, SENTINEL, and CALI-Net combined social data, case data, and news streams to generate predictive, adaptive, multi-source surveillance models (Wang et al. 11).

This research paper seeks to evaluate this central question: How effective is social media data in tracking, predicting, and contextualizing disease outbreaks globally, and how can it be integrated with traditional epidemiological tools to enhance real-time public health surveillance? This evaluation is critical because social media platforms have become increasingly embedded in public discourse during epidemics. Furthermore, their role in structured disease modeling and public health infrastructure remains inconsistently applied and poorly understood. This review seeks to fill several gaps in our current understanding by synthesizing recent research on digital infectious disease surveillance methods, identifying where social media data complements or diverges from traditional metrics and surveillance tools, and evaluating how AI-driven models can filter noise, detect misinformation, and provide geographically and linguistically sensitive insights. By comparing early theoretical expectations with pandemic-era implementations, this study offers a critical analysis of the promises, limitations, and future directions of integrating social media into mainstream epidemiological practice.

II. METHODS

This review synthesized research published between 2015 and 2023 to evaluate how social media and internet-based platforms were utilized for infectious disease surveillance and prediction globally. The scope of the review encompassed studies that examined major global infectious disease outbreaks such as COVID-19, SARS-CoV-2, Influenza, Ebola, Zika, and MERS.

The eligibility criteria for finding papers were as follows: (1) was published in English, (2) used a text-based social media or search platform as a data source, and (3) addressed the use of these platforms for disease surveillance, outbreak prediction, or public health communication for an infectious disease outbreak.

Only platforms that allowed users to create and share their own content, such as posts,comments, or search queries, were considered eligible for inclusion in this review. In addition, these platforms had to provide some form of accompanying metadata that could support analysis, such as the time a post was made (timestamp), the user's location (when available), or how others interacted with the content (e.g., likes, shares, or comments). This ensured that the data could be used for tracking how diseases spread across time and space or how public engagement evolved. Qualitative studies were also included in the review, but only if they examined these digital signals specifically in the context of public health, either by analyzing how social media data was used to monitor disease trends or by investigating the spread and impact of health misinformation online.

Papers were identified through systematic searches across multiple academic databases and repositories, including PubMed, Scopus, Web of Science, IEEE Xplore, and Google Scholar. Keywords used in the search included combinations of: "digital epidemiology," "social media,"

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25jul887

"disease surveillance," "Twitter," "Google Trends," "Reddit," "misinformation," "COVID-19," "Influenza," "Ebola," "Zika," "Dengue," "MERS," "SARS," "H1N1," "H7N9," and "infectious disease modeling." Reference lists from key review papers were also examined to ensure comprehensive coverage. Studies were selected based on their methodological alignment with the review's objectives. Each study was screened to determine whether it used social media data as a predictive indicator or as part of a real-time or near-real-time surveillance framework. Studies that only described public perception on social media without linking it to epidemiological trends were excluded.

To ensure relevance and consistency with the aims of this paper, only studies that employed social media data as a predictive measure or as part of a real-time monitoring framework were included. Special attention was given to research that compared social media–driven analyses with traditional epidemiological data sources, such as those from government health agencies or clinical settings. To address the larger research question, the papers were examined with the following sub-questions in mind:

- How have data collection and modeling practices evolved through social media surveillance?
- What are the methodological and ethical considerations in using social media as a data source?
- To what extent do prior theoretical claims about digital epidemiology hold up during real-world stress tests like COVID-19
- How can researchers design future-ready systems that combine social, clinical, and environmental data to provide robust, real-time surveillance?

III. RESULTS

A total of 15 studies were identified and included in this review. These papers were published between 2015 and 2023, capturing the evolution of digital epidemiology before the COVID-19 pandemic and its applications during the global crisis. In terms of country representation, the United States was the most frequently studied location, particularly in analyses involving Twitter, Reddit, and Google Trends. Several studies also focused on China, utilizing platforms such as Baidu Index and Sina Weibo to monitor the early spread of COVID-19 in Wuhan and Hubei Province. Other countries represented include Brazil, where researchers examined correlations between Wikipedia page views and dengue cases; Saudi Arabia, where deep learning was applied to Arabic-language tweet classification; and the United Kingdom, Italy, Spain, Malaysia, India, Canada, and Nigeria, either as focal points or within multi-country datasets. As for diseases, the majority of studies centered on COVID-19, while several others addressed Influenza, Zika, Dengue, Ebola, and MERS, providing a comparative view of digital surveillance across multiple infectious disease contexts. See Table 1 below for full details on the included studies.

