A real-time biosurveillance mechanism for early-stage disease detection from microblogs: a case study of interconnection between emotional and climatic factors related to migraine disease

Sarsam, S. M., Al-Samarraie, H., Ismail, N., Zaqout, F. & Wright, B.

Original citation & hyperlink:
Sarsam, SM, Al-Samarraie, H, Ismail, N, Zaqout, F & Wright, B 2020, 'A real-time biosurveillance mechanism for early-stage disease detection from microblogs: a case study of interconnection between emotional and climatic factors related to migraine disease', *Network Modeling Analysis in Health Informatics and Bioinformatics*, vol. 9, no. 1, 32.
https://doi.org/10.1007/s13721-020-00239-6

DOI 10.1007/s13721-020-00239-6
ISSN 2192-6662
ESSN 2192-6670

Publisher: Springer

The final publication is available at Springer via http://dx.doi.org/10.1007/s13721-020-00239-6

A real-time biosurveillance mechanism for early-stage disease detection from microblogs: A case study of interconnection between emotional and climatic factors related to migraine disease

Abstract
For many years, certain climatic factors have been used to predict potential disease outcomes of relevance to humans. This is because early discovery of disease (or its symptoms) would help people or healthcare professionals to take the necessary precautions. Since microblogs can be used to create new connections and maintain existing relationships, disease detection in microblogs is still considered a serious problem for many healthcare systems, especially for establishing a successful epidemic recognition procedure. To tackle this issue, this study proposed a novel tracking approach to diagnose illnesses in microblogs. It is based on the interconnection between certain emotional type and climatic factors associated with a specific disease (e.g., migraine). In this study, detailed migraine data were collected from Twitter. We used K-means and Apriori algorithms to extract migraine-related emotions and investigate the potential associations between migraine symptoms and climatic factors. The results showed that sad emotions were highly interrelated with migraine symptoms. The classification results showed that Sequential Minimal Optimization (SMO) was efficient (95.53% accuracy) in detecting the migraine symptoms from Twitter. The proposed mechanism can be used efficiently in biosurveillance systems due to its capability in identifying the hidden symptoms of a sickness on microblogs. This study paves the way to discover disease-related features using both emotional and climatic factors.

Keywords: biosurveillance, disease detection, Twitter; machine learning; migraine

1. Introduction

Health organizations require accurate and timely disease surveillance techniques to respond to various emerging issues. Therefore, there has been increasing attention on the use of social media and the Internet as important resources for general disease surveillance (Priyadarshi & Saha, 2020). This, in turn, implies a new generation of surveillance strategies that were developed in order to assist the automatic detection process of the emerging infections (Zadeh, Zolbanin, Sharda, & Delen, 2019) that could be effectively analyzed via machine learning algorithms, due to their capability in handling big data and extracting the latent information (Richter & Khoshgoftaar, 2020). Thus, disease surveillance relied on Internet data and social media like Twitter, to perform an early-stage disease recognition due to the ability of such systems to be used in conjunction with conventional surveillance systems to enhance early warning of public health threats. Technically, the overall process of disease analysis involves collecting biosurveillance data, integrating structured and unstructured data, applying advanced machine learning algorithms, and evaluating the prediction result (Erraguntla, Zapletal, & Lawley, 2019). For early-stage disease detection, artificial intelligence methods are required to perform fast and accurate prediction (such as in the case of outbreaks of infectious diseases) which can be implemented using either supervised or unsupervised learning methods (Jordan et al., 2019). For instance, Molaei, Khansari, Veisi, and Salehi (2019) used a supervised learning technique to predict the early-stage of influenza-like illness based on real-time data derived from Twitter. The authors found that deep neural network and entropy-based methods can help minimize the mean average error by up to 25% compared to other nonlinear methods. Another study by Joshi et al. (2020) was carried out to predict the thunderstorm asthma from Twitter messages using a library for support vector machines and the support vector machine Pref.

Despite the number of studies on the use of machine learning for disease detection, very few studies revealed the potential of unsupervised machine learning tools in detecting or predicting early-stage disease using real-time data. For example, Lim, Tucker, and Kumara (2017) identified real-world hidden infectious diseases by mining social media data. The authors considered hidden infectious diseases to be communicable disease that have not yet been recognized by national public health institutes. They applied an unsupervised learning method on textual data together with the temporal information in an attempt to provide a bottom-up technique for latent infectious disease discovery in a given location. Our review of the literature showed the use of unsupervised machine learning tools in modeling clinical risk stratification (Huang, Dong, & Duan, 2015). This includes learning laboratory test reference intervals by laboratory results and coded diagnoses (Poole, Schroeder, & Shah, 2016). In addition, the use of Natural Language Processing (NLP) methods for early-stage disease recognition has shown a great potential in processing and understanding event narratives. According to Liu, Weng, and Yu (2019), NLP can be effectively used to understand individuals' health-related issues by analyzing syntax, semantics, morphology, pragmatics and discourse (Șerban, Thapen, Maginnis, Hankin, & Foot, 2019).

In order to further explore the potential of unsupervised machine learning techniques in real-time disease detection, migraine was used as a model disease in this study. Migraine is a headache disorder characterized by intense and debilitating headaches accompanied by visual or/and sensory symptoms and a prodromal phase (Cioffi et al., 2017). This type of disease is usually correlated with climatic factor changes (Hoffmann et al., 2015; Ravat, Chaudhari, & Chafekar, 2019; Zebenholzer et al., 2011). From a clinical perspective, migraine is unilateral in location and linked with nausea mostly (Chai, Rosenberg, & Peterlin, 2012; Ravat et al., 2019). Migraine is related to specific climatic conditions or changes over short periods, which includes seasonal changes, bright sunshine, strong winds, hot weather, cold weather or thunderstorms. Nevertheless, other weather conditions, including ambient temperature and humidity, may differ based on whether the individual is outdoor or indoor. It is, therefore, essential to evaluate such weather conditions using portable monitors in everyday life settings.

