Social Media-Based Surveillance Systems for Healthcare using Machine Learning

Dr. Chetanpal Singh, Dr. Rahul Thakkar, and Jatinder Warraich

Abstract — One of the most popular domains that have caught the attention of researchers is real-time surveillance in the health and informatics segment. Many initiatives have been discovered due to this real-time surveillance surrounding public health informatics. Real-time surveillance in the health and informatics field has used the information from social media to predict the outbreak of diseases as well as to look after the diseases. There is no doubt in the fact that the availability of the data from social media in the recent past, especially the data from Twitter, has offered the researchers real-time syndromic surveillance in making quick analyses and conclusions in investigating the disease outbreak. The paper will get to know about the recent work of machine learning trends and text classification that has been utilized by the surveillance system by using the data from social media in the field of healthcare. Apart from this, the paper has also discussed the various limitations and challenges by taking into account the future direction that can be considered in this domain further.

Keywords — Disease Prediction, Health Prediction, Instagram, Machine Learning, Outbreak, Social Media, Surveillance Systems, Twitter.

I. INTRODUCTION

To enhance public health surveillance, the use of health information available on the internet has been seen as an opportunity. The health and surveillance system has always depended on the established system of mandatory as well as voluntary reporting of infectious diseases by the doctors in the laboratories [1]. As of now, social media data enables direct access to the data that would help in the surveillance epidemiology used to monitor the various public health threats like new diseases or pandemic-related early-level warnings. However, whether data from social media and the internet would help analyze potential public health threats remains a question. The paper has highlighted that there is a wide-reaching application in the field of public health surveillance in this century with the challenges of utilizing the imaging surveillance system for the sake of infectious disease epidemiology such as the specific resource needed, the technical essentials, and the acceptance of the public health practitioner as well as a policymaker [2].

There are many machine learning algorithms such as the deep neural network, Naive Bayes, Multinomial Naive Bayes, etc. have been utilized and proposed for the sake of the epidemic prediction classification approach at the time of looking at the surveillance system in the health informatics dormant [3].

A. Research Objective

In the paper, one will get to deal with the latest trend based on social media in the field of healthcare. Apart from this, the overview of the machine learning algorithm used for monitoring the data is also discussed in the paper [3].

Listed below are the Research questions that you should go through.

RQ1: what type of machine learning has become popular among the authors of the various research papers at the time of developing social media-based surveillance systems in the field of the health sector?

RQ2: what are some of the most popular social media data that are used for civilians in the field of healthcare domain?

RQ3: what is the implementation of a social media-based surveillance system in the field of health informatics?

RQ4: whether there are any challenges experienced by the syndromic surveillance system by the inclusion of the data from social media?

B. Research Motivation

One upon reading this paper will get to have a look at the following contribution that was not there in the previous paper. The concept of article selection query taken from different digital library databases for choosing the relevant article is there in this paper [4]. The research paper has discussed the overview of the popular machine learning classification algorithm related to social media-based surveillance systems in the field of healthcare. The paper has also conducted its statistical analysis on the social media platform as well as health topics that have been studied by the particular articles [5].

C. Research Gap

Social media data has a major role to play in making healthcare decisions even though social media has a lot of usage in other sectors. To reinforce the capability of the traditional syndromic surveillance system and the early detection of the disease and immediate public health response, there is a need for new approaches and technologies. To review the different surveillance systems that utilize social media data, many research papers have been prepared. All these papers have successfully covered the various data sources, technologies, algorithms, application, and evaluations. According to a recent review of the surveillance system in the field of health informatics using social media, the researchers have not been much impressed with the development [2]. The paper will give a complete

Submitted on October 10, 2022.
Published on November 07, 2022.
C. Singh, VIT, Australia.
(e-mail: Chetanpal.singh@vit.edu.au)
R. Thakkar, VIT, Australia.
(e-mail: rahul.thakkar@vit.edu.au)
J. Warraich, Holmesglen, Australia.
(e-mail: Jatinder.warraich@holmesglen.edu.au).

DOI: http://dx.doi.org/10.24018/eje2022.7.6.2914

Vol 7 | Issue 6 | November 2022
analysis of machine learning technology and its approaches in this field specifically in the recent past. Apart from this one will also get to know about the challenges and future directions of it.

II. LITERATURE REVIEW

A. Machine Learning Methods Utilized by Surveillance Systems to Process Social Media Data

There is no doubt in fact about the growing popularity of machine learning in the recent past in detecting the various patterns in images on raw data. According to [6], he has concluded the progression in machine learning offers epidemiologists the to mine with the help of a broad set of digital data. To detect the personal health experience as well as the deep granulater approach to enhance decisions at the time of applying to the independent test set, there has been a study of several supervised machine learning algorithms by [7]. Another researcher [8] has provided a detailed analysis of the conjunction of natural language processing as well as machine learning with the various platforms of social media to assist in the analysis of huge datasets for the sake of population-level mental health research. Moreover, some architecture is still popular among the various methodological variations of machine learning.

There is no question of the requirement of labeled data sets in predicting the output in an unsupervised algorithm like it was required in a supervised classification algorithm. This is the reason why the unsupervised classification method is known to be a more popular alternative in analyzing the text; however, this method is challenging in achieving the same percentage of accuracy as a supervised method. The same thing is seen when [9] provided tweet classification with the help of supervised and unsupervised methods. [9] has discussed topic modeling which is known to be one of the best-supervised techniques that present control over topic contents in contrast to the old classified specifically when it is a naturally noisy media channel.

1) Multinomial naive bayes

To classify the Twitter content, one of the most popular supervised classification approaches was followed which is known as Multinomial Naive Bayes. [10] has come up with a real-time allergy surveillance system that has helped in the classification of tweets as either positive or negative and when it is the positive tweet it highlights the person or any other person who is the person that beholds all the allergy symptoms. If it mentions things such as news, and advertisements for general awareness of allergies then the tweet is classified as negative. The author has come to the conclusion that the Naive Bayes Multinomial model with an F measure is the best solution for the text classification performance. On the other hand, [11] has utilized machine learning methods so that they can classify the tweets based on personal or news related.

The author also went with classification of personal tweets into a couple of categories that are negative or neutral tweets. The NBM has provided the best result and is known to be better than the other two techniques used. The classifier such as in Naive Bayes and SVM has not produced satisfactory results as Naive Bayes Multinomial model [12].

<table>
<thead>
<tr>
<th>Methods used</th>
<th>Application</th>
<th>Outcome</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multinomial Naive Bayes</td>
<td>Real-time allergy surveillance system for classification of tweets as either positive or negative</td>
<td>Naive Bayes Multinomial model with an F measure is the best solution for the text classification performance</td>
<td>[10]</td>
</tr>
<tr>
<td>Multinomial Naive Bayes, Naive Bayes and SVM</td>
<td>Classification of tweets based on personal, or news related</td>
<td>NBM has provided the best result</td>
<td>[11], [12]</td>
</tr>
</tbody>
</table>

2) Support vector machine

One upon reading the paper will get to know that the dependency of input parameter and application is high for the performance of the classification algorithm, however, when the classification task was taken into account, SVM is best suited. [13] with the help of the SVM classification model, was successfully able to classify sick microblog and non-sick microblog posts. The author also highlighted the time consumption by SVM that is required for the classification task was not affected at the time of arranging the microblog increment in the consumption of time by KNN in completing the classification task. According to [13], the best classification method was known to be an SVM when it was differentiated from the various other techniques of machine learning. The SVM method plays a crucial role in classifying the various data from social media on a range of health issues.

According to [14], the SVM classifier is best suited when accuracy in the prediction of the class of tweets was taken into account. At the same time, as per a study by [15], this algorithm can reach 90% accuracy when the tweets from social media were segregated as epidemiological and non-epidemiological.
### 3) Deep neural network

The convolutional deep neural network has a crucial role in the classification of text in the field of health. [16] has utilized the various types of DNN that is the convolutional neural network as well as bidirectional long short-term memory by combining machine learning approaches that would help in the classification of measles-related tweet classification tasks and the researchers have pointed out that the convolutional neural network has provided remarkable result.

![Fig. 3. Deep Neural Network](image)

### TABLE III: CNN BLSTM APPROACH OUTCOME

<table>
<thead>
<tr>
<th>Methods used</th>
<th>Application</th>
<th>Outcome</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Convolutional neural network (CNN), bidirectional long short term memory (BLSTM)</td>
<td>Classification of measles-related tweet classification tasks</td>
<td>CNN has provided remarkable result</td>
<td>[16]</td>
</tr>
</tbody>
</table>

### 4) Decision tree

For predicting positive as well as negative tweets surrounding the personal health experience, the decision tree classifier has played a crucial role and has performed well [17]. Apart from this, the approach of the decision tree classifier was also utilized by [18] to differentiate tweets surrounding the swine flu. [19] has achieved a result that is average with the help of a decision tree classifier for the sake of classifying the personal health experience tweets.

#### 1) Logistic regression

Logistic regression is another popular choice; it is used for data classification tasks among the other classification algorithms. In a study by [20], Logistic regression towards updating a record showed better F1 measure and recall as compared to the SVM in terms of classification of relevant and irrelevant tweets regarding asthma. Apart from this, the usage of a maximum entropy classifier is also seen in the research paper which is significantly utilized for the classification of text [21]. [17] has also used the logistic regression classifier for the research. Moreover, the illness tweets were monitored with the help of the maximum entropy [22], and another study performed for tweet classification also used the maximum entropy [23].

![Fig. 4. Decision Tree Algorithm](image)

### TABLE IV: DECISION TREE CLASSIFIER APPROACH OUTCOME

<table>
<thead>
<tr>
<th>Methods used</th>
<th>Application</th>
<th>Outcome</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Decision tree classifier</td>
<td>Predicting positive as well as negative tweets surrounding the personal health experience</td>
<td>Decision Tree gives good performance</td>
<td>[17]</td>
</tr>
<tr>
<td>Decision tree classifier</td>
<td>To differentiate tweets surrounding the swine flu</td>
<td>Good performance by decision tree</td>
<td>[18]</td>
</tr>
<tr>
<td>Decision tree classifier</td>
<td>For classifying the personal health experience tweets</td>
<td>Decision Tree gives average result</td>
<td>[19]</td>
</tr>
</tbody>
</table>

![Fig. 5. Logistic Regression](image)
The author has taken into account all the swine Flu-related words and identified that the Naive Bias and SVM has given the best outcome with a measure of 0.77. The classification algorithm has become popular, and the authors have considered using this for the text classification that shows the average performance when it is compared to the various other classifiers [24]. The best performance was given by the Naive Bayes classifier when the dengue-suspected tweet was taken into account and marked as irrelevant or relevant. For this, various bigrams, emojis, trigrams, and location information were also considered [23].

3) Random forest

Social media text classification is utilized by the random forest approach combined with conventional machine learning approaches. [25] has experimented with the various approaches with the naive Bayes classifier and the outcome of the result came to be that the former is far better than the Naive Bayes method. The various types of machine learning approaches were also experimented with to deal with text mining such as clustering, k means, etc. [13], [24]. The grounds of similar words were utilized to group the tweet and the tweet can also be differentiated based on the similarity measure.

![Diagram of Random Forest](image)

Fig. 7. Random Forest [25].

### Table V: Logistic Regression Approach Outcome

<table>
<thead>
<tr>
<th>Methods used</th>
<th>Application</th>
<th>Outcome</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>logistic regression</td>
<td>Text classification</td>
<td>-</td>
<td>[17]</td>
</tr>
<tr>
<td></td>
<td>For classification of relevant and</td>
<td>Logistic regression</td>
<td></td>
</tr>
<tr>
<td></td>
<td>irrelevant tweets regarding asthma</td>
<td>showed better F1 measure and</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>recall as compared to SVM</td>
<td></td>
</tr>
<tr>
<td>Logistic regression</td>
<td>Text classification</td>
<td>-</td>
<td>[20]</td>
</tr>
<tr>
<td>Maximum entropy</td>
<td>Text classification</td>
<td>-</td>
<td>[21], [23]</td>
</tr>
<tr>
<td>classifier</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Maximum entropy</td>
<td>Monitoring illness tweets</td>
<td>-</td>
<td>[22]</td>
</tr>
<tr>
<td>classifier</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Table VI: Naive Bayes and SVM Approach Outcome

<table>
<thead>
<tr>
<th>Methods used</th>
<th>Application</th>
<th>Outcome</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM and Naive Bayes</td>
<td>classification of data sets into mosquito-borne</td>
<td>Further classification of tweets was possible</td>
<td>[14]</td>
</tr>
<tr>
<td></td>
<td>To differentiate swine Flu-related text from the noise of all the tweets</td>
<td>Naive Bias and SVM has given the best outcome with a measure of 0.77 as compared to decision tree and random forest</td>
<td>[18]</td>
</tr>
<tr>
<td>Naive Bayes, SVM</td>
<td>For text classification</td>
<td>Naive Bayes gives average performance compared to other classifiers</td>
<td>[24]</td>
</tr>
<tr>
<td></td>
<td>Dengue suspected tweet was marked as irrelevant or relevant</td>
<td>Naive Bayes classifier gave the best performance</td>
<td>[23]</td>
</tr>
</tbody>
</table>

### Table VII: Random Forest, Naive Bayes, Random Forest and K Means Approach Outcome

<table>
<thead>
<tr>
<th>Methods used</th>
<th>Application</th>
<th>Outcome</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random Forest Classifier, Naive Bayes</td>
<td>Social media text classification</td>
<td>Random Forest gave better performance</td>
<td>[25]</td>
</tr>
<tr>
<td>Random Forest Classifier, clustering, k means</td>
<td>Text mining for differentiating the tweets</td>
<td>Random Forest gave better performance</td>
<td>[13], [24]</td>
</tr>
</tbody>
</table>

4) K nearest neighbor

The utilization of KNN with Naive Bayes, SVM, and Naive Bayes multinomial was done by [12] to figure out and monitor messages reporting and discuss various types of allergies. The author has come up with the conclusion that the k-NN has better precision as compared to the other approaches used by the author in identifying and assigning the tweets whether it is an actual incident of allergy or an awareness tweet.
processing approach has been considered [30]. [31] has performed the study to identify the potential of the social media platform as a new way of sharing information. The various posts from Twitter are trusted by millions of users and the media postings are also considered as a fast source to identify the incidence of diseases in the population and hence the researchers feel it's important to find an efficient method so that the health-related tweets can be examined and processed easily.

Moreover, for the partition vector representation, the unsupervised method was utilized by [24]. [10] has done research based on the allergy activities in their collection of tweets that have allergy-related tweets mentioned in them. Twitter was also helpful in detecting health problems like respiratory, gastrointestinal, health-related problems. The data from Twitter was used to study a variety of public health issues like allergy, mosquito-borne disease, dengue etcetera [5].

2) Instagram

In 2010, another popular social media platform was founded known as Instagram. Instagram is a popular photo and video sharing platform and since it was founded in 2010 the number of registered users rose to 800 hundred million [32]. The reason why Instagram has provided satisfactory results is that it is a photo and video sharing platform and the data that was there in this platform can be a good source for the surveillance of the disease [32], [33] have studied the Ebola-related social media posts on a couple of social media platforms is Twitter and Instagram and the outcome has highlighted that the best platform among the two for communication and reaching the people at the time of health crisis is Instagram.

C. Application of Social Media-Based Surveillance System

This section talks about the recent application of the popular surveillance system in the health informatics domain. The various recent applications include the prediction of disease tracking misinformation and global awareness.

<table>
<thead>
<tr>
<th>Methods used</th>
<th>Application</th>
<th>Outcome</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>KNN with Naive Bayes, SVM, and Naive Bayes Multinomial (NBM)</td>
<td>To figure out and monitor messages identifying tweets as actual incident</td>
<td>K-NN has better precision for of allergy or awareness tweets</td>
<td>[12]</td>
</tr>
</tbody>
</table>

B. Various Kinds of Social Media Data Sources for Data Collection

The number of social media users is increasing with time and the various social media users share different information thus the researcher needs to track the significant information so that they can monitor the various activities in social media related to public health purposes. Social media also has exposure to different kinds of topics apart from public health.

There is no doubt in the fact that social media platform is one of the best platforms where one can get a lot of information about public health. But some researchers are still in doubt that the data from social media would play a crucial role in detecting an outbreak and analyzing the content of social media for healthcare data [26]. Various posts on social media and online search behaviour can act as a very crucial source of information related to health outbreaks.

1) Twitter

One of the most popular microblogging services is Twitter which has a lot of users who are posting to tweet, and it's related to various posts that the unregistered user can also read. Twitter is one of the leading microblogging services has more than 300 million monthly active users and this is the reason why the social media platform can be trusted and can identify the various incidents of diseases in mankind. A set of seven terms were used by [27] to gather tweets of more than 50000 and study and classify them by analyzing cardiac arrest and resuscitation. For the sake of surveillance of disease, some of the factors such as location, volume, time as well as public perceptions are taken into account [28].

There is recent work that was done by [29] for the various health organizations where the information was collected from the social media platform and was used to figure out the information at the time of the epidemic which has been very helpful to the various health organizations.

To get the Ebola-related tweets that are considered in 4 topics such as risk factors, prevention, education, disease trends as well as compassion, the usage of a natural language processing approach has been considered [30]. [31] has performed the study to identify the potential of the social media platform as a new way of sharing information. The various posts from Twitter are trusted by millions of users and the media postings are also considered as a fast source to identify the incidence of diseases in the population and hence the researchers feel it's important to find an efficient method so that the health-related tweets can be examined and processed easily.

 Moreover, for the partition vector representation, the unsupervised method was utilized by [24]. [10] has done research based on the allergy activities in their collection of tweets that have allergy-related tweets mentioned in them. Twitter was also helpful in detecting health problems like respiratory, gastrointestinal, health-related problems. The data from Twitter was used to study a variety of public health issues like allergy, mosquito-borne disease, dengue etcetera [5].

2) Instagram

In 2010, another popular social media platform was founded known as Instagram. Instagram is a popular photo and video sharing platform and since it was founded in 2010 the number of registered users rose to 800 hundred million [32]. The reason why Instagram has provided satisfactory results is that it is a photo and video sharing platform and the data that was there in this platform can be a good source for the surveillance of the disease [32], [33] have studied the Ebola-related social media posts on a couple of social media platforms is Twitter and Instagram and the outcome has highlighted that the best platform among the two for communication and reaching the people at the time of health crisis is Instagram.

C. Application of Social Media-Based Surveillance System

This section talks about the recent application of the popular surveillance system in the health informatics domain. The various recent applications include the prediction of disease tracking misinformation and global awareness.

<table>
<thead>
<tr>
<th>Methods used</th>
<th>Application</th>
<th>Outcome</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>KNN with Naive Bayes, SVM, and Naive Bayes Multinomial (NBM)</td>
<td>To figure out and monitor messages identifying tweets as actual incident</td>
<td>K-NN has better precision for of allergy or awareness tweets</td>
<td>[12]</td>
</tr>
</tbody>
</table>

B. Various Kinds of Social Media Data Sources for Data Collection

The number of social media users is increasing with time and the various social media users share different information thus the researcher needs to track the significant information so that they can monitor the various activities in social media related to public health purposes. Social media also has exposure to different kinds of topics apart from public health.

There is no doubt in the fact that social media platform is one of the best platforms where one can get a lot of information about public health. But some researchers are still in doubt that the data from social media would play a crucial role in detecting an outbreak and analyzing the content of social media for healthcare data [26]. Various posts on social media and online search behaviour can act as a very crucial source of information related to health outbreaks.

1) Twitter

One of the most popular microblogging services is Twitter which has a lot of users who are posting to tweet, and it's related to various posts that the unregistered user can also read. Twitter is one of the leading microblogging services has more than 300 million monthly active users and this is the reason why the social media platform can be trusted and can identify the various incidents of diseases in mankind. A set of seven terms were used by [27] to gather tweets of more than 50000 and study and classify them by analyzing cardiac arrest and resuscitation. For the sake of surveillance of disease, some of the factors such as location, volume, time as well as public perceptions are taken into account [28].

There is recent work that was done by [29] for the various health organizations where the information was collected from the social media platform and was used to figure out the information at the time of the epidemic which has been very helpful to the various health organizations.

To get the Ebola-related tweets that are considered in 4 topics such as risk factors, prevention, education, disease trends as well as compassion, the usage of a natural language processing approach has been considered [30]. [31] has performed the study to identify the potential of the social media platform as a new way of sharing information. The various posts from Twitter are trusted by millions of users and the media postings are also considered as a fast source to identify the incidence of diseases in the population and hence the researchers feel it's important to find an efficient method so that the health-related tweets can be examined and processed easily.

Moreover, for the partition vector representation, the unsupervised method was utilized by [24]. [10] has done research based on the allergy activities in their collection of tweets that have allergy-related tweets mentioned in them. Twitter was also helpful in detecting health problems like respiratory, gastrointestinal, health-related problems. The data from Twitter was used to study a variety of public health issues like allergy, mosquito-borne disease, dengue etcetera [5].

2) Instagram

In 2010, another popular social media platform was founded known as Instagram. Instagram is a popular photo and video sharing platform and since it was founded in 2010 the number of registered users rose to 800 hundred million [32]. The reason why Instagram has provided satisfactory results is that it is a photo and video sharing platform and the data that was there in this platform can be a good source for the surveillance of the disease [32], [33] have studied the Ebola-related social media posts on a couple of social media platforms is Twitter and Instagram and the outcome has highlighted that the best platform among the two for communication and reaching the people at the time of health crisis is Instagram.

C. Application of Social Media-Based Surveillance System

This section talks about the recent application of the popular surveillance system in the health informatics domain. The various recent applications include the prediction of disease tracking misinformation and global awareness.
1) Global awareness of the event
The surveillance system plays a crucial role in monitoring general public awareness and also provides perception regarding health events once the detection of the event has been completed. The social media platform has user-generated sentiment regarding the outbreak situation that talks about the knowledge, attitude, and perception of the people [34]. The users on social media can share their sentiments, opinion, and response at the time of the outbreak.

2) Syndromic surveillance-based disease prediction
One of the best tools for predicting the outbreak for public health purposes with the help of data that is gathered from different sources is syndromic surveillance. That is all that was acquired from the tool is targeted to minimize the exposure of the disease to the population and to take proper measures and prevent it. The information from social media has been used in the past few years widely to study disease incidences and to figure out the outbreak of the disease. The data that was taken from social media is beneficial for the officials of public health in detecting the outbreak earlier than traditional means. Studies have shown that a surveillance system in the healthcare domain helps to predict diseases for the concerns of public health and the data of the surveillance system is in the type of self-reported symptom. For early warning and outbreak detection, the data from Twitter was utilized as a tool for predicting the swine flu, tuberculosis, Ebola, and syphilis [35]-[38]. In another study, the examination of disease incidence such as dengue as well as typhoid fever was taken into account [24].

3) Event-based surveillance and disease prediction
Event-based surveillance is a process where the data is captured very fast and in a proper manner about the various events that are at serious risk to the health of the public. The collection of data can come from diverse internet sources such as reports from the media, online discussion platforms, routine reporting systems, personal information or it can even from rumors. Talking about the web forum contacts, the definition of an event is defined as excessive news posting. For early warning and outbreak detection, the data from Twitter was utilized as a tool for predicting the swine flu, tuberculosis, Ebola, and syphilis [35]-[38]. In another study, the examination of disease incidence such as dengue as well as typhoid fever was taken into account [24].

4) Magnitude estimation of disease over sometime
The magnitude of the issue can be easily determined with the help of a surveillance system. The estimation of the future of the various diseases can easily be done by planning the allocation of resource treatment and prevention [2]. Moreover, surveillance system analysis can play a significant role in figuring out the disease level over a certain period and the assessment can be made on behalf of that.

III. METHODS AND MATERIALS
The research aims are to look into surveillance of social media-based systems by using the technique of machine learning to forecast illness in real-time or the situation that arises in the near real-time. The research article selection criteria were established to include were published in the year 2010 and 2018.

To compile thorough research in a bibliography format the publications on social media- which is a surveillance-based system in the area of healthcare, the following scientific databases were searched: IEEE Xplore, Science Direct, PubMed, and ACM Digital Library.

Now following and describing the query based in IEEE Xplore database:
The following query was formed using an advanced search of the IEEE Xplore database: ((("Abstract": surveillance) OR "Document Title": leadership OR "Abstract": outbreak)) When filters were applied, 656 items (Journals & Magazines and conferences) were found.

A. Describing and Running the Query-Based in the ACM Digital Library System
ACM Digital Library searched for query: record in a similar way. (((outbreak OR surveillance) OR acmdlTitle(\+surveillance) AND (health* OR illness) 265 articles were found using the search terms; in addition, we conducted an advanced search of the ScienceDirect database for the following terms in the title, abstract, and keywords: (surveillance OR outbreak) AND (health* OR illness) AND “social media”. As a result of the search keywords, 75 articles were retrieved.

Similarly, PubMed, which accesses the MEDLINE database, was utilized to find papers. (surveillance [Title/Abstract]) OR (((epidemic [Title/Abstract]) AND ((health[Title/Abstract]) OR ((disease[Title/Abstract])) AND (socialmedia[Title/Abstract]) AND (health[Title/Abstract]) OR (disease[Title/Abstract]) AND (socialmedia[Title/Abstract]) AND A total of 1240 articles were discovered for further research, out of which we acquire roughly 244 articles.

In addition, they are about our research and concluding the results and statistics of the same from Google Scholar were used for the same reflected recent trends. The words “surveillance system”, “social media”, “machine learning”, and “health informatics” were used in the study. Over time, this graph clearly illustrates a rise in the number of articles in healthcare. The use of social media data and the various machine learning algorithms is the central area of concern considered in the surveillance system.

Each of the papers found in our research which was around 1240 articles, was separately vetted based on the abstract, and titles were also considered by each of the paper's authors. We accepted them for further inquiry if the abstract or header, or both, explained social media or web-based monitoring; otherwise, they were dismissed. The second thing we did was look at the papers that included the algorithm of the machine learning techniques in their methodology.
IV. RESULTS AND DISCUSSION

While considering the system of surveillance, which was based on the data from social media for detection of the outbreak or a break-even point or health events that have improved early identification of epidemics and related events, other researchers have questioned the effectiveness of these monitoring systems for the following reasons:

A. Privacy Issues

Issues that arise from getting data from social media accounts, such as datasets that are private, are obtained for health purposes utilizing social media. Even though social media data is publicly available, individuals may not want their postings or data to be used for the study [41]. Users’ expectations, based on public data and privacy was considered a significant factor.

B. Verification of the Data Set

An issue with gathering data from social media is that it must be validated. Standardization, verification, and control issues may arise if unofficial data from social media is used [42]. The truth connected with a vast quantity of diverse data from social media was validated by [43].

The dataset is an essential part of the prediction model. The dataset has a significant impact on the outcomes of prediction models of the following:

i. Historical data
ii. Training data
iii. Testing data

All the dataset mentioned above is included in the prediction models. A considerable quantity of training data is necessary for testing predictions based on model training to forecast models.

C. Noise

Noise is one of the most significant issues encountered during data collection. The information gathered through social networking sites might include unrelated data to the goal. Such information on sickness words has no bearing on one's health. For example, posts featuring the keyword “Irish Flu” may trigger a slew of flu-related activity [20]. Unfortunately, there are situations when a user posts a status and is mistakenly assumed to be infected when they are not. In this way, false information might impact illness management in the public health department.

To eliminate such noise in data; nevertheless, additional training is required to obtain the relevant data for further analysis.

D. Bias Based on Demographics

Although social media can help collect demographic information like age, gender, and race. The determination is complex and managing the algorithm so that public healthcare efforts are directed in such a direction makes the task complex.

The research also supports the semi-demographic factor and excludes those who are not active on social media and the old children who are least involved in such platforms [44]. One of the few studies [45] looked at the users’ profiles looking at the Facebook comments posted about sex and discovered that males wrote more posts per person than women. The majority of social media users are under the age of thirty. The discovery that social media data is weighted towards the frequently active user and the data from young people further supports the bias.

E. Variability in Lexicon and Language

Though communication via social media aids in extracting healthcare data, it is difficult to evaluate the language semantically. Due to the informal and imprecise nature of social media communications, it results in an incomplete result. This constraint has been studied and adequately researched by [46].

F. Low Confidence

Low confidence is another issue that occurs when using social media data. The research presents a conspiracy concerning the Zika virus pandemic on Reddit during a public health crisis [47]. And for this, more training is required to reach and approach the algorithm with the classification feature.

According to a recent study [48], official websites are a more reliable source of vaccination information than social media. The quality of health-related data available on the internet varies. Many social media data analysis tools may indicate hyped data that something significant is happening. However, this might reflect panic rather than actual illness outbreaks.

Also, users may claim to have the flu when they have a regular cold, or others may discuss the sickness owing to heightened media coverage.

V. CONCLUSION

According to the data, Twitter was in the platform which was most searched. SVM was also the most often utilized classification approach. Furthermore, when data were categorized into two classes, SVM was the best classifier. This research looks at the most recent trends in public health monitoring systems that use different algorithms. Compared to traditional methods, it is found that social media-based surveillance systems outperform them. The paper has also spoken about how data collected can be further used to improve monitoring systems in the field of public health.

A. Future Work

The combination of internet data with the current circumstances such as weather, demographic data, and so on to improve forecast outcomes.

Combining the factors such as sentiment, comments, locations, and other input characteristics from user postings with text material for improved the overall analysis.

Sorting user postings into multiple categories that give varying weights to each class to increase predicting accuracy and text analysis of images can be extended. There is a lot of room for development in topic modeling to generate more precise findings.

REFERENCES

Preventing and combating mosquito surveillance: Word embedding based clustering method for Twitter named entity recognition using an aggregate classifier

Korkontzelos Mike, Yang Ji, A., Van der Meulen Janua, Bates M.


Kumar V., Kumar S. An Effective Approach to Track Levels of Influenza-A (H1N1) Pandemic in India Using Twitter. Procedia Computer Science, 2015; 70: 801-807.


Limospatham, Collier N. Towards the semantic interpretation of personal health messages from social media. Proceedings of the ACM First International Workshop on Understanding the City with Urban Informatics, 2015.