Table 1 Overview of Studies Included in the Review: Platforms, Diseases, Geographic Coverage, and Methodological Approaches (2015-2023)

Authors (Year)	Platform(s) Used	Disease Focus	Method(s) Used	Key Findings
Bernardo et al. (2013)	Google, Twitter, Yahoo	Influenza, foodborne illnesses	Scoping review, tool taxonomy	Most tools accurate but struggled with foodborne illness; Google-based tools dominated.
Fung et al. (2015)	Twitter, Weibo, Baidu	H7N9, Ebola, Breast Cancer, SARS	Event-based surveillance, sentiment tracking	Social media effective in communication, but subject to misinformation and data gaps.
Al-Garadi et al. (2016)	Twitter, Facebook, OSNs	H1N1, Zika, Dengue	Support Vector Machines (SVM), Logistic Regression, Topic Modeling	High correlation with CDC data; Twitter useful for early detection.
Aiello et al. (2020)	Google Trends, Twitter, FluNearYou	Flu, COVID-19	Search trends, participatory data	Hybrid models improve outbreak tracking; GFT overestimated flu in 2013.
Bisanzio et al. (2020)	Twitter (historical)	COVID-19	Geospatial tracking, correlation analysis	$\rho = 0.71$ correlation with case spread; historic mobility data predicted outbreaks.
El Azzaoui et al. (2021)	Twitter	COVID-19	NLP, Symptom extraction, Misinformation detection	Predicted outbreaks 7 days in advance with $r = 0.989$; identified misinformation.
Alsudias & Rayson (2021)	Twitter (Arabic)	COVID-19, Influenza	Multilabel classification, AraBERT, NER	F1 > 90% for classification; dialectal Arabic improves accuracy by 15%.
Wilson et al. (2021)	Twitter, Google Trends, Facebook	Zika, Ebola, COVID-19	Search query spikes, sentiment trends	Google/Twitter trends preceded outbreaks; misinformation spikes noted.
Kazijevs et al. (2022)	Twitter	COVID-19	LSTM, USE embeddings, Time series forecasting	$R^2 > 0.7$; tweets predicted cases up to 8 days in advance.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25jul887

Wang et al. (2023)	Twitter, Reddit, Wikipedia, Weibo, Baidu, TikTok	COVID-19, Influenza, Ebola, Zika	AI models (Deepluenza, MVEDL, CALI-Net), ensemble systems	Ensemble models highly accurate (up to 99%); misinformation filtering essential.
Tanner et al. (2023)	Twitter	COVID-19 (mental health context)	Tweet timeline extraction, cohort integration	Collected 700K tweets; integrated social data with longitudinal cohorts.

Each platform was examined separately prior to making inferences across platforms. The reason for such scrutiny was to analyze their distinct data structures, user behaviors, and surveillance capabilities thoroughly. Treating each platform as a unique digital environment allowed for a more accurate assessment of how different types of data, ranging from search patterns to user-generated content, can inform disease surveillance. This approach ensured that comparisons and conclusions were grounded in the specific affordances and limitations of each system rather than assuming uniformity across platforms.

➢ *Twitter*

Twitter, a microblogging platform where users share posts limited to 280 characters, has become the most widely studied social media site for disease surveillance due to its publicly accessible API, high-frequency user engagement, and rapid dissemination of health-related content. Its unique data structure, which includes real-time timestamps, hashtags, mentions, and occasionally geotagged posts, provides a granular and dynamic stream of public discourse. Researchers have used Twitter extensively to monitor symptom expression, detect misinformation, model outbreak trajectories, and capture shifts in public sentiment during health crises such as COVID-19, Zika, Influenza, and Ebola.

The types of data extracted from Twitter for epidemiological use have been diverse and detailed. Text content was mined for symptom-related keywords such as “dry cough,” “shortness of breath,” “fever,” “loss of taste,” “loss of smell,” and “chest pain.” These keywords were typically identified through dictionary-based symptom lexicons or machine learning–assisted keyword expansion. Studies also tracked disease-specific hashtags like #COVID19, #StayHome, and #flu to capture broader topic trends. Metadata, including timestamps and user geolocation, was incorporated when available, although only a small subset of tweets (~2%) were geotagged. Nevertheless, researchers used indirect geolocation techniques such as user profile inference and regional keyword clustering to augment spatial analyses.

Analytical approaches to Twitter data have become increasingly sophisticated. Early studies relied on bag-of-words models, which represent text by counting the frequency of individual words without considering grammar or word order, and simple frequency tracking, which involves monitoring how often specific keywords or hashtags appear over time. More recent work has utilized advanced Natural Language Processing (NLP) techniques such as tokenization, lemmatization, Named Entity Recognition (NER), and word embeddings. For predictive modeling, Kazijevs et al. employed Long Short-Term Memory (LSTM) neural networks in conjunction with the Universal Sentence Encoder (USE) to semantically embed tweet content into a temporal forecasting model (Kazijevs et al. 7). Through the use of LSTM, Kazijevs et al. achieved strong predictive performance, with city-specific COVID-19 case counts predicted up to eight days in advance and coefficient of determination scores (R^2) exceeding 0.7 in multiple urban regions. Similarly, El Azzaoui et al. processed over 10,000 English-language tweets using NLP to identify co-occurring symptomatic keywords (El Azzaoui et al. 4). Their NLP outbreak model, which included filtering steps for misinformation and noise, demonstrated a correlation of 0.989 between tweet-based symptom signals and confirmed COVID-19 case data.

Sentiment analysis played a key role in measuring emotional and behavioral responses to the pandemic. Tools like VADER (Valence Aware Dictionary and sEntiment Reasoner), TextBlob, and Bidirectional Encoder Representations from Transformers (BERT)-based models were used to classify tweets as positive, neutral, or negative, and in some studies, as indicative of anxiety, fear, anger, or trust. These sentiment scores were then aggregated temporally and geographically. Wang et al. found that surges in negative sentiment often preceded rises in reported cases or hospitalization, and that spikes in polarity scores were frequently associated with vaccine skepticism or public backlash to health mandates (Wang et al. 12). Alsudias and Rayson, in a study of Arabic-language tweets, further demonstrated that dialectal adaptation was essential for accurate classification (Alsudias and Rayson 13). Their use of the AraBERT model improved multi-label classification F1 scores by up to 15%, highlighting the importance of linguistic and cultural nuance in NLP models applied to social media.

Geospatial analysis has also been a key focus, though constrained by data sparsity. A study by Bisanzio et al. capitalized on a historical Twitter dataset from 2013 to 2015 to track the movement of users geotagged in China (Bisanzio et al. 6). Bisanzio et al.’s study analyzed the post-departure mobility of 4,669 users who had tweeted from various regions in China and then moved to other regions. When this data was compared with national COVID-19 case emergence in 2020, the researchers found a strong Spearman’s correlation of $\rho = 0.71$ between historic mobility patterns and confirmed outbreak regions. This suggested that even historical geotagged data could serve as a proxy for human mobility and disease importation risk in the absence of real-time spatial tracking.

However, despite its utility, Twitter's value for epidemiological modeling is tempered by the need to address misinformation and demographic bias. The platform

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25jul887

is overrepresented by younger, urban, and tech-savvy populations, which limits its representativeness to the general public. Misinformation also poses a substantial threat to signal validity. Studies such as those by Wilson et al. and Wang et al. documented how false claims, including those promoting unproven treatments, COVID-19 hoaxes, and anti-vaccine conspiracy theories, often surged in response to policy announcements and health agency warnings. These misinformation spikes introduced volatility into keyword and sentiment tracking, which, if not filtered, could distort outbreak forecasts. To combat this, El Azzaoui et al. incorporated misinformation detection pipelines into their NLP model, using external fact-checking databases and machine learning classifiers to flag and exclude false content (Wilson et al. 5; Wang et al. 10; El Azzaoui et al. 3). Others deployed bot detection algorithms and social network analysis to exclude tweets from suspected automated accounts.

Twitter has proven to be a highly responsive and informative platform for digital epidemiology. Its strengths lie in real-time accessibility, thematic richness, and scalability, while its limitations, chiefly demographic skew, low geotag penetration, and vulnerability to misinformation, necessitate cautious interpretation and robust methodological safeguards. When combined with advanced modeling techniques and ethical safeguards, Twitter data can serve as a powerful component of modern outbreak surveillance and public health response.