Performing a real-time biosurveillance process to recognize a particular disease is a crucial task to be accomplished in social media platforms. Internet biosurveillance utilizes the data sources found on the Internet (such as news and social media) to improve certain detection, situational awareness, and forecasting of epidemiological events (Hartley et al., 2013; Sabatovych, 2019a; Zaeem, Liau, & Barber, 2018). In this context, people generally use Twitter, a form of social media that became a key communication medium and reliable source of information in recent years that addresses general real-life events (Naslund et al., 2019; Nejad, Delghandi, Bali, & Hosseinzadeh, 2020; Sabatovych, 2019b; Weng & Lee, 2011). These characteristics motivated several scholars to adopt Twitter data and delve into people's health-related issues. For example, Culotta (2010) analyzed over 500 million tweets and found that tracking a small number of flu-related keywords can help healthcare organizations forecast influenza rates with high accuracy. Aramaki, Maskawa, and Morita (2011) proposed a Twitter-based influenza epidemics detection method by using NLP and Support Vector Machine (SVM) algorithm to find out the negative influenza tweets posted by individuals who did not catch the disease. They found that posts generated from the Twitter platform resulted in a precise detection of influenza. Besides, Lamb, Paul, and Dredze (2013) presented a novel approach using a log-linear model for analyzing differences between flu infection and concerned awareness tweets. Byrd, Mansurov, and Baysal (2016) developed an approach with a set of methods to collect data from Twitter by classifying the obtained tweets based on their sentiment characteristics to identify Twitter users who are affected by a disease outbreak.

The important aspect for accurate social media surveillance is in identifying the geographic location of each tweet with high accuracy (Burton, Tanner, Giraud-Carrier, West, & Barnes, 2012; Savini et al., 2018). Broniatowski, Paul, and Dredze (2013) developed an influenza infection detection algorithm that automatically recognizes disease-relevant tweets in a location-specific manner. So, to predict the influenza predominance, the researchers only used the geolocation system (e.g., Carmen) to detect the location of 22% of the collected tweets. Based on these observations, one can observe that a major challenge facing event detection from Twitter streams is in the separation of irrelevant information from other relevant details (Atefeh & Khreich, 2015).

Hence, this study proposed a new mechanism for event detection based on the interconnection between certain emotional and climatic factors. It used migraine disease as the disease model. The following questions are dealt with:

1. What types of emotions are associated with migraine in social media?
2. What are the climatic factors related to the migraine disease?
3. How to detect the detecting the migraine disease on social media?

In this study, we have three main objectives: extracting the sentimental features related to migraine disease from Twitter data; identifying the climatic factors associated with migraine; and finding the best classification scheme for the migraine detection task. We believe that the proposed approach can facilitate the detection process of migraine disease by identifying the relevant disease-related emotions and its climatic factors from microblogs.

2. Literature review

Surveillance data can provide benchmarks to assess the intervention measures for monitoring public health and setting health policies. However, the current technological advances are extensively used to extract health information from big data via social media content on the Internet (Richter & Khoshgoftaar, 2020). Hence, the relationship between social media and different health-related trends can be addressed in several ways. For instance, information collected from social media platforms has been used as an essential source for the existing outpatient, hospital, and laboratory-based systems (Aiello, Renson, & Zivich, 2020). Also, an example of the relation between social media and disease-related issues can be observed in the work of Harris et al. (2017) who used Twitter messages to identify tweets related to food poisoning. The authors used the Web-based Dashboard (HealthMap Foodborne Dashboard) to recognize and respond to tweets about food poisoning from St Louis City residents. Also, extracting disease symptoms from social media messages can contribute significantly to the development of disease surveillance systems (Wang et al., 2020). For example, Boit and El-Gayar (2020) extracted topical themes from Twitter data using Crimson Hexagon's ReadME algorithm to classify and analyze topical emergent themes as well as monitoring malaria trends. The authors found that extracting personal characteristics from social media users plays a significant role in understanding their health conditions, thus fostering the detection and prevention of disease outbreaks.

It is common that people's mood can vary widely over time. Weather was believed to have an influence on emotions (Kööts, Realo, & Allik, 2011). For example, there is a general belief among people that sunshine is associated with positive states of mind, whereas rain is associated with negative states of mind (Kööts et al., 2011). Also, it was believed that weather may contribute to the individual's mood state and behavior (Bujisic, Bogicevic, Parsa, Jovanovic, & Sukhu, 2019). From a physiological perspective, the relationship between weather and emotions has been explored. Schneider et al. (2008) found that Systolic blood pressure may decrease immediately on more humid days. In this sense, if temperature is low, blood pressure to some extent could increase due to the catecholamine secretion which may lead to vasoconstriction (Jehn, Appel, Sacks, & Miller, 2002).

The influence of certain atmospheric factors on people's emotional state is a topical research interest in environmental psychology that received a great deal of attention recently. Different weather patterns led to changes in people's emotional states (Spasova, 2012). Much of the prior work on climate/weather has studied its relation to certain psychological and behavioral responses, including increases in the occurrence of suicide during the early summer. This has motivated Petridou, Papadopoulos, Frangakis, Skalkidou, and Trichopoulos (2002) to explore whether exposure to sunshine can trigger suicidal behavior among adults. The authors concluded that the sunshine could have a triggered effect on suicide rates. They reported a remarkably consistent pattern of seasonality with a peak incidence around June in the northern hemisphere and December in the southern hemisphere. It is still argued that the sunlight can impose some effect on individuals' feelings based on changes in brain serotonergic activity (i.e., the rate of production of serotonin by the brain rises rapidly with increased luminosity). Makris et al. (2016) found that there is evidence that tends to favor an increase in suicide risk among both men and women with increasing sunshine hours. However, previous studies that tend to focus on psychological outcomes showed that air pollution exposure might be a risk factor for depression (Kim et al., 2016; Shin, Park, & Choi, 2018). This is evident from the work of Tian, Zhang, and Zhang (2018) who also stated that the mood-depressing effect of air pollution is mostly associated with various behavioral changes. Power et al. (2015) pointed out that one important environmental exposure which may be strongly associated with anxiety is air pollution.

The rise and fall in temperature through the year is believed to have a notable impact on human emotions in which an increase in temperature was found to be linked to both negative and positive feelings (Kööts et al. 2011). In this context, the researchers used an experience sampling method to assess online users' emotional changes where they found that warmer temperatures increase the frequency of both positive and negative emotions, whereas higher humidity has the reverse effect. When comparing mood experienced at average daily temperatures (10–16°C or 50–60°F) with temperatures above 70°F (21°C), Noelke et al. (2016) observed that high temperatures reduce positive emotions (e.g., joy/happiness), increase negative emotions such as stress or anger, and increase fatigue – feeling tired, having low energy. In addition, some researchers have examined the effects of season and weather on emotion and travel satisfaction. For example, Ettema, Friman, Olsson, and Gärling (2017) found that positive and negative emotion is known to be influenced by certain atmospheric factors. This was supported by the work of Kämpfer and Mutz (2013) who found an influence of sunshine on life satisfaction and argued that this effect is mediated by mood (defined as frequencies of experienced positive vs. negative emotions).

Based on these observations, it can be anticipated that factors like sunlight, temperature, and air pollution are the major mood-influential variables that may potentially lead to a change in human behavior. It can also be noted that there is a large body of literature that encompassing various physiological and psychological disciplines which have sought to understand and explain the relationship between people and their interaction with the environment. Those studies showed a possible relationship between certain climatic factors and people emotion. Nevertheless, in practice, to gain a clear understanding about viral diseases in the population, and to evaluate the health condition of society, surveillance systems are required for monitoring the general health of the population.

Surveillance systems have been recognized as the cornerstone of public health efforts that were established to tackle general public health needs such as early identification of epidemics and control of infectious diseases (Zadeh et al., 2019). Traditional surveillance systems primarily relied on data reported by medical institutions which requires a long time to process and learn, thus increasing the uncertainty of the decision-making process (Fang & Chen, 2016). Besides that, according to Chen, Brown, Hu, King, and Chen (2011), the failure of public health agencies to provide effective measures for detecting and managing viral diseases is typically underpinned by the limited reach of traditional surveillance systems to achieve regional or global efficiencies. In light of this, the current surveillance systems are required to tackle the recent explosive growth in digital data and its growing role in real-time decision support systems, so as to monitor the health status of a population at a given time (Zadeh et al., 2019). For this reason, social media platforms can provide a great deal of valuable information about various health-related topics (Paul et al., 2016). Online trends and discussions being carried out between users can provide a crucial clue for understanding the epidemiologic intelligence that health officials and clinical administrators use to properly make actions and come up with efficient decisions (Santillana et al., 2016). This involves recognizing the relevant symptoms of a disease shared by members of a given population in the social media network. In the literature, several studies have been found to extensively rely on social media data to detect different health-related trends using platforms such as Twitter. Twitter is a very popular platform for expressing opinions and interacting with people in the online community (Barnaghi, Ghaffari, & Breslin, 2016). An example of such studies was the one conducted by Kitagawa, Komachi, Aramaki, Okazaki, and Ishikawa (2015) who proposed the use of modality features to improve influenza detection by using Twitter messages. Specifically, the authors annotated online tweets with a binary label (positive and negative). If a tweet is found to contain information about influenza, then the label will be positive, and vice versa. Another example can be found in the work of Clark et al. (2018) who identified tweets related to breast cancer using supervised learning algorithms with natural language processing. They also examined the tweet content with a hedonometric sentiment analysis method to extract emotion-related topics. Karami, Dahl, Turner-McGrievy, Kharrazi, and Shaw (2018) analyzed the features of the public's opinions related to diabetes, diet, exercise and obesity (DDEO) purely based on Twitter data. For this purpose, a multi-component semantic and linguistic framework was developed to discover topics of interest about DDEO in which the topics that they predicted were analyzed using a latent Dirichlet allocation algorithm.

As such, this study investigated the possibility of detecting critical diseases through the use of social media. We proposed a novel mechanism for detecting migraine disease based on the interrelation between users' emotion and climatic factors from tweets. The following section explains the study procedure in detail.

3. Procedure

In this study, data collection, data pre-processing, and feature extraction phases were performed (see Figure 1) to develop a predictive model for migraine disease and evaluate its performance in a real-time world. The following sections describe these phases in detail.

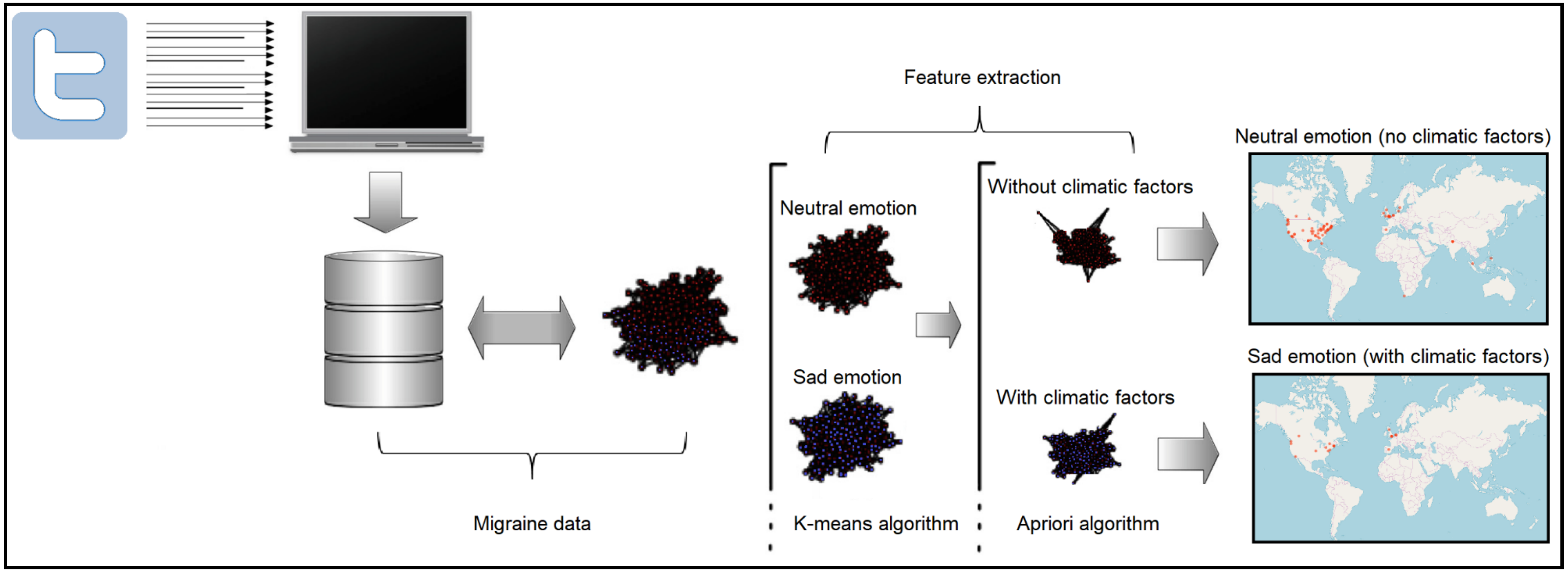

Figure 1: The study phases

3.1 Data collection

Since this work aims to propose a new mechanism for detecting early diseases in microblogs based on certain emotional and climatic factors, the geo-spatial location for each tweet posted by a user was assumed to offer an in-depth understanding of various emotional information (e.g., polarity). Here, the proposed mechanism depends on the existing associations between weather and disease forecasting. A total of 238,506,796 English tweets were collected between September 17, 2018 and November 1, 2018 using Twitter's API. For data collection, we utilized the following search keywords: 'migraine sickness', 'migraine disease', and 'migraine and weather'. In addition, the acquired tweets were labeled according to their geo-spatial location information (longitude and latitude coordinates).

3.2 Data pre-processing

Several pre-processing steps were applied to the tweets before performing the classification process. At this stage, re-tweets were removed during the data pre-processing stage. However, since most of the common machine learning algorithms cannot directly process the data in its original form, we converted the collected tweets into a more manageable representation. For this purpose, the bag-of-words (or set-of-words) model was used. The "Tokenization" technique was also applied to extract the main features and assign a specific weight to each of them using a complex weighting scheme (e.g., the classical term frequency–inverse document frequency (TF-IDF)). The TF-IDF weight was calculated as follows (Bhattacharjee, Srijith, & Desarkar, 2019):

$$TF - IDF_{std}(t) = \ tf_d^t \times \log\frac{N}{df^t}$$

Where the tf_d^t denotes the number of times the term *t* occurred in document *d*. And *N* denotes the total number of documents in the corpus. df^t denotes the number of documents in which the term *t* occurred. After the tokenization process, all the extracted words for each tweet were converted to a lowercase format, followed by normalizing the length of each tweet using the L2 norm. All informal messaging convention (e.g., #hashtag and links) were removed from all tweets and the generated dictionary. In addition, the stop word list was used to characterize word-related information in a specific tweet. Based on these measures, a total of 237,098,462 tweets were used in this study. To clearly understand users' tweet statements, the part-of-speech (POS) tagging technique was applied using a Stanford parser to split the tweets into words and combine related tags into noun, verb, adjective and adverb.

3.3 Feature extraction

K-means clustering algorithm was applied to cluster the processed data into independent groups (clusters) due to its flexibility, popularity, and simplicity (Patowary, Sarmah, & Bhattacharyya, 2020). It places the instances of a particular set of tweets in the relevant cluster based on the similarities and differences in each cluster (Sarsam & Al-Samarraie, 2018; Sarsam, Al-Samarraie, & Omar, 2019). In this study, the K-means algorithm produced two different clusters. We asked three experts to categorize these clusters in order to determine the overall polarity type based on a subset of features for each cluster. After that, the two clusters (each with 118,549,231 tweets) were labelled as 'Sad' and 'Neutral'. Those two labels were used to specify the class attribute. Then, the association rules mining approach (using Apriori algorithm) was applied to extract the features of a disease that are associated with certain climatic factors. Apriori algorithm is commonly used to provide meaningful insights as well as inferences from within a set of items or selections (Kamsu-Foguem, Rigal, & Mauget, 2013; Singh, Ram, & Sodhi, 2013). This algorithm creates association rules essential for providing minimum support and confidence thresholds. It aims to identify frequent item sets, building sequentially longer candidate item sets from shorter ones. We configured the Apriori algorithm using the Waikato Environment for Knowledge Analysis (Weka). More precisely, we set the delta value to 0.05 as the value of lower bound for minimum support. In addition, the minimum metric score was set to 0.9. The result of the Apriori algorithm is explained in section 4.1.

3.4 Classification task

In general, classification algorithms use the information provided from a given input to predict a particular outcome (Choubey, Kumar, Tripathi, & Kumar, 2020). After extracting and understanding the nature of the associated features in each of the two clusters (sad and neutral), four different classification algorithms were applied to detect migraine ('Yes' was used as a labeling target for migraine-related tweets and 'No' for non-migraine tweets) from the collected tweets (each tweet was associated with its longitude and latitude coordinates). These tweets were then fed into four classifiers (Sequential Minimal Optimization (SMO) or SVM (Platt, 1998) using the Radial Basis Function Kernel (RBF) (Ring & Eskofier, 2016), decision tree (J48) (Salzberg, 1994), 1-rule classifier (OneR) (Holte, 1993), and instance-based learner (IBk) (Aha, Kibler, & Albert, 1991)) to compare and identify the best performing one. These classifiers were executed from the Weka environment (Waikato Environment for Knowledge Analysis). The stratified tenfold cross-validation method was used to evaluate the learning process of each classifier. In addition, several evaluation metrics were used to examine the classification performance (see section 4.1). The result showed that SMO classifier had the best performance. Hence, it was examined further on a real-time world map (see section 4.2).

4. Result

This section is divided into three sub-sections: the first section explains the result of the Apriori algorithm; the second section demonstrates the result of the disease classification process; and the third section evaluates the proposed approach on a real-time world map.

4.1 Result of the Apriori algorithm

Figure 2 shows the Apriori results for each cluster. Figure 2(a) presents the resultant rules and their support and confidence for sad emotions. Our results showed that the confidence level for sad emotions was ranging between 80% and 100%. In addition, Figure 2(b) presents the resultant rules and their support and confidence for neutral emotions where the confidence level was found to be between 20% and 75%. The findings from the Apriori algorithm were that climatic factors and migraine symptoms are significantly associated with sad statements.

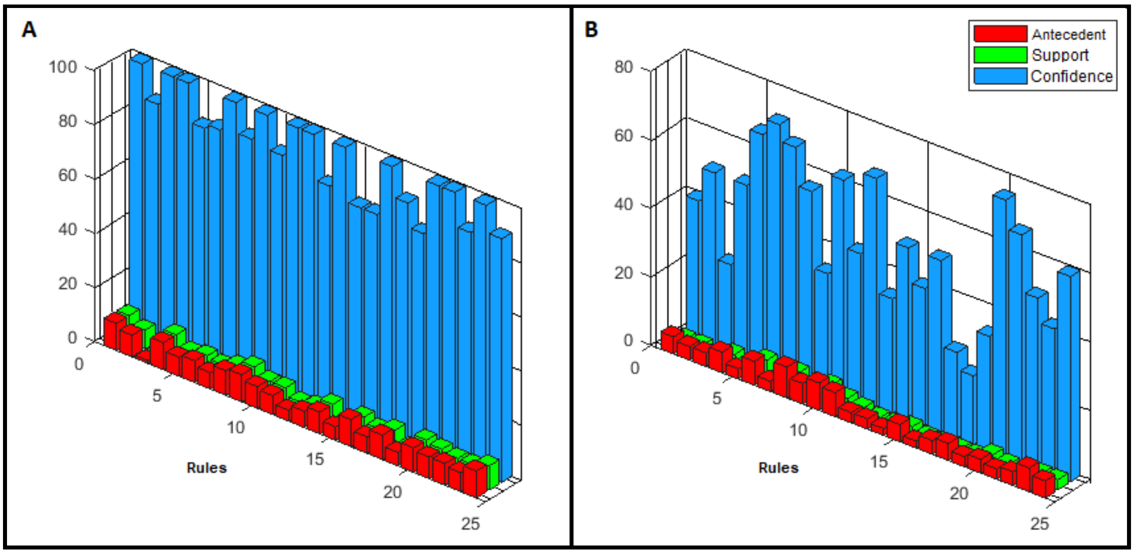

Figure 2: Apriori results for each cluster (a) the climatic factors that are associated with sad emotions; (b) the features associated with neutral emotions (no climatic factors)

Figure 3 shows a serial of networks of connected items and rules. These networks focus on how the established rules can be used to represent individual items and sets. The size of each circle represents the support value, while the color represents the level of confidence of the association rules. In Figure 3(a), we can observe that the rules for sad emotions can be established by using a combination of climatic factors and migraine disease symptoms. In addition, the figure shows that certain climatic factors, such as humidity, air pressure, cold, dry, and wind, were highly associated with the detection of migraine. For example, migraine symptoms due to these climatic changes were mostly associated with vomiting, nausea, sneeze, blurred vision, sensitivity towards light and sound. In contrast, Figure 3(b) shows less potential for the neutral emotions in establishing networked relationships between disease symptoms (migraine) and climatic factors.

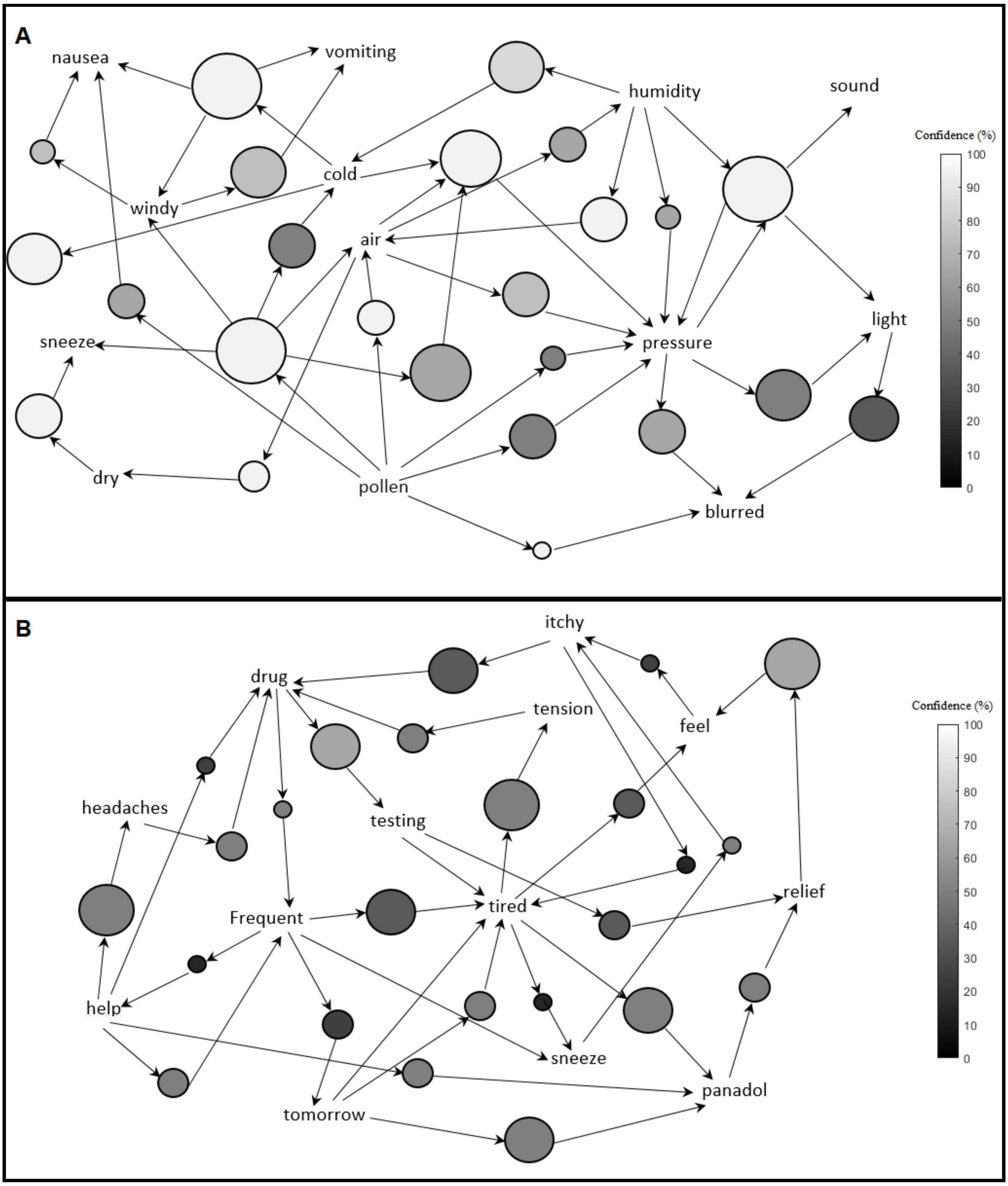

Figure 3: Association rules for sad and neutral emotions

4.2 Results of classification task

In order to predict the migraine symptoms from the 237,098,462 tweets, a stratified tenfold cross-validation method was utilized to evaluate the learning process of SMO, J48, OneR, and IBk. The data were divided randomly into 10 parts, in which the class was represented in approximately the same proportions as in the full dataset. Each part is held out in turn and the learning scheme trained on the remaining nine-tenths (training set which is about 213,388,616). The error rate was calculated for each algorithm based on the holdout set. The learning procedure was executed for 10 times on different training sets. Then, the 10 error estimates were averaged to yield an overall error estimate. Accuracy (Acc.), Kappa statistic (Ks.), Root Mean Squared Error (RMSE), and Receiver Operating Characteristic (ROC) were used to assess the performance of these classifiers. The result is summarized in Table 1. From the table, it can be said that the SMO classifier had the highest ROC value (80%) (see Figure 4a and b), followed by J48 (64%), OneR (56%), and IBk (46%), respectively. The SMO classifier had the highest accuracy value (95.53%) and the highest kappa statistic value (47.58%), followed by J48 (**Acc. =** 61.49%, **Ks. =** 23.42%), OneR (**Acc. =** 55.27%, **Ks. =** 12.47%), and IBk

(**Acc. =** 50.93 %, **Ks. =** 7.34%), respectively. Furthermore, SMO in this study produced the lowest number of classification error (42.97%), followed by J48 (52.32%), OneR (66.87%), and IBk (69.92%) (see Figure 4c). Based on these results, it can be concluded that the SMO algorithm can be used efficiently to detect a specific disease using the associated climatic and emotional features in a microblog platform.

Table 1: Performance results of the four classifiers

No.	Classifier	Acc. (%)	Ks. (%)	RMSE (%)
1.	*SMO*	95.53	47.58	42.97
2.	*J48*	61.49	23.42	52.32
3.	*OneR*	55.27	12.47	66.87
4.	*IBk*	50.93	7.34	69.92

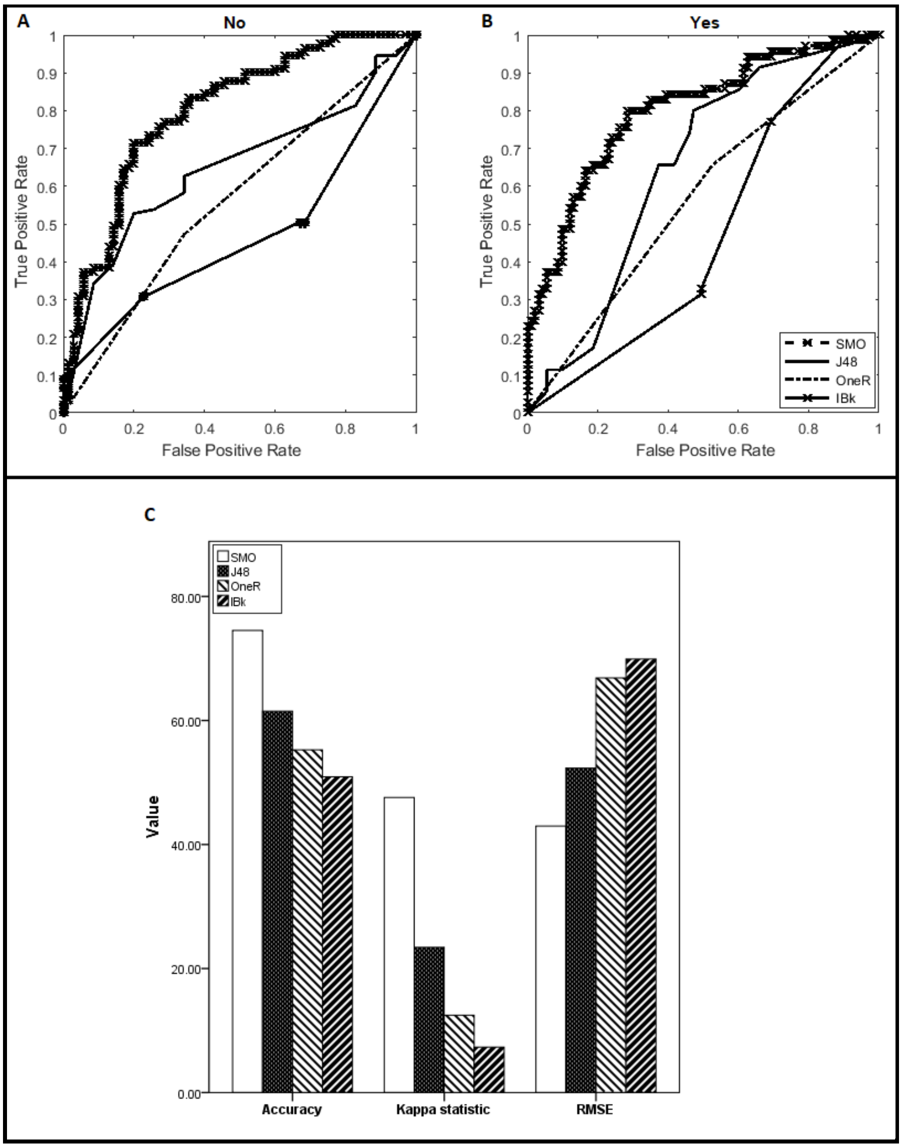

Figure 4: Evaluation metrics of the four algorithms

4.3 Evaluation

After selecting the best predictive model, the geographic distribution of migraine symptoms was used for evaluation purposes. We used the original dataset to locate migraine symptoms on a real-time world map. Since the processed dataset carries all the relevant information about the users' posts, we used certain climatic factors of longitude and latitude coordinates which are found in the sentiments of emotional tweets. Figure 5 shows the distribution of tweets related to the emotional features of the disease.

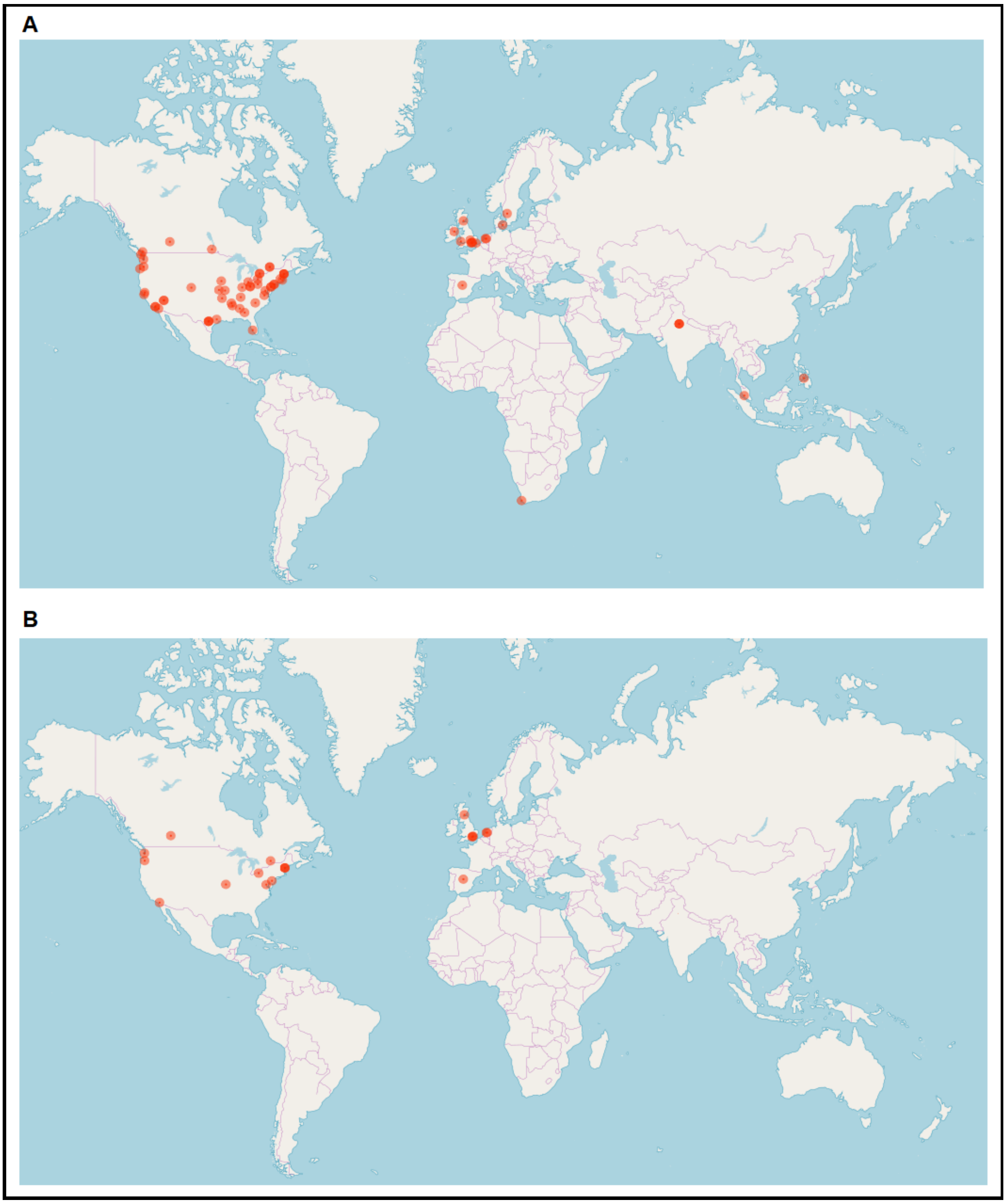

Figure 5: A real-time biosurveillance detection of migraine disease

5. Discussion

In this study, sad emotions were found to play a key role in disease recognition on a real-world map. This type of emotions can be highly interrelated with the early-stage of a disease. In other words, patients with migraine are likely to express sad emotions in their published tweets. Such findings seem to be reasonable because migraine has a strong impact on people's mood by making them more irritable and aggressive (Banciu & Bouleanu, 2018). In addition, migraine is commonly associated with other anxiety and mood disorders that show substantial physical symptoms (Peres, Mercante, Tobo, Kamei, & Bigal, 2017). For example, irritability, concentration problems, and agitation are commonly used to characterize anxiety in individuals affected by migraine, while fatigue, concentration, sleep and appetite are used to characterize depression (Cioffi et al., 2017). Our findings support the work of Mannix et al. (2016) who revealed that emotional responses, such as frustration, experiencing depressed mood, and anxiety, might be associated with migraine. In addition, chronic stress is considered one of the common risk factors for both depression and chronic migraine (Lantéri-Minet, Duru, Mudge, & Cottrell, 2011). Most migraine symptoms that appear in users' tweets can be effectively used in the sentiment analysis. Therefore, it is assumed that using tweets with sad emotions can potentially help to explain individuals' health and propensity for disease.

In this study, the text mining result of sad tweets showed how certain climatic factors can be associated with migraine. We found that some climatic factors such as humidity, air pressure, cold, dry, and wind, were associated with the main symptoms of migraine (vomiting, nausea, sneeze, blurred vision, sensitivity towards light and sound). In addition, a low-pressure environment may increase local vasoconstriction and ischaemia. In line with that, cold environmental conditions could favor the release of peripheral mediators, inducing vasoconstriction, local ischaemia and promoting the release of algesic substances by non-neuronal cells. This would clarify the existing associations between low temperature and incidents of migraine headaches in individuals (Yang, Fuh, Huang, Shia, & Wang, 2015). Meanwhile, higher levels of fine particulate matter ($PM_{2.5}$), nitrogen dioxide (NO_2), sulfur dioxide (SO_2), ozone (O_3), and carbon monoxide (CO) are associated with a higher number of hospitalization episodes for migraine headache (Dales, Cakmak, & Vidal, 2009). Previous studies (e.g., Buse et al., 2013; Capi et al., 2018; Schroeder et al., 2018) have discussed the role of symptoms like nausea, visual aura, blurred vision, photophobia, phonophobia, and vomiting, in increasing migraine severity. These symptoms were highly associated with weather-related factors, which could explain the associations between migraine disease symptoms and climatic factors. Such rules are helpful for the recognition and dissemination of the disease geographically. This also could justify the high classification result in this study. This is supported by Vioulès, Moulahi, Azé, and Bringay (2018) who proposed a new method to quantify suicide-warning signs and predict suicide-related posts.

This study introduces the potential of supervised and unsupervised machine learning methods to enhance the applications of biosurveillance within social media platforms. This approach can be applied to inform the public when a certain sickness breaks out. The proposed mechanism facilitates the identification of emerging public health conditions rather than a specific illness that can be embedded in twitter messages. This mechanism has the potential to characterize the early-stage disease and extract the possible relationships between the identified symptoms and environmental factors. In addition, the extracted relationships can be used to identify simultaneous diseases existing in the same location at the same time, thus facilitating timely actions by government and health organizations who want to decrease the number and cost of illnesses and deaths in society. Furthermore, current work may also inspire researchers to carry out more extensive research on this topic, which includes the use of textual analytics for analyzing disease-related tweets. The use of latent sentiments in the textual content of microblogs can speed up the process of locating a sickness, which could facilitate fast preparation for epidemics that is extremely useful for both patients and doctors.

6. Implications

The proposed mechanism contributes to the disease detection process and improves existing surveillance systems in resource-limited settings. From the health perspective, our mechanism has the potential to characterize disease by looking at the relationship between disease symptoms and relevant environmental factors. In addition, the proposed method can be an effective alternative for managing real-world data where several diseases are disseminated. Thus, finding the association between certain climatic factors and a disease will contribute significantly to public health initiatives. This can be understood by means of facilitating timely actions by government and health organizations to prevent illnesses and deaths in a society. Technically, the proposed approach contributes to the current disease recognition methods by discovering the micro-elements associated with the sickness and accurately predicts simultaneous illnesses within a geographical area. Furthermore, the proposed method can handle critical cases related to the detection of multiple diseases in a population.

7. Limitation and future works

Our study faces some limitations that should be acknowledged. The feasibility of the proposed mechanism was focused only on a specific disease. Another limitation was on the use of "crawling" software for data collection from twitter, which used specific keywords to filter users' posts. The prediction results were based on K-means clustering technique that was invoked to obtain users' emotional data based on two types of emotions (sad and neutral). Based on these limitations, researchers may consider using other disease-related information. In addition, annotated corpus, a corpus, which is mainly based on medical-related resources, can be used in the future to label the extracted tweets. Future studies may also consider the potential of adopting users' psychological factors and explore their effects on the performance of the predictive model. Other clustering methods can be also used to group users' tweets that share similar attributes during the early stage of a disease or infection. This includes classifying other types of emotions such as boredom, happiness, and cold anger which may facilitate the disease classification procedure.

8. Conclusion

In this work, we proposed a novel real biosurveillance mechanism to detect early-stage disease from Twitter. A K-means clustering method was used to extract the emotional characteristics (sad and neutral features) that were deduced from users' tweets. An Apriori algorithm was then applied to extract the migraine features that are associated with specific climatic factors in each of these emotions. The result of the Apriori algorithm revealed that sad emotions and climatic factors were sufficient for the detection of migraine. For the prediction purposes, the SMO classifier was applied by using migraine-related emotions in conjunction with certain climatic factors related to disease symptoms. The SMO model was then examined geographically on a real-time world map using geo-spatial location information (e.g., longitude and latitude coordinates). Our results showed the potential for sad emotions (e.g., frustration, experiencing depressed mood, and anxiety) in detecting a disease. It was also found that certain climatic factors (e.g., humidity, air pressure, cold, dry, and wind) can be used to establish a benchmark for identifying migraine symptoms such as vomiting, nausea, sneeze, blurred vision, sensitivity towards light and sound. In sum, the process of disease recognition in microblogs can be established by understanding the interconnection between certain emotional and climatic factors. The proposed approach serves as a surveillance system that provides an opportunity for decision-makers to significantly enhance public health.

References

Aha, D. W., Kibler, D., & Albert, M. K. (1991). Instance-based learning algorithms. *Machine Learning, 6*(1), 37-66.

Aiello, A. E., Renson, A., & Zivich, P. N. (2020). Social media–and internet-based disease surveillance for public health. *Annual Review of Public Health, 41*(1). doi:10.1146/annurev-publhealth-040119-094402

Aramaki, E., Maskawa, S., & Morita, M. (2011). Twitter catches the flu: Detecting influenza epidemics using twitter. In *Proceedings of the conference on empirical methods in natural language processing* (pp. 1568-1576). Association for Computational Linguistics.

Atefeh, F., & Khreich, W. (2015). A survey of techniques for event detection in twitter. *Computational Intelligence, 31*(1), 132-164.

Banciu, A., & Bouleanu, E. L. (2018). The experience of persons living with migraine. *Acta Medica Transilvanica, 23*(2), 27-29.

Barnaghi, P., Ghaffari, P., & Breslin, J. G. (2016). Opinion mining and sentiment polarity on twitter and correlation between events and sentiment. In *2016 IEEE Second International Conference on Big Data Computing Service and Applications (BigDataService)* (pp. 52-57). IEEE.

Bhattacharjee, U., Srijith, P., & Desarkar, M. S. (2019). Term specific tf-idf boosting for detection of rumours in social networks. In *2019 11th International Conference on Communication Systems & Networks (COMSNETS)* (pp. 726-731). IEEE.

Boit, J., & El-Gayar, O. (2020). Topical mining of malaria using social media. A text mining approach. In *Proceedings of the 53rd Hawaii International Conference on System Sciences.* doi: 10.24251/HICSS.2020.466

Broniatowski, D. A., Paul, M. J., & Dredze, M. (2013). National and local influenza surveillance through twitter: An analysis of the 2012-2013 influenza epidemic. *PloS One, 8*(12), e83672.

Bujisic, M., Bogicevic, V., Parsa, H., Jovanovic, V., & Sukhu, A. (2019). It's raining complaints! How weather factors drive consumer comments and word-of-mouth. *Journal of Hospitality & Tourism Research, 43*(5), 656-681.

Burton, S. H., Tanner, K. W., Giraud-Carrier, C. G., West, J. H., & Barnes, M. D. (2012). "Right time, right place" health communication on twitter: Value and accuracy of location information. *Journal of Medical Internet Research, 14*(6), 34-52.

Buse, D. C., Loder, E. W., Gorman, J. A., Stewart, W. F., Reed, M. L., Fanning, K. M., . . . Lipton, R. B. (2013). Sex differences in the prevalence, symptoms, and associated features of migraine, probable migraine and other severe h eadache: Results of the a merican migraine prevalence and prevention (ampp) study. *Headache: The Journal of Head and Face Pain, 53*(8), 1278-1299.

Byrd, K., Mansurov, A., & Baysal, O. (2016). Mining twitter data for influenza detection and surveillance. In *Proceedings of the International Workshop on Software Engineering in Healthcare Systems* (pp. 43-49).

Capi, M., Gentile, G., Lionetto, L., Salerno, G., Cipolla, F., Curto, M., . . . Martelletti, P. (2018). Pharmacogenetic considerations for migraine therapies. *Expert Opinion on Drug Metabolism & Toxicology, 14*(11), 1161-1167.

Chai, N. C., Rosenberg, J. D., & Peterlin, B. L. (2012). The epidemiology and comorbidities of migraine and tension-type headache. *Techniques in Regional Anesthesia and Pain Management, 16*(1), 4-13.

Chen, Y.-D., Brown, S. A., Hu, P. J.-H., King, C.-C., & