International Journal of Public Health Science (IJPHS)
Vol. 14, No. 1, March 2025, pp. 479~492
ISSN: 2252-8806, DOI: 10.11591/ijphs.v14i1.24632

Utilising consumer reviews for passive surveillance of foodborne illnesses: insights and challenges from the Indian restaurant

Akash Prabhune[1,2], Vinay Sri Hari[2], Neeraj Kumar Sethiya[1], Mansi Gauniyal[3]
[1]Faculty of Pharmacy, School of Pharmaceutical and Populations Health Informatics, DIT University, Dehradun, India
[2]Institute of Health Management Research Bangalore, Bengaluru, India
[3]Foundation of Healthcare Technology Society, New Delhi, India

Article Info

Article history:

Received Feb 14, 2024
Revised May 10, 2024
Accepted Jun 5, 2024

Keywords:

Aspect based sentiment analysis
Consumer reviews analysis
Foodborne illness surveillance
Passive surveillance systems
Public health monitoring

ABSTRACT

This study explores the feasibility of leveraging consumer reviews for passive surveillance of foodborne illnesses, drawing parallels with pharmacovigilance systems. Utilizing a mixed-methods approach, including a literature review, qualitative interviews with key stakeholders, and quantitative analysis of 56,000 restaurant reviews using a bidirectional encoder representation from transformers (BERT-based) aspect-based sentiment analysis model, the study seeks to understand the potential of social media data in enhancing food safety regulations. Despite the promise shown by the aspect-based analysis in identifying hygiene-related concerns, challenges such as data reliability, language diversity, and the absence of verifiable evidence limit the direct application of consumer reviews for regulatory purposes. Key informants highlighted the need for a tailored surveillance system, considering the Indian restaurant industry's diversity and scale. The quantitative analysis revealed hygiene concerns in a significant portion of reviews, yet the specificity and verifiability of these complaints remain issues. The study suggests a cautious yet optimistic approach towards integrating social media data into foodborne illness surveillance, emphasizing the enhancement of traditional reporting mechanisms, targeted awareness campaigns, and industry collaboration. The findings advocate for a comprehensive strategy that includes developing a common platform for stakeholder communication and strengthening policy frameworks to support a collaborative surveillance system, particularly in developing contexts like India.

Corresponding Author:

Akash Prabhune
Faculty of Pharmacy, School of Pharmaceutical and Populations Health Informatics, DIT University Dehradun, India
Email: prabhunesky@gmail.com

1. INTRODUCTION

Sentiment analysis, especially aspect based sentiment analysis (ABSA), has proven effective in summarizing consumer sentiments in various domains, including food services. The research aims to extend these analyses to aid regulatory bodies by providing real-time insights into potential public health risks [1]–[5]. The sentiment analysis technique has been popularly used to analyze the sentiment behind the review and provide cumulative insights about the consumer sentiments of a product or establishment. ABSA further refines the sentiments into various aspects like quality, quantity, price, and hygiene. These natural language processing-based (NLP) tools have been tried and tested across various databases including Amazon Product Reviews, Google Maps Places Reviews, and Yelp reviews.

The ABSA-based Sentiment analysis has been able to effectively summarize consumer sentiment objectively. Eat at right place initiative (ERPI) is a rule aspect-based sentiment analysis platform based on the bidirectional encoder representations from transformers (BERT) model trained and validated to analyse the sentiments of Indian consumers reviewed for restaurants. The BERT model was developed using supervised learning and the F square value of the model was estimated to be 0.86, superior to the unsupervised learning-based approach. The ERPI platform is, an user application-based interface of the BERT model developed based on the fusion of agile scrum methodology and its mHealth adoption for user-centric mHealth application development [6], [7].

During the development of the ERPI platform, a critical question arose among the research team, developers, and advisors: could the platform be adapted to signal regulators about public health violations, such as improper food handling, pest infestations, or adverse health events following food consumption at specific restaurants? This inquiry is pivotal given the significant role of passive surveillance systems in managing public health risks. The literature underscores the necessity of such systems, particularly highlighted by agencies like the Food Standards and Safety Authority of India (FSSAI), which oversees restaurant and food business operations. The urgency for implementing passive surveillance stems from the highly fragmented and unorganized nature of India's food industry coupled with the FSSAI’s limited resources [8]–[12].

The current study presents a comprehensive approach to learning from a pharmacovigilance perspective wherein a passive surveillance system has been tested using data from consumer insights. Examples-web-recognizing adverse drug reaction (WEB RADR) by innovative medicines initiative [13], Vigi4Med Project [14], and Eudra Vigilance [15], are some of the well-established adverse drug event reporting programs utilising data from social media, and online forums. Insights indicates public health perspectives from the ERPI platform indicate the limitations of review data to support such a system and insights from the key informants’ interviews regarding developing passive surveillance systems using purely secondary data.

The objective of the current study is to: i) draw the parallels between passive surveillance for Pharmacovigilance and passive surveillance for food safety and define the parameters to build a passive surveillance system for Foodborne Illness; ii) determine the pragmatic perspectives on how data from social media could add value to regulators in terms of foodborne illness surveillance from key informants; and iii) based on the sentiment analysis insights determine the suitability of consumer reviews data to support a passive surveillance system for food safety.

The paper is structured to methodically explore the integration of pharmacovigilance techniques into foodborne illness surveillance through several key sections. Initially, a literature review compares the methodologies used in pharmacovigilance with those in foodborne illness surveillance to develop a framework for data requirements in passive surveillance systems. This is followed by qualitative interviews with key informants, which aim to align the proposed surveillance framework with existing food safety guidelines and policies, both within India and globally. Lastly, a quantitative analysis evaluates the data derived from the ERPI platform to determine if these data requirements can be effectively met in practical scenarios, thereby validating the framework’s applicability to real-world conditions.

2. METHOD

The current study was designed as a mixed methods study with the key methodologies: i) literature review for defining parameters of the passive surveillance system for foodborne illnesses; ii) qualitative analysis for understanding the stakeholder’s perspectives on setting up the proposed surveillance system; and iii) quantitative analysis for understanding the data feasibility of setting up the proposed surveillance system. The study was approved by the Institutional Review Board of DIT University Dehradun, vide letter number DITU/UREC/2022/04/16. The researchers have obtained written informed consent from all the participants in the study. The primary and secondary data collected for the study purpose were stored and analysed after removing any personal identifiers. All the participants were informed about their right to confidentiality and privacy. For the restaurants only geographic identifiers were preserved.

2.1. Literature review methodology

To establish parallels between passive surveillance in pharmacovigilance and foodborne illness surveillance studies and delineate the essential parameters for constructing a passive surveillance system for foodborne illness surveillance studies, a comprehensive literature review methodology was adopted. The methodology commenced with an extensive search across digital databases, including PubMed, Medline, Embase, Google Scholar, and IEEE, to identify pertinent articles. Additionally, grey literature sources, such as conference proceedings and pre-print servers like Research Square and Research Gate, were consulted to ensure a comprehensive review.

Two distinct search strategies were meticulously formulated: one tailored to pharmacovigilance and another to foodborne illness surveillance studies. Key terms for Pharmacovigilance encompassed concepts such as “surveillance,” “adverse drug reaction reporting,” and “pharmacovigilance,” while those for foodborne illness surveillance studies included terms like “surveillance,” “food safety,” “foodborne illness,” “food poisoning,” and “social media for food safety.” Employing appropriate Boolean operators in adherence to the PICO format, these terms were combined to facilitate targeted searches.

This extensive effort yielded a total of 3,073 articles (1,645 on pharmacovigilance studies and 1,428 on foodborne illness surveillance studies). Following primary screening, which focused on article titles and abstracts containing key search terms, 2,878 articles were excluded for not meeting the initial criteria. Subsequently, a second-level screening was conducted on 195 articles, guided by specific inclusion and exclusion criteria. Included articles involved passive surveillance for drug or foodborne illness surveillance studies, utilizing secondary data or a combination of secondary and primary data. Excluded were articles related to active surveillance or sentinel surveillance, as well as those concerning drugs or medical devices undergoing clinical trials for regulatory approval. Articles presenting conceptual frameworks without proof of concept were also omitted.

Furthermore, the definitions of active and passive surveillance, as defined by the World Health Organization, were integrated into the methodology. Active surveillance was characterized as the “collection of case study information as a continuous pre-organized process,” while passive surveillance was described as “a system that gathers information constantly, rather than actively pursuing targeted results” [16]. This rigorous approach culminated in the selection of 18 articles (14 related to pharmacovigilance studies and 4 to foodborne illness surveillance studies) for the literature review (outlined in Figure 1)

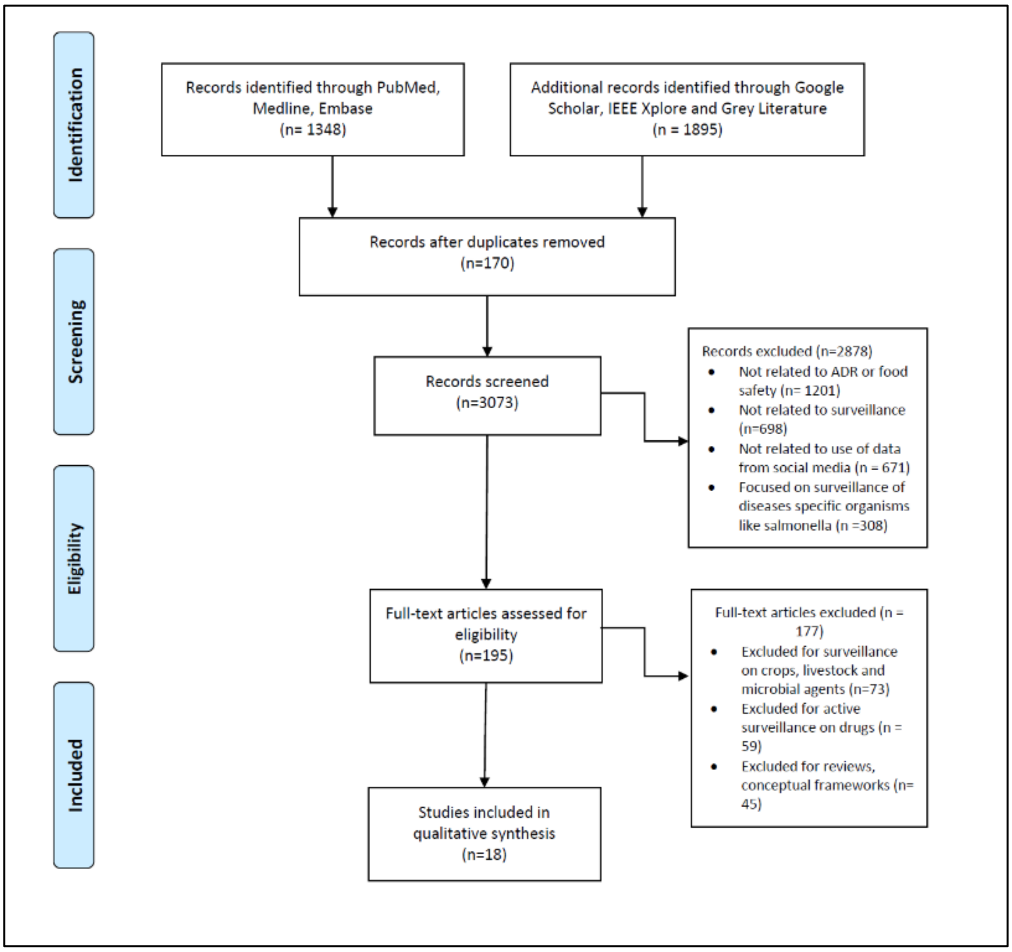

Figure 1. PRISMA flow diagram of search for passive surveillance methodologies on pharmacovigilance studies and foodborne illness surveillance studies

2.2. Qualitative research methodology

The qualitative research methodology was meticulously designed to specifically address the objective of discerning the pragmatic viewpoints on how social media data could augment regulators' efforts in ensuring foodborne illness surveillance studies, as articulated by key informants. A series of key informant interviews were conducted, with representation from a spectrum of stakeholders directly relevant to the Indian restaurant industry. These interviews aimed to delve into several critical aspects, beginning with an exploration of the imperative need for implementing a passive surveillance system tailored to the nuances of the Indian restaurant sector. A substantial focus was directed towards elucidating the multifaceted role of regulatory agencies in safeguarding food safety standards. Equally pivotal was the assessment of the feasibility and reliability of harnessing social media as an invaluable data source for passive surveillance within the restaurant industry. Furthermore, the interviews sought insights into the potential applications and advantages of leveraging outputs generated by a passive surveillance system, with benefits both for regulators and consumers at large. The sampling technique employed the snowball method, with an initial index respondent identified within each stakeholder group through professional networks. These index respondents, in turn, were requested to refer additional participants from their respective stakeholder circles, thereby diversifying the pool of interviewees. Data collection involved the use of a semi-structured questionnaire, ensuring a systematic approach during the interviews. Prior to the interviews, participants granted their consent, and audio recordings were employed to preserve the accuracy of the conversations. Subsequently, the recorded data underwent transcription and translation into English for subsequent thematic analysis.

In total, 12 interviews were conducted, providing representation from a range of stakeholder categories, including consumer advocacy groups, union representatives of restaurant owners, and officials from the FSSAI. Additionally, insights were gathered from industry consultants specializing in restaurant and food business branding, marketing, and strategic positioning. This methodological reconfiguration underscores a deliberate and focused effort to garner practical insights from key informants, shedding light on the potential of social media data to bolster food safety regulation within the context of the Indian restaurant industry. We used the sampling criteria developed by Prades *et al.* [17], wherein 12 in-depth interviews were conducted amongst diverse group of stakeholders focusing on profile diversity with and objective of defining how the social media could be used in a better way to communicate food risks and benefits. We aimed to achieve similar number of 12 interviews along with data saturation.

2.3. Quantitative research methodology

A total of 56,408 reviews from 76 restaurants spread across four Indian cities of Bengaluru, Pune, Dehradun, and Bhubaneswar were acquired using Web Scarper API. We didn't undertake formal sample size estimation for quantitative analysis since our model was already tested and validated. Most studies in literature usually foucssed on a single city or geogpraphy. Considering the vastnees and diversity of India, we decided to represent at least one city from each direction. This four India cities Bengalur (South), Pune (West), Dehradun (North) and Bhubaneswar (East) were selected using lottery method from the list of Tier 1 and Tier 2 cties classification used by Reserve Bank of India. We intended to collect data form 150 to 180 restaurants aliging to paper published by Lee [18]. However, considering the cost incurred on web-scrapping and budget constraints, we decided to cover half of the proposed number by Lee [18]. Thus, we targeted to scrape data from 20 restaurants per city (total 80 across India) selected and following selection criteria was devised and employeed.

The selection criteria for a restaurant were: i) the restaurant should have a minimum of 1,000 reviews from customers over 12 months (October 2022 to October 2023), and ii) the restaurant should have a dine-in option and only delivery or cloud kitchens were excluded. The restaurants were selected as follows.

- Two administrative wards were selected from each city using a lottery method.
- A list of the top 5 best-rated restaurants and top 5 worst-rated restaurants from each ward was acquired from the Google Maps Local Guides portal.
- The selected restaurants were screened for eligibility, a total of 10 restaurants were selected per ward and 20 restaurants per city with a total of 80 restaurants.

Four (4) worst-rated restaurants from Bhubaneshwar were excluded as they didn't record 1,000 reviews between October 2022 to October 2023. Thus, a total of 76 restaurants were included in the study. All the 1,000 reviews were web scrapped from each restaurant. For each review, the total word count including spaces, and punctuation marks should be more than 10 words for inclusion in the study. A total of 76,000 web reviews were scrapped from Google Maps, and 19,592 were excluded from analysis as the total word count was less than 10 characters. These reviews included one-word reviews like good, best, bad, thumbs up or thumbs down smiley.

The data from 56,000 reviews was analysed through the machine learning-based aspect-based sentiment analysis (ASBA) model. The BERT model is a state-of-the-art natural language processing (NLP) model that utilizes deep learning techniques. It is pre-trained on large amounts of text data and can be fine-tuned for specific NLP tasks such as sentiment analysis. There are commercially available open-source BERT models like scala consultant's ASBA [19], allowing Git Cloning for commercial and non-commercial use. The rule-based aspect assignment approach allows for targeted analysis of sentiment towards different aspects, providing valuable insights into customer opinions and feedback. By applying rule-based aspect determination, we can systematically identify and analyse sentiment patterns associated with specific aspects, enabling more granular and focused sentiment analysis.

The development methodology employed web scraping and an ABSA model to analyse 3,596 restaurant reviews from Bengaluru, with additional validation from Dehradun, Pune, and Bhubaneshwar. The ABSA model, with aspects like "service, food, price, ambiance," was tested on 500 reviews, providing baseline performance metrics. A rule-based system was developed using 3,096 labelled reviews, covering aspects like "ambiance, hygiene, service, time, food, variety, price, and accessibility." The adapted model demonstrated improved performance, evaluated through precision, recall, specificity, and sensitivity, showcasing its efficacy in handling diverse aspects of Indian consumer sentiment across multiple cities. Table 1 presents the details of the performance of the ABSA model. The details of the model are published in an IEEE conference [7].

Table 1. Details on the performance of the model for indian consumer data

Measure	Precision	Recall	F square	Accuracy (%)
Price	0.923	0.901	0.913	91
Food	0.915	0.905	0.910	86
Ambience	0.957	0.985	0.971	97
Staff	0.842	0.988	0.914	96
Hygiene	0.790	0.875	0.636	93
Variety	0.833	0.789	0.811	77
Accessibility	0.845	0.931	0.869	92

The crucial focus of this quantitative analysis is the significance of hygiene as a key attribute in a public health surveillance system. The FSSAI defines the hygiene aspect in terms of overall restaurant maintenance, staff personal hygiene, and pest control. We identified 32 keywords related to hygiene to facilitate a comprehensive analysis.

List of 32 keywords *""hairs, nails, stones, worms, insects, hygiene, washrooms, stinks, dirty, clean, unhygienic, bathroom, stomach upset, vomited, leakage, air conditioner, sanitized, super clean, cockroaches, felt ill, puking, trash, pan stains, smelly, plates, cutlery, smells bad, smells fresh, flies, mosquitoes, pest control".*

Analysis objectives included.
- Frequency analysis: Assess the frequency of the Hygiene aspect identified by the ABSA Model per 100 reviews, recognizing that reviews may encompass multiple aspects.
- Surveillance loop components: Evaluate if classified data under the Hygiene aspect includes essential components to initiate the surveillance loop.
- Regulatory compliance test: Test if the available data fulfils regulatory agency requirements for lodging food operator-related complaints, specifically focusing on three key data points: i) place details, ii) concern specifics, iii) evidence (geotagged photograph, laboratory report, or medical practitioner report linking health events with consumed food).

Illustrative example: A customer review for a restaurant was classified by the ABSA algorithm with negative sentiment and three aspects: food (-0.99), hygiene (-0.98), and ambiance (-0.99).

> "*We went with family yesterday for dinner! Such a bad decision. The place was a dungeon, which invited us with the foul smell of the dirty washroom! The food was horrible! The quality of the barbeque was pathetic - charred when they brought …"*

The review highlighted concerns about horrible food (food aspect), a bad decision (ambiance aspect), and foul smells and a dirty washroom (hygiene aspect). Although place details were available from Google Maps metadata and the concern was identified as poor hygiene, the essential evidence, such as a geotagged photograph, laboratory report, or medical practitioner report, was missing.

The analysis was conducted for each review, presenting findings in terms of both the number and frequency of reviews meeting all three data requirements. This structured approach ensures a comprehensive examination of the suitability of reviews data for effective public health surveillance and regulatory compliance. The final triage was undertaken to understand from the literature, qualitative interviews and available data to make evidence backed appeal on investing resources for setting up passive surveillance systems for Foodborne Illness Surveillance Studies using consumer reviews on social media.

3. RESULTS AND DISCUSSION

3.1. Findings of literature review

The literature review focussed on understanding data source, research aim, primary approaches for analysis of the data, evaluation methodology. The analysis of literature review is presented in Table 2 (see in Appendix) [18], [20]–[36]. The literature on passive surveillance systems for Pharmacovigilance Studies have standardised the framework with the processing being; i) data collection, ii) data pre-processing, iii) data processing, and iv) analysis. The same framework is also adopted by papers on foodborne illness surveillance studies. There are some similarities and differences in the process. Both foodborne illness surveillance studies and pharmacovigilance studies exhibit significant overlaps in their methodologies and objectives, emphasizing the critical role of online data in public health monitoring. Leveraging data from various online sources including social media platforms, online health forums, and healthcare communities, both types of studies aim to contribute to public health by identifying potential health risks. They employ a diverse array of data analysis techniques such as sentiment analysis, NLP, machine learning algorithms, and statistical methods to extract valuable insights from the collected data. Moreover, both types of studies rely on annotations to enhance the reliability and accuracy of their analyses, often employing evaluation methodologies involving quantitative comparisons against pre-validated models to assess the performance of their approaches. The utilization of advanced models and techniques like bidirectional encoder representations from transformers (BERT model), Bayesian models, support vector machines (SVM), and proportional reporting ratio underscores the sophistication of their analyses. Ultimately, both foodborne illness surveillance and pharmacovigilance studies share a common emphasis on signal detection, whether it pertains to adverse drug reactions or potential foodborne outbreaks, highlighting their shared commitment to safeguarding public health.

In addition to the similarities outlined, there are distinct differences between pharmacovigilance studies and foodborne illness surveillance studies that underscore their unique approaches and objectives. Pharmacovigilance studies typically rely on existing literature and databases, such as FDA data on known drug side effects, to validate the causation between a drug and an adverse event. This reliance on established references helps confirm the likelihood of an adverse drug reaction. In contrast, foodborne illness surveillance studies face a significantly broader range of potential combinations, given the multitude of food items, pathogens, and environmental factors involved. Moreover, the aims of these studies differ: pharmacovigilance studies aim to generate signals related to drug safety, focusing on detecting and analysing adverse drug reactions, while foodborne illness surveillance studies are geared towards projecting the next possible outbreak, often in terms of geographic regions or specific food items. These distinctions highlight the specialized focus and methodologies employed in each field, tailored to address the unique challenges and objectives inherent to pharmacovigilance and foodborne illness surveillance.

Based on the literature a seven pointers framework is proposed for foodborne illness surveillance

- Data sources: Social media platforms (e.g., Twitter, Yelp) and online communities are primary sources for collecting data related to foodborne illness.
- Research aim: The primary aim is to identify and project possible foodborne outbreaks, focusing on monitoring and detecting potential health risks associated with food consumption.
- Analytical techniques: Analytical techniques used include sentiment analysis, NLP, machine learning-based text mining, and statistical methods. Models such BERT model, are applied to analyse the collected data.
- Annotations and validation: Annotations are often used to enhance the reliability and accuracy of data analysis. Evaluation methodologies involve mixed methods, including model validation using measures like F measures and expert opinion validation.
- Evaluation methodology: Quantitative comparisons are conducted within models of the same class using measures such as F measures to assess performance. Descriptive analysis is employed to characterize the data explored from social media and other sources, often utilizing pre-existing models for validation.
- Manual processes: In some cases, manual processes are involved in selecting specific tweets or data points for further inspection and reporting to regulatory bodies.
- Outcome: The passive surveillance system aims to provide insights into foodborne illness trends, identify potential outbreaks, and assess the risk associated with specific food products.

This framework was shared with experts during qualitative interviews to understand their perspectives on setting up proposed foodborne illness surveillance system.

3.2. Results from the qualitative interviews with Key informants

The key informants, representing a spectrum of stakeholders, provided valuable insights into various critical aspects related to this endeavour. Firstly, there was a consensus among the stakeholders regarding the imperative need for implementing a tailored passive surveillance system for the Indian restaurant sector. As articulated by one key informant from a consumer advocacy group, "Given the diversity and scale of the Indian restaurant industry, there is a pressing need for a surveillance system that can effectively monitor and mitigate food safety risks."

Secondly, the role of regulatory agencies, particularly the FSSAI, was highlighted as pivotal in safeguarding food safety standards. An official from FSSAI emphasized, "Regulatory agencies play a crucial role in setting and enforcing food safety standards. However, with the evolving landscape of the restaurant industry, there is a need for innovative approaches to complement traditional regulatory mechanisms." However, despite recognizing the potential benefits of leveraging social media data for passive surveillance, key informants also highlighted several challenges and limitations. One industry consultant specializing in food business branding noted, "While social media offers a vast repository of data, its reliability and accuracy for foodborne illness surveillance remain questionable. The abundance of unverified information and the lack of standardized reporting mechanisms pose significant challenges."

Furthermore, concerns were raised regarding the feasibility of establishing a comprehensive passive surveillance system within the Indian restaurant industry. A union representative of restaurant owners commented, "Implementing a passive surveillance system requires significant resources and infrastructure, which may not be feasible for smaller establishments. Moreover, there are privacy and data protection concern that need to be addressed."

Thematic analysis undertaken highlights as follows.

Data challenges:

- Limited reach: Not everyone with foodborne illness uses social media, potentially leading to underreporting and inaccurate data. "*This system might miss a significant portion of cases, especially from individuals who are less tech-savvy or don't use social media regularly." (Representative from consumer advocacy group*).
- Language barriers: With English not the primary language for many Indians, accurately capturing and analysing sentiment and relevant information could be challenging *"Analysing data in multiple languages requires specialized tools and expertise, which might not be readily available or cost-effective." (Consultant for Industry)*
- Verification difficulty: Distinguishing genuine complaints from fake reviews, malicious intent, or cultural nuances in expression could be difficult, leading to false positives *"Separating genuine concerns from negativity unrelated to foodborne illness requires significant manual effort and verification, which could be resource intensive." (Official from FSSAI)*

Privacy concerns:

- Ethical implications: Collecting and analyzing personal data from social media raises privacy concerns and requires careful consideration of ethical guidelines and user consent *"Implementing such a system requires robust data privacy measures and transparency to gain public trust." (Consultant for Industry)*
- Cultural sensitivities: Cultural interpretations of illness and online expression might differ, leading to misinterpretations and potential biases in the system "*The system should be mindful of cultural sensitivities and avoid drawing conclusions based on solely online expressions*." (*Consultant for Industry)*

Regulatory and practical hurdles:

- Resource constraints: Regulatory agencies might lack the resources or expertise to manage and analyse large volumes of social media data effectively *"Scaling up such a system for national monitoring might require significant infrastructure and personnel investment." (Official from FSSAI)*
- Integration challenges: Integrating social media data with existing food safety surveillance systems might require complex technical solutions and collaboration between multiple stakeholders *"Smooth integration with existing reporting systems requires clear communication and data standardization across entities." (Official from FSSAI)*
- Limited actionability: Even if outbreaks are identified, tracing them back to specific restaurants and taking effective action could be challenging due to anonymity and limited online information *("The system might not be able to pinpoint specific establishments responsible for outbreaks, hindering quick intervention." (Representative from consumer advocacy group)*

Alternative approaches suggested by key informants include: i) strengthening traditional reporting systems: Focusing on improving existing reporting mechanisms from restaurants and healthcare providers might be more efficient and reliable; ii) targeted awareness campaigns: Educating consumers on food safety and encouraging them to report suspected illness through established channels could be beneficial; and iii)

Collaboration with industry stakeholders: Partnering with restaurant associations and food delivery platforms to implement self-regulation and data sharing initiatives could provide valuable insights.

3.3. Quantitative data analysis results were as follows

The data from 56,000 reviews was analysed using validated BERT model. The frequency of aspects reported per 100 reviews is presented in Table 3. Since multiple aspects could be identified per review the total frequencies would be higher than 100, thus total is presented in the table. The aspect pertaining to food borne illness surveillance was attribute to “Hygiene” which had second least frequency. Indicating the customer are lesser concerned about the Hygiene aspect when sharing reviews. A total of 12,880 (23%) Hygiene related issues were noted across the 56,000 datasets.

Table 3. Frequency of aspect identified by BERT Model per 100 reviews from total of 56,000 reviews

Aspects	Positive sentiment frequency	Neutral sentiment frequency	Negative sentiment frequency
Food quality	90	25	57
Staff	85	4	41
Ambience	72	5	31
Variety	45	1	16
Price	33	2	27
Hygiene	13	1	9
Accessibility	5	1	11

The quantitative analysis aimed to assess whether current data meets the requirements of a surveillance framework, particularly in linking food items to reported health issues, such as illnesses, pests, or other hygiene-related concerns. This analysis involved examining 12,880 restaurant reviews that mentioned hygiene. Out of these, 3,670 reviews (28%) did not mention any specific food item and primarily discussed the general cleanliness of the establishment or the personal hygiene of the staff. Another 2,852 reviews (22%) focused on the hygiene of the washrooms. There were 2,424 reviews (19%) that reported the presence of insects or pests in the meal, but these did not specify any particular food item. Finally, 3,934 reviews (30%) explicitly mentioned specific food items in the context of hygiene concerns. In the quantitative analysis, the final point of examination was whether the dataset complies with the criteria set by regulatory agencies for filing complaints related to food service providers. This evaluation concentrated on three essential elements: i) details about the establishment, ii) specifics of the concern, and iii) supporting evidence, which could include geotagged photos, lab reports, or medical documentation linking adverse health outcomes to the consumed food.

For the entire dataset of 12,880 reviews, location information was readily accessible through Google Maps metadata, ensuring the first criterion was fully met. The specificity of concerns was also adequately addressed in all reviews, with an average length of 32 words per review, ranging from a minimum of 12 to a maximum of 231 words, indicating clear communication of issues encountered. Regarding the third criterion, evidence, only 1,562 reviews (approximately 12%) included photographs that depicted issues such as unsanitary conditions in washrooms or dining areas, or visible pests. However, there were no laboratory or medical reports provided in any of the reviews, which signifies a gap in the type of evidence available to support health-related complaints directly linked to food consumption.

Restaurant-wise frequency analysis was not undertaken to preserve the anonymity of restaurants. Analyzing data from social media and online community platforms has revealed a wealth of information regarding foodborne illnesses, with 12,880 hygiene-related reviews serving as a significant dataset for passive surveillance within the Indian restaurant industry. Despite the fact that 30% of these reviews specify particular food items, highlighting potential sources of illness, a considerable portion of the data lacks this specificity, offering insights into general hygiene practices rather than pinpointing exact sources. Utilizing BERT models and other NLP techniques has proven effective for sentiment analysis, though the challenge remains in linking health risks to specific foods or practices due to the absence of corroborative medical reports. Stakeholders recognize the utility of social media data for early detection of food safety risks but are cautious about its reliability and the practical challenges of integrating it with traditional regulatory frameworks, especially given India's linguistic diversity and the ethical concerns surrounding personal data.

The feasibility of employing social media for passive surveillance is acknowledged, yet the current infrastructure's limitations and the difficulty in obtaining verifiable evidence pose significant barriers. While the identification of hygiene-related issues offers promise, the ability to act on these findings is constrained by the anonymity of social media and the scarcity of detailed reports, underscoring the need for further investment and development to enhance the efficacy of such surveillance systems.

3.4. Discussion

The analysis underscores the vast potential of leveraging restaurant reviews as a data source for passive surveillance of foodborne illnesses. With the ubiquity of online platforms and the tendency of patrons to share dining experiences, restaurant reviews offer a rich repository of real-time, consumer-generated data. This approach aligns with the growing trend of utilizing digital epidemiology tools for public health surveillance, as seen in previous studies [37], [38]. A paper by Nsoesie *et al.* [39] have presented working model on signal detection and prediction of possible outbreaks geo-spatially using data from social media and customer reviews however the work needed ground truthing support and collaboration with regulatory agencies to act on signal generated. In context of the current study wherein the Indian settings are examined, the signal generation is not a priority for regulatory agencies and even various other stakeholders. The primary concern is the reliability and accuracy of the data, given the subjective nature of reviews and the potential for misinformation. Similar concerns were raised by the participants of qualitative interviews published by Velsen *et al.* [40] when asked about the used of social media data for outbreak investigations due to food-borne illnesses.

Additionally, the lack of verifiable evidence, such as laboratory reports or medical documentation, in the reviews limits the ability to establish causation between the consumption of specific food items and reported illnesses. A systematic review by Overbey *et al.* [41] reviewed papers focussing on social media for risk communication for food safety indicated social media could compliment existing systems and used to create a communication channel between consumers, restaurants and regulatory agencies.

Another aspect covered by this paper is the frequency of user reporting on the Hygiene aspects, out of the total 56,000 reivews analysed only 12,880 (23%) were foucssed on Hygiene aspect which is higher the 15.7% reported by Siering [42] from data of 2,875 restaurant reviews reviews of Las Vegas restaurants yelp reviews. Earlier study published earlier by the authors undertook survey amongst 280 consumers from a southern India city, hygiene aspects like cleanliness of premises, pest control, absence of external objects in food were ranked as third data point consumers would use to make decision about visiting a restaurant. First data point was quality of food and second was staff behaviour. Indicating the consumer mindset does have concerns about food safety however it can be canvased if food quality and staff behaviour is better.

The aspect-based machine learning approach, particularly models like BERT and SVM, has shown promise in extracting relevant information from unstructured text data. This technique allows for the categorization of reviews based on specific aspects such as food quality, hygiene, and staff behaviour, which are pertinent to foodborne illness surveillance. The effectiveness of such models in sentiment analysis and feature extraction has been well-documented evident over three decades with latest work of Devlin *et al.* [43] presenting accuracy of 86 to 92%. The parallels drawn between the proposed system and existing surveillance frameworks in pharmacovigilance and foodborne illness monitoring indicate need for clear framework to set up surveillance systems for food safety. While there are similarities in methodologies, the unique challenges posed by the diverse and complex nature of foodborne illnesses necessitate a tailored approach to surveillance in the restaurant industry.

Based on the findings, the manuscript advocates for a cautious yet optimistic approach to integrating restaurant reviews into foodborne illness surveillance systems. It suggests a multi-faceted strategy involving the enhancement of traditional reporting mechanisms, targeted awareness campaigns, and collaboration with industry stakeholders to ensure a comprehensive and effective surveillance system. A step wise approach could be adopted by developing countries like India with building a platform as common floor for communication between various stakeholders utilizing social media followed by strengthening and cohesive development of policies and frameworks to support a collaborative passive surveillance system for foodborne illness.

4. CONCLUSION

This study demonstrates the potential of leveraging digital platforms, especially social media and online review systems, for passive surveillance of foodborne illnesses. By examining methodologies from pharmacovigilance, where advanced tools like sentiment analysis, NLP, and machine learning are used for adverse drug reaction detection, valuable insights can be applied to food safety monitoring. While pharmacovigilance benefits from curated databases, foodborne illness surveillance faces challenges due to the diverse range of food items, pathogens, and environmental interactions. Digital tools can complement traditional systems by offering real-time, large-scale data for early signal detection and trend analysis. Key insights reveal that employing aspect-based machine learning models like BERT and SVM can effectively categorize consumer reviews according to food quality, hygiene, and staff behaviour, critical parameters for foodborne illness surveillance. However, challenges remain, including data reliability and establishing causation due to the subjective nature of consumer-generated data. Despite these obstacles, the study suggests a cautious yet optimistic approach, advocating for a multi-faceted strategy that includes enhancing traditional reporting systems, targeted awareness campaigns, and industry collaboration. Future efforts should focus on developing a communication platform for stakeholders, policies, and frameworks to support a collaborative passive surveillance system, particularly in developing countries like India. Building such a platform will allow for improved communication between regulatory agencies, consumers, and restaurant stakeholders, leading to more effective foodborne illness surveillance.

ACKNOWLEDGEMENTS

Author thanks Director IIHMR Bangalore for providing support for the project under internal grant FY 2023-24, vide letter IIHMR-B/DIR/2023-2024/190.

APPENDIX

Table 2. A summary table showing approaches and evaluation methodologies for surveillance approaches based on secondary data for drug saftey and foodborne illness surveillance studies

Serial	Author details	Year of publication	Type of surveillance drug/food	Data source	Research aim	Primary approaches for analysis of the data	Broad model	Annotations used (YES/NO)	Evaluation methodology
1	[20]	2012	Drug	Online healthcare community (MedHelp)	Signal detection of adverse drug reactions	Association mining using leverage, lift and proportional reporting ratio	Sentiment analysis	YES	Quantitative: Compared performances against pre-validated model of same class using F measures
2	[21]	2020	Drug	Online healthcare community (Daily Health)	Enhancing signal detection of adverse drug reactions	Combination of syntactic dependencies and ADR lexicon for feature extraction and deep linguistic processing model for data analysis	Sentiment analysis	YES	Quantitative: Compared performances against pre-validated model of same class using F measures
3	[22]	2022	Drug	Social media (X formerly twitter)	Merge and process data from multiple sources for processing	NLP	Only data processing demonstrated	NO	Visualisation of data processing
4	[23]	2022	Drug	Social media (Facebook and X formerly twitter)	create more reliable ADE extraction models	Bidirectional Encoder Representations from Transformers (BERT model) with negation and speculation	Sentiment Analysis	YES	Quantitative: Compared performances against pre-validated model of same class using F measures
5	[24]	2017	Drug	Social media and online health forums	signal detection of adverse drug reactions	Bayesian model for the authenticity and credibility aware signalling of potential ADEs from social media (AC-SPASM)	Bayesian model	YES	Quantitative: Compared performances against pre-validated model of same class using F measures
6	[25]	2016	Drug	Social media (Facebook and X formerly twitter)	signal detection of adverse drug reactions	Proportional reporting ratio	Sentiment analysis	YES	Quantitative: Compared performances against pre-validated model of same class using F measures
7	[26]	2017	Drug	Online health forums	signal detection of adverse drug reactions	Named Entity Recognition (NER)	Sentiment Analysis	YES	Quantitative: Compared performances against pre-validated model of same class using F measures
8	[27]	2021	Drug	Social media and online health forums	enhancing signal detection of adverse drug reactions	Named Entity Recognition (NER)	Sentiment analysis	YES	Quantitative: Compared performances against pre-validated model of same class using F measures

Table 2. A summary table showing approaches and evaluation methodologies for surveillance approaches based on secondary data for drug saftey and foodborne illness surveillance studies (*continue*)

Serial	Author details	Year of publication	Type of surveillance drug/food	Data source	Research aim	Primary approaches for analysis of the data	Broad model	Annotations used (YES/NO)	Evaluation methodology
9	[28]	2017	Drug	Social media (X formerly twitter)	enhancing signal detection of adverse drug reactions	Named Entity Recognition (NER)	Sentiment Analysis	YES	Quantitative: Compared performances against pre-validated model of same class using F measures
10	[29]	2016	Drug	Online health forums (AskaPatient.com)	signal detection of adverse drug reactions	Granger analyses	Bayesian model	YES	Quantitative: Compared performances against pre-validated model of same class using F measures
11	[30]	2023	Drug	Online health forums (AskaPatient.com, webmd.com, and iodine.com)	extraction of ADR from social media data	NLP	Sentiment Analysis	NO	Descriptive analysis on the characteristics of data explored from social media analysed on pre-existing model
12	[31]	2017	Drug	Social media and online health forums	Extraction of ADR from social media data	Semi supervise learning based ADE extraction	Sentiment analysis	YES	Quantitative: Compared performances against pre-validated model of same class using F measures
13	[32]	2017	Drug	Social media (X formerly twitter)	Extraction of ADR from social media data	Big data analysis and two-class support vector machine (SVM)	Sentiment analysis	YES	Quantitative: Compared performances against pre-validated model of same class using F measures
14	[33]	2018	Drug	Social media (X formerly twitter)	Signal detection of adverse drug reactions	Proportional reporting ratio	Sentiment analysis	YES	Quantitative: Compared performances against pre-validated model of same class using F measures
15	[34]	2023	Food	Social media (X formerly twitter)	Building a data pipeline for projecting possible food outbreak	Bidirectional encoder representations from transformers (BERT model	Sentiment analysis	NO	Descriptive analysis on the characteristics of data explored from social media analysed on pre-existing model
16	[18]	2023	Food	Social media (Yelp)	building a data pipeline for projecting possible food outbreak	Random Forest, Decision Tree, Support Vector	Sentiment analysis	YES	Quantitative: Compared performances within the models of same class using F measures
17	[35]	2018	Food	Social media (X formerly twitter)	identify and respond to food poisoning tweets on Twitter and increased food poisoning reporting by customers	Manual	NA	NA	Descriptive process of selecting specific tweets and promoting reporting incidents to regulatory bodies for further inspection
18	[36]	2020	Food	Amazon Customer reviews	Profile food items and assess risk, building a post market decision support system to identify hazardous food products	Machine learning based text minning	NLP	YES	Mixed methods with model validation using F measures and expert opinion to validate the final findings

REFERENCES

[1] A. Parikh, C. Behnke, M. Vorvoreanu, B. Almanza, and D. Nelson, “Motives for reading and articulating user-generated restaurant reviews on Yelp.com,” *Journal of Hospitality and Tourism Technology*, vol. 5, no. 2, pp. 160–176, Aug. 2014, doi: 10.1108/JHTT-04-2013-0011.

[2] A. Falaki, “Sentiment analysis by fine-tuning BERT [feat. Huggingface’s Trainer class],” Medium. Accessed: Jul. 07, 2023. [Online]. Available: https://nlpiation.medium.com/sentiment-analysis-by-fine-tuning-bert-feat-huggingfaces-trainer-class-97c5635035f7.

[3] D. S. Batista, “Aspect-based sentiment analysis.” Accessed: Jul. 07, 2023. [Online]. Available: https://github.com/davidsbatista/Aspect-Based-Sentiment-Analysis.

[4] A. Karimi, L. Rossi, and A. Prati, “Adversarial training for aspect-based sentiment analysis with BERT,” *arXiv preprint arXiv: 2001.11316*, Jan. 2020.

[5] N. U. Pannala, C. P. Nawarathna, J. T. K. Jayakody, L. Rupasinghe, and K. Krishnadeva, “Supervised learning-based approach to aspect-based sentiment analysis,” in *2016 IEEE International Conference on Computer and Information Technology (CIT)*, IEEE, Dec. 2016, pp. 662–666, doi: 10.1109/CIT.2016.107.

[6] A. Prabhune, V. R. Srihari, N. K. Sethiya, and M. Gauniyal, “Agile fusion: developing ‘Eat at Right Place’ sentiment analysis tool,” *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 34, no. 1, Apr. 2024, doi: 10.11591/ijeecs.v34.i1.pp602-619.

[7] A. G. Prabhune, V. R. S. Hari, and N. K. Sethiya, “Adaptation of a BERT model to the India restaurant data using rule-based approach for aspect-based sentiment analysis,” in *2023 Second International Conference on Smart Technologies for Smart Nation (SmartTechCon)*, IEEE, Aug. 2023, pp. 457–461, doi: 10.1109/SmartTechCon57526.2023.10391510.

[8] J. Bishop and A. Tritscher, “Food safety surveillance and response,” *Western Pacific Surveillance and Response Journal*, vol. 3, no. 2, pp. 1–3, Apr. 2012, doi: 10.5365/wpsar.2012.3.2.013.

[9] C. Kohli and S. Garg, “Food safety in India: An unfinished agenda,” *MAMC Journal of Medical Sciences*, vol. 1, no. 3, 2015, doi: 10.4103/2394-7438.166308.

[10] B. J. McCabe-Sellers and S. E. Beattie, “Food safety: Emerging trends in foodborne illness surveillance and prevention,” *Journal of the American Dietetic Association*, vol. 104, no. 11, pp. 1708–1717, Nov. 2004, doi: 10.1016/j.jada.2004.08.028.

[11] Public Health Agency of Canada, Canada, and Health Canada, “Framework for tools for evaluating health surveillance systems,” Ottawa, Ont.: Centre for Surveillance Coordination Population and Public Health Branch. Accessed: Jun. 21, 2019. [Online]. Available: http://ra.ocls.ca/ra/login.aspx?inst=centennial&url=https://www.deslibris.ca/ID/202166.

[12] International Life Sciences Institute-India (ILSI), “A surveillance and monitoring system for food safety for India (Under food safety and standards act-2005),” International Life Sciences Institute-India (ILSI), Accessed: Mar. 11, 2024. [Online]. Available: https://www.ilsi-india.org/PDF/Report_on_Food_Safety_Surveillance_and_Monitoring_System_for_India.pdf.

[13] R. Ghosh and D. Lewis, “Aims and approaches of Web-RADR: a consortium ensuring reliable ADR reporting via mobile devices and new insights from social media,” *Expert Opinion on Drug Safety*, vol. 14, no. 12, pp. 1845–1853, Dec. 2015, doi: 10.1517/14740338.2015.1096342.

[14] B. Audeh, F. Bellet, M.-N. Beyens, A. Lillo-Le Louët, and C. Bousquet, “Use of social media for pharmacovigilance activities: key findings and recommendations from the Vigi4Med project,” *Drug Safety*, vol. 43, no. 9, pp. 835–851, Sep. 2020, doi: 10.1007/s40264-020-00951-2.

[15] “2022 annual report on EudraVigilance for the European Parliament the Council and the Commission,” European Medicines Agency. [Online]. Available: https://www.ema.europa.eu/en/documents/report/2022-annual-report-eudravigilance-european-parliament-council-and-commission_en.pdf.

[16] P. Nsubuga *et al.*, *Public Health Surveillance: A Tool for Targeting and Monitoring Interventions*, Disease Control Priorities in Developing Countries. 2nd edition. Washington (DC): The International Bank for Reconstruction and Development/The World Bank; 2006.

[17] J. Prades, J. Ferré, and J. L. Gonzalo, “Journalists and bloggers. Professional identities and practices in food risk/benefits communication in Spain,” *Communication and Society*, vol. 27, no. 1, pp. 1–21, Dec. 2019, doi: 10.15581/003.27.36003.

[18] C. K. H. Lee, “Predicting food safety violations via social media to improve public health surveillance,” *European Conference on Social Media*, vol. 10, no. 1, pp. 109–116, May 2023, doi: 10.34190/ecsm.10.1.1009.

[19] “Aspect-based-sentiment-analysis,” Scalac. Accessed: May 09, 2024. [Online]. Available: https://github.com/ScalaConsultants/Aspect-Based-Sentiment-Analysis

[20] C. C. Yang, H. Yang, L. Jiang, and M. Zhang, “Social media mining for drug safety signal detection,” in *Proceedings of the 2012 international workshop on Smart health and wellbeing*, New York, NY, USA: ACM, Oct. 2012, pp. 33–40, doi: 10.1145/2389707.2389714.

[21] Y. Zhang, S. Cui, and H. Gao, “Adverse drug reaction detection on social media with deep linguistic features,” *Journal of Biomedical Informatics*, vol. 106, Jun. 2020, doi: 10.1016/j.jbi.2020.103437.

[22] C. Kakalou, S. Dimitsaki, V. K. Dimitriadis, and P. Natsiavas, “Exploiting social media for active pharmacovigilance: The PVClinical social media workspace,” in *Studies in Health Technology and Informatics*, 2022, doi: 10.3233/SHTI220176.

[23] S. Scaboro, B. Portelli, E. Chersoni, E. Santus, and G. Serra, “Increasing adverse drug events extraction robustness on social media: Case study on negation and speculation,” *Experimental Biology and Medicine*, vol. 247, no. 22, pp. 2003–2014, Nov. 2022, doi: 10.1177/15353702221128577.

[24] T. Hoang *et al.*, “Authenticity and credibility aware detection of adverse drug events from social media,” *International Journal of Medical Informatics*, vol. 120, pp. 101–115, Dec. 2018, doi: 10.1016/j.ijmedinf.2018.09.002.

[25] G. E. Powell *et al.*, “Social media listening for routine post-marketing safety surveillance,” *Drug Safety*, vol. 39, no. 5, pp. 443–454, May 2016, doi: 10.1007/s40264-015-0385-6.

[26] X. Chen *et al.*, “Mining adverse drug reactions in social media with named entity recognition and semantic methods,” *Studies in Health Technology and Informatics*, vol. 245, pp. 322–326, 2017, doi: 10.3233/978-1-61499-830-3-322.

[27] Z. Li, Z. Yang, L. Wang, Y. Zhang, H. Lin, and J. Wang, “Lexicon knowledge boosted interaction graph network for adverse drug reaction recognition from social media,” *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 7, pp. 2777–2786, Jul. 2021, doi: 10.1109/JBHI.2020.3042549.

[28] A. MacKinlay, H. Aamer, and A. J. Yepes, “Detection of adverse drug reactions using medical named entities on Twitter,” *AMIA Annual Symposium proceedings. AMIA Symposium*, pp. 1215–1224, 2017.

[29] M. S. Duh *et al.*, “Can social media data lead to earlier detection of drug-related adverse events?,” *Pharmacoepidemiology and Drug Safety*, vol. 25, no. 12, pp. 1425–1433, Dec. 2016, doi: 10.1002/pds.4090.

[30] O. Oyebode and R. Orji, "Identifying adverse drug reactions from patient reviews on social media using natural language processing," *Health Informatics Journal*, vol. 29, no. 1, Jan. 2023, doi: 10.1177/14604582221136712.

[31] J. Liu, S. Zhao, and G. Wang, "SSEL-ADE: A semi-supervised ensemble learning framework for extracting adverse drug events from social media," *Artificial Intelligence in Medicine*, vol. 84, pp. 34–49, Jan. 2018, doi: 10.1016/j.artmed.2017.10.003.

[32] J. Bian, U. Topaloglu, and F. Yu, "Towards large-scale twitter mining for drug-related adverse events," in *Proceedings of the 2012 International Workshop on Smart Health and Wellbeing*, New York, NY, USA: ACM, Oct. 2012, pp. 25–32, doi: 10.1145/2389707.2389713.

[33] O. Caster *et al.*, "Assessment of the utility of social media for broad-ranging statistical signal detection in pharmacovigilance: results from the WEB-RADR project," *Drug Safety*, vol. 41, no. 12, pp. 1355–1369, Dec. 2018, doi: 10.1007/s40264-018-0699-2.

[34] D. Tao *et al.*, "A novel foodborne illness detection and web application tool based on social media," *Foods*, vol. 12, no. 14, Jul. 2023, doi: 10.3390/foods12142769.

[35] J. K. Harris *et al.*, "Evaluating the implementation of a twitter-based foodborne illness reporting tool in the City of St. Louis Department of Health," *International Journal of Environmental Research and Public Health*, vol. 15, no. 5, Apr. 2018.

[36] D. M. Goldberg, S. Khan, N. Zaman, R. J. Gruss, and A. S. Abrahams, "Text mining approaches for postmarket food safety surveillance using online media," *Risk Analysis*, vol. 42, no. 8, pp. 1749–1768, Aug. 2022, doi: 10.1111/risa.13651.

[37] J. S. Brownstein, C. C. Freifeld, and L. C. Madoff, "Digital disease detection-harnessing the web for public health surveillance," *New England Journal of Medicine*, vol. 360, no. 21, pp. 2153–2157, May 2009, doi: 10.1056/NEJMp0900702.

[38] J. K. Harris, R. Mansour, B. Choucair, J. Olson, C. Nissen, and J. Bhatt, "Health department use of social media to identify foodborne illness-Chicago, Illinois, 2013-2014," *MMWR. Morbidity and Mortality Weekly Report*, pp. 681–685, 2014.

[39] E. Nsoesie, J. Hawkins, G. Tuli, S. Kluberg, and J. Brownstein, "The use of social media and business reviews for foodborne illness surveillance," *International Journal of Infectious Diseases*, vol. 53, Dec. 2016, doi: 10.1016/j.ijid.2016.11.177.

[40] L. van Velsen, J. E. W. C. van Gemert-Pijnen, D. J. M. Beaujean, J. Wentzel, and J. E. van Steenbergen, "Should health organizations use Web 2.0 Media in times of an infectious disease crisis? An in-depth qualitative study of citizens' information behavior during an EHEC outbreak," *Journal of Medical Internet Research*, vol. 14, no. 6, Dec. 2012, doi: 10.2196/jmir.2123.

[41] K. N. Overbey, L.-A. Jaykus, and B. J. Chapman, "A systematic review of the use of social media for food safety risk communication," *Journal of Food Protection*, vol. 80, no. 9, pp. 1537–1549, Sep. 2017, doi: 10.4315/0362-028X.JFP-16-345.

[42] M. Siering, "Leveraging online review platforms to support public policy: Predicting restaurant health violations based on online reviews," *Decision Support Systems*, vol. 143, Apr. 2021, doi: 10.1016/j.dss.2020.113474.

[43] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv: 1810.04805v2*, Oct. 2018.

BIOGRAPHIES OF AUTHORS

Akash Prabhune is a health researcher with 5+ years of experience in health systems strengthening, digital health, and AI/ML applications in healthcare management. He has worked with various state governments, the Indian Council of Medical Research, and several national and international funding agencies to strengthen healthcare delivery using data analytics and build a comprehensive healthcare system. He has conducted training sessions for ASHA workers, nurses, and AYUSH physicians for using digital tools in primary healthcare. He has published 12+ research papers and multiple technical reports, collaborated with researchers across various countries, and worked in multisectoral teams. He can be contacted at email: prabhunesky@gmail.com.

Vinay Sri hari is a data scientist at Institute of Health Management Research (IIHMR) in Bengaluru, focusing on public healthcare, supply chain optimization, machine learning, and data analytics. With a Master's in machine learning from Ramaiah University and a BE in Electronics and Communication Engineering from JSSATE, Bengaluru, Vinay has been instrumental in key projects like SDG goals and healthcare personnel optimization in Karnataka. His work includes the computation of indices like MPI, Health Index, and GII, and building models for health infrastructure optimization and drug procurement. Previously at Cognizant, he developed websites using Java Spring and Hibernate. Vinay's contributions are marked by a commitment to leveraging data science for social good. He can be contacted at email: vinay.rs@iihrmbangalore.edu.in.

Neeraj Kumar Sethiya joined Faculty of Pharmacy, DIT University on 2nd July 2019 in the capacity of Senior Assistant Professor and on August 2022 he has been promoted as Associate Professor. He had also worked as Head Department, to establish Centre for Innovation, Incubation, Entrepreneurship and Startup at DIT University. Prior to this, he has worked in the position of Senior Innovation Fellow, at National Innovation Foundation-India (An Autonomous Body of DST-Govt of India), Formulation Scientist, at Regen Biocorps, Vadodara and Teaching Assistant at Pharmacy Department, The MS University of Baroda, Vadodara, Gujarat. He has total 15 year of mixed experience including teaching, research and innovation. He is recipient of Bharat Gaurav Purushkar, Academic Excellence Award, Excellence in Performance Award, SPER Young Scientist Award, Dr. PD Sethi Best Research Paper and UGC JRF Research Fellowship (for perusing both Post Graduation and Ph.D.). During his tenure, he achieved H. index-22, and total citation of 2100 (Google Scholar Score) with 48 peer reviewed paper (with<100 impact factor), 6 book chapter, 7 books and 15 patents**.** He can be contacted at email: neeraj.sethiya@dituniversity.edu.in.

Mansi Gauniyal is a distinguished Anthropologist with a Ph.D. from the University of Delhi, specializing in Physical Anthropology. Her expertise lies in molecular genetic research, particularly regarding India's tribal populations. Her notable doctoral work on the Siddi tribe in Karnataka involved comprehensive genetic admixture analysis. She has significantly contributed to Public Health research, focusing on HIV/AIDS and Gestational Diabetes, and has a keen interest in chronic non-communicable diseases, including diabetes and cardio-metabolic conditions. Her proficiency in both qualitative and quantitative methodologies for epidemiological studies sets her apart in her field. At FHTS, she embraces a SMAART approach to evidence-based public health research, contributing to community well-being. Additionally, she mentors Ph.D. scholars, shaping the future of public health professionals. She works exemplifies her dedication to advancing health research and improving diverse population health. She can be contacted at email: mansigauniyal@gmail.com.