➢ *Facebook*

Facebook is a globally dominant social networking platform designed for long-form content sharing, group-based interaction, and threaded discussions. It supports various multimedia formats and offers both public and private posting structures, making it a rich source for understanding collective narratives and behavioral responses during public health crises. However, Facebook's data is considerably less accessible than Twitter's due to privacy restrictions and limited API functionality, which has resulted in fewer large-scale epidemiological studies utilizing the platform. Al-Garadi et al. noted that despite Facebook's vast user base and potential for capturing in-depth user behavior and sentiment, it has been underutilized in public health surveillance due to these access limitations (Al-Garadi et al. 5). They emphasized that the lack of open, real-time data streams has restricted their application in predictive modeling and real-time outbreak monitoring, even though it remains a promising source for studying health communication dynamics.

Despite these barriers, Facebook has played a critical role in tracking the diffusion of health-related misinformation and gauging vaccine hesitancy. Researchers analyzing public group posts and comment threads have identified thematic patterns of health behavior, misinformation spread, and anti-government sentiment. In particular, Wilson et al. demonstrated that Facebook misinformation during the 2014–2015 Ebola outbreak contributed to a measurable decrease in treatment-seeking behavior in West Africa (Wilson et al. 3). The study found that communities exposed to false claims about Western biomedicine exhibited reduced trust in health authorities. This sentiment translated into behavioral outcomes that hindered outbreak response.

Wang et al. expanded on this finding during the COVID-19 pandemic, observing that misinformation clusters on Facebook, especially within closed or private groups, served as echo chambers for anti-vaccine rhetoric and conspiracy theories (Wang et al. 9). Topics included claims about microchipping through vaccines, immune system "boosting" alternatives, and the denial of SARS-CoV-2 as a real pathogen. While these insights were derived through manual or semi-automated content analysis due to Facebook's data access constraints, they revealed how algorithmically curated content streams can reinforce pseudoscientific beliefs and potentially escalate public health risk.

Analytically, Facebook content was commonly examined using qualitative content analysis, topic modeling techniques such as Latent Dirichlet Allocation (LDA), and sentiment scoring algorithms to detect emotional tone and thematic clustering. While many studies focused on vaccine discourse, government trust, and public reactions to health policy, others identified more specific themes, such as concerns over vaccine safety, references to alternative medicine, and first-person narratives of illness. These thematic trends were often analyzed longitudinally to observe how narratives evolved in response to external events like outbreak surges or policy announcements. Although Facebook's privacy restrictions and limited API access prevented its widespread use in real-time symptom tracking or predictive modeling, functions more commonly seen on Twitter or Google Trends, it remained a valuable environment for understanding social contagion, behavioral risk, and evolving public sentiment. Some researchers also examined engagement metrics such as likes, shares, and comments to assess which types of content gained traction, particularly within private groups, and how misinformation or emotionally charged posts were amplified. This highlighted Facebook's potential not just as a site of misinformation, but as a nuanced lens into the psychosocial and communicative dimensions of public health behavior at scale.

➢ *Google Trends*

Google Trends is a publicly available web analytics tool that provides real-time and historical data on the popularity of search queries across regions and languages. As a proxy for public attention and information-seeking behavior, Google Trends has become one of the earliest and most frequently used digital tools in the surveillance of infectious diseases, notably influenza, COVID-19, Zika, and Ebola. Researchers typically analyze keyword volume data related to disease symptoms or treatments, using terms such as "fever", "cough", "flu", "symptoms", "loss of taste", and "COVID test." These terms are tracked longitudinally to identify search volume surges that might indicate community-level concern or early symptom reporting. Spikes were reported in relevant search queries that

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25jul887

consistently preceded reported increases in disease incidence by several days to weeks, with a typical lag range of 5–14 days (Aiello et al. 4; Al-Garadi et al. 8).

In recent years, Google Trends has been integrated with other surveillance tools to mitigate the spread of diseases. For instance, McGough et al. combined Google Trends data with Twitter and HealthMap to predict the spread of the Zika virus across six Latin American countries, Brazil, Colombia, El Salvador, Honduras, Mexico, and Venezuela, during the 2015–2016 outbreak (McGough et al. 5). The researchers used weekly Zika case counts from Pan American Health Organization (PAHO) as the ground-truth metric (data that is known to be true) and applied lagged cross-correlation to compare these with spikes in search volume, tweet frequency, and HealthMap alerts. Google Trends provided keyword data for terms such as “Zika virus,” “mosquito bites,” and “Zika symptoms,” while Twitter data captured public discourse and personal symptom mentions. HealthMap added an event-based layer, cataloging news alerts and official outbreak reports. These three data streams were aligned temporally using Autoregressive Integrated Moving Average (ARIMA) models to forecast case trajectories up to 3 weeks in advance. The study found that the multi-source model improved predictive accuracy (R^2 = 0.76) compared to models based on Google Trends or Twitter alone. Additionally, geospatial overlays were used to map query volumes by subnational regions, which were cross-validated against local case reports to identify hotspots. This layered approach demonstrated how combining search behavior with real-time social media communication and structured surveillance reports enhances both the sensitivity and specificity of digital disease forecasting.

Although Google Trends does not provide demographic identifiers or individual-level metadata, its strength lies in its ability to rapidly signal shifting public concerns and guide resource allocation for testing and public communication. When used as an early warning system in conjunction with traditional surveillance and other digital tools, it improves the timeliness and scope of monitoring.

➢ *Reddit*

Reddit is a community-oriented, forum-based platform composed of topic-specific subreddits where users engage in threaded, often long-form discussions. Known for its anonymity and deep user engagement, Reddit is an emerging but potent resource for digital epidemiology, particularly for qualitative and sentiment-based analysis.

Reddit content has been used to capture first-person narratives of illness, user-shared symptom logs, vaccine experiences, and public reactions to health policy. Data sources include post titles, body text, timestamps, comment threads, and upvotes/downvotes, which collectively provide insight into user engagement and credibility perception. Subreddits such as r/COVID19, r/Coronavirus, r/AskDocs, and r/NoNewNormal have served as rich environments for collecting health-relevant discourse.

Analytical methods applied to Reddit data include unsupervised topic modeling (e.g., LDA and Non-negative Matrix Factorization (NMF)), transformer-based sentiment classification (e.g., BERT, Robustly optimized BERT approach (RoBERTa)), and sequential thread analysis. Reddit data were integrated into ensemble models like SENTINEL and DEFENDER to analyze vaccine hesitancy, track misinformation narratives, and monitor public mood concerning evolving health guidelines (Wang et al. 13). They found that Reddit’s longer post format and voting structure made it ideal for surfacing nuanced discussions and identifying shifts in collective sentiment that might not emerge in shorter, more reactionary platforms like Twitter.

Studies also identified misinformation themes and behavioral insights within Reddit’s structure. For instance, antivaccine ideas often appeared in pseudoscientific content, referencing discredited studies or appealing to “natural immunity.” Users frequently cited each other’s personal experiences, adding a layer of narrative legitimacy that contributed to the viral nature of false claims.

Though Reddit lacks the real-time data velocity of Twitter or Google Trends, it offers significant advantages in the depth, quality, and context of user-generated content. Its anonymous environment further encourages frank disclosure of symptoms, risk behaviors, and treatment decisions.

➢ *Wikipedia*

Wikipedia is a collaboratively edited online encyclopedia that, although not a social media platform in the conventional sense, provides valuable insight into public information-seeking behavior through its page view analytics. These data reflect collective attention patterns and are used to infer awareness and concern around ongoing health events.

Researchers analyze the number of views per day for disease-specific articles such as “COVID-19 pandemic,” “Dengue fever,” or “Ebola virus.” These patterns are then compared with official case reports using correlation analysis and time-series alignment. Marques-Toledo et al. found that Wikipedia page views closely tracked dengue incidence in Brazil, with spikes in views corresponding with increases in reported cases. Such alignment indicates that Wikipedia usage can function as a public awareness indicator during outbreaks.

Wikipedia data are typically analyzed using simple temporal modeling and visualization tools. Its page view data are freely available and updated in near real-time, enabling researchers to observe surges in attention as news events break. The platform’s content also serves as a reference point for gauging public exposure to accurate versus outdated medical information, though Wikipedia’s open editing model requires cautious interpretation. Because Wikipedia lacks user-level metadata and does not support original content creation in the same way as social media platforms, its epidemiological utility is limited to awareness tracking rather than direct symptom or behavior monitoring.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25jul887

➢ *Baidu and Sina Weibo*

In China, where Western platforms such as Google, Twitter, and Facebook are restricted, Baidu and Sina Weibo serve as primary digital platforms for search and social interaction, respectively. Both platforms have been actively studied for their roles in real-time health surveillance during regional and global outbreaks.

Baidu, the country's dominant search engine, provides keyword frequency data through the Baidu Index. Studies have used this data to monitor population-level search behavior related to symptoms like "fever," "pneumonia," "diarrhea," and "loss of smell." During the early stages of the COVID-19 outbreak, search interest in respiratory symptoms began to rise approximately one week before confirmed case spikes in Hubei Province. Lag correlation analyses confirmed a predictive window of six to ten days, reinforcing the potential of the Baidu Index as an early-warning system (Qin et al. 10; Li et al. 9).

Sina Weibo, China's largest microblogging platform, functions similarly to Twitter, allowing users to post short updates, images, and hashtags. It has been widely used to monitor real-time expressions of health concerns, policy reactions, and requests for assistance. Researchers extracted posts containing symptom mentions, geolocated metadata (where available), and repost networks to assess public response and identify emerging hotspots. During the early Wuhan outbreak, spatial analysis of Weibo posts revealed clusters of concern and self-reported symptoms, which aligned with official outbreak zones later confirmed by the Chinese CDC.

Methodologically, studies used NLP to extract symptoms and sentiment from Weibo text, regression modeling to align search and post frequencies with case data, and Geographic Information System (GIS) mapping to visualize regional concerns. Despite concerns over censorship and government influence, both platforms provided early, localized signals that complemented official surveillance and helped predict outbreak trajectories.

Together, Baidu and Sina Weibo represent powerful tools for health monitoring in contexts in China, specifically where Western platforms are inaccessible. Their integration into epidemiological workflows has expanded the global applicability of digital disease surveillance, particularly in China, which is a non-English-speaking and high-density urban population.

IV. DISCUSSION

Social media and internet-based platforms have emerged as critical tools in supplementing traditional public health surveillance systems. Studies across Twitter, Google Trends, Reddit, and Wikipedia consistently demonstrate that spikes in public activity, such as symptom-related tweets, search queries, or page views, can precede official epidemiological data by several days to weeks. This lead time gives public health agencies a window for earlier intervention, risk communication, and policy adaptation, especially during fast-moving outbreaks like COVID-19, Zika, and Dengue.

However, while these platforms offer speed and scale, they cannot fully replace traditional contact tracing, which relies on verified, individual-level contact and diagnostic data. The key strength of social media lies in its ability to sense public reaction and early signals, particularly before lab-confirmed cases are widely reported.

Each platform analyzed in this study provides different types of data shaped by its architecture, user demographics, and accessibility, which, in turn, affects its epidemiological utility:

- Twitter is most effective for real-time signal detection due to its open-access API, rapid content flow, and rich metadata (e.g., hashtags, timestamps). It supports keyword tracking, symptom detection, sentiment analysis, and misinformation detection at scale. However, it suffers from low geotagging (~2%) and demographic skew toward urban, younger, and tech-savvy populations.

- Facebook offers access to longer-form posts and threaded discussions, which makes it highly valuable for qualitative insights into vaccine hesitancy, misinformation diffusion, and public sentiment. Its data are more representative of older and more diverse populations. However, Facebook is much more restrictive in terms of data access, especially in private or closed groups, limiting its utility for real-time surveillance or large-scale predictive modeling.

- Google Trends is one of the earliest tools used in digital epidemiology, particularly strong in tracking population-level information-seeking behavior. Search spikes for symptom-related terms (e.g., "cough," "COVID test") often precede official case reports. While it lacks individual-level metadata, its strength lies in scale, speed, and geographic coverage, especially when combined with clinical or social media data in multi-source models.

- Reddit excels at capturing in-depth community narratives, behavioral disclosures, and sentiment evolution. Its anonymity allows users to express personal illness experiences or opinions more openly, making it ideal for studying psychosocial dynamics and public trust. However, Reddit is less suitable for real-time outbreak detection due to slower content velocity and limited metadata, but it offers richer content context than platforms like Twitter.

- Sina Weibo functions similarly to Twitter but is culturally and linguistically tailored to China. It is especially effective in early localized outbreak detection in Chinese cities, often ahead of government-confirmed cases. When paired with Baidu search data, Weibo enables region-specific monitoring where Western platforms are inaccessible

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25jul887

- Wikipedia provides a unique lens into public attention and awareness. Page view data have been shown to correlate with case spikes in diseases like Dengue and COVID-19. Though not a social platform in the traditional sense, it serves as a barometer of public concern and can complement other platforms in hybrid surveillance models.

Together, these platforms represent complementary surveillance ecosystems. Real-time detection is strongest on Twitter and Baidu; sentiment and misinformation are best studied on Facebook and Reddit; population-level attention is most evident via Google Trends and Wikipedia, and region-specific signals can be captured through Weibo.

Predictive modeling has evolved from simple keyword frequency analysis to sophisticated AI systems incorporating Natural Language Processing (NLP), machine learning, and ensemble frameworks. Models using tweet embeddings (e.g., LSTM with USE) have predicted city-specific COVID-19 trends up to 8 days in advance. Dialect-sensitive models like AraBERT improved classification accuracy by 15%, demonstrating the importance of linguistic adaptability.

➢ *Improvements in Predictive Power have come from:*

- Multi-source integration (e.g., Twitter + Google Trends + HealthMap)
- Context-sensitive NLP (e.g., sarcasm and misinformation detection)
- Geographic Information System (GIS) based mapping of query clusters and user locations
- Real-time misinformation filtering using external fact-checking databases
- Studies have shown that combining data streams produced higher R^2 values and better regional hotspot identification than single-platform models, confirming the benefit of hybrid frameworks (McGough et al. 14).

Digital surveillance is most effective in epidemics with widespread media coverage and rapid transmission. Outbreaks like COVID-19, Influenza, and Ebola fit this pattern. However, the effectiveness of digital surveillance declines for chronic diseases or those with low public visibility. Reddit and Facebook have been more effective for tracking long-term mental health trends or chronic illness narratives, but they require distinct analytical strategies in order to produce optimal results.

➢ *There are also Demographic and Geographic Limitations:*

- Urban and younger users are overrepresented on Twitter and Reddit
- Older and rural populations are more present on Facebook, but harder to study due to privacy constraints
- Language and regional biases affect model accuracy (e.g., models trained on English tweets perform poorly on Arabic or Chinese posts without adaptation)
- These biases highlight the need for more inclusive datasets and culturally tuned models.

Misinformation poses a significant challenge for digital disease surveillance by distorting data signals and undermining public health responses. Across platforms like Twitter, Facebook, and Reddit, false narratives, such as those linking 5G cellular networks to COVID-19 or promoting vaccine conspiracies, frequently surged in response to policy announcements, skewing keyword frequencies and sentiment metrics. These surges can mimic outbreak-related signals, leading to misleading forecasts if not properly filtered.

To mitigate these effects, recent studies have adopted advanced machine-learning techniques for misinformation detection. Approaches now include fact-checking classifiers, bot detection systems, and network analysis to identify and remove manipulated or low-credibility content. Models like those developed by El Azzaoui et al. (2021) integrate external fact-checking databases to improve accuracy, while ensemble frameworks such as MVEDL, Deepluenza, and CALI-Net use multi-source data to separate genuine health signals from misinformation noise.

This review contributes a timely synthesis of research across multiple digital platforms and infectious diseases, highlighting the evolving role of social media in disease surveillance. One of its key strengths lies in its cross-platform approach, allowing for a more comprehensive comparison of each platform's surveillance capacity. Additionally, the inclusion of both pre-pandemic and pandemic-era studies offers a longitudinal view of how methodologies have evolved and how theoretical expectations held up under real-world conditions.

However, the study is not without limitations. Its scope was constrained by time, resulting in a sample size of 15 papers, which, while diverse, may not capture the full breadth of the rapidly expanding literature in this field. The focus was limited to English-language publications, potentially overlooking valuable regional studies in non-English contexts. Finally, the analysis, while detailed, was largely qualitative and would benefit from a future, larger-scale meta-analysis that quantitatively compares model performance across platforms and diseases. Despite these limitations, the findings of this review underscore the importance of integrating social media data into future-ready public health surveillance systems, particularly those that are multilingual, cross-platform, and sensitive to behavioral and cultural nuance.

This study's key strength lies in its cross-country, multi-platform scope and its inclusion of multiple disease contexts. However, its scale and timeframe were constrained, and future work should pursue a broader meta-analysis to quantitatively evaluate model performance across platforms and regions. With expanded research, these digital tools could become integral to early warning systems and global health preparedness.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25jul887

V. CONCLUSION

This review demonstrates that social media and internet-based platforms have become essential complements to traditional disease surveillance, offering earlier detection, broader population reach, and valuable insights into public sentiment and behavior. Across platforms like Twitter, Google Trends, Reddit, Wikipedia, and Sina Weibo, studies consistently showed that spikes in user activity, such as symptom-related tweets, search queries, or article views, preceded official case reporting by several days to weeks. This lead time creates a critical window for public health agencies to implement early interventions, target communication, and allocate resources. Future research should standardize data collection, strengthen AI and NLP methods, and expand access to underrepresented populations and languages. Collaborations between public health agencies and tech companies will be key to developing misinformation-resistant models and ethically grounded, privacy-conscious tools.

ACKNOWLEDGEMENTS

The author would like to express sincere gratitude to the researchers and authors whose studies formed the foundation of this review, particularly those who have pioneered the integration of social media data into epidemiological surveillance systems. Their innovative work has advanced our understanding of digital epidemiology and its potential to enhance public health responses.

REFERENCES

[1]. **Charles-Smith, Lauren E., et al.** "Using social media for actionable disease surveillance and outbreak management: A systematic literature review." *Online Journal of Public Health Informatics*, vol. 9, no. 1, 2017, doi:10.5210/ojphi.v9i1.7481.
[2]. **Schneider, Jan David, et al.** "Social Media Data for Omicron Detection from Unscripted Voice Samples." *Big Data and Cognitive Computing*, vol. 7, no. 2, 2023, p. 72. *MDPI*, https://www.mdpi.com/2504-2289/7/2/72. Accessed 17 May 2025.
[3]. **Tanner, Mark W., et al.** "Epicosm: Linking Twitter Data With Longitudinal Population Health Data." *International Journal of Epidemiology*, vol. 52, no. 3, 2023, pp. 952–961. *Oxford Academic*, https://academic.oup.com/ije/article/52/3/952/7058977. Accessed 17 May 2025.
[4]. **Aiello, Allison E., et al.** "Design and Methods of the Social Factors, Epigenomics and Lupus in African American Women (SELA) Study." *Frontiers in Public Health*, vol. 8, 2020, article 573866. *National Center for Biotechnology Information*, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7454796/. Accessed 17 May 2025.
[5]. **El Azzaoui, Amine, et al.** "Social Media Mining for COVID-19 Symptoms and Misinformation Detection." *JMIR Medical Informatics*, vol. 9, no. 9, 2021, article e27670. https://medinform.jmir.org/2021/9/e27670/. Accessed 17 May 2025.
[6]. **Alsudias, Lama, and Paul Rayson.** "COVID-19 and Influenza Classification Using Arabic Tweets: Deep Learning and Natural Language Processing Approach." *Frontiers in Computer Science*, vol. 4, 2022, article 949045. *National Center for Biotechnology Information*, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9490453/. Accessed 17 May 2025.
[7]. **Kazijevs, Vladislavs, et al.** "Predicting COVID-19 Trends with Social Media Embeddings." *IEEE Journal of Biomedical and Health Informatics*, vol. 27, no. 3, 2023, pp. 1398–1408. *PubMed*, https://pubmed.ncbi.nlm.nih.gov/36168466/. Accessed 17 May 2025.
[8]. **Bernardo, Theresa M., et al.** "Scoping Review on Search Queries and Social Media for Disease Surveillance: A Chronology of Innovation." *Journal of Medical Internet Research*, vol. 15, no. 7, 2013, article e147. *PubMed*, https://pubmed.ncbi.nlm.nih.gov/23896182/. Accessed 17 May 2025.
[9]. **Bisanzio, Donal, et al.** "Use of Twitter Data to Improve Zika Virus Disease Surveillance." *American Journal of Tropical Medicine and Hygiene*, vol. 102, no. 5, 2020, pp. 944–949. *PubMed*, https://pubmed.ncbi.nlm.nih.gov/31905322/. Accessed 17 May 2025.
[10]. **Fung, Isaac Chun-Hai, et al.** "Ebola and the Social Media." *Lancet*, vol. 2, no. 10, 2015, e566–e567. *National Center for Biotechnology Information*, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4542478/. Accessed 17 May 2025.
[11]. **Al-Garadi, Mohammed Ali, et al.** "Predicting Infectious Disease Occurrences Using Social Media." *Health Informatics Journal*, vol. 22, no. 1, 2016, pp. 1–18. *PubMed*, https://pubmed.ncbi.nlm.nih.gov/27224846/. Accessed 17 May 2025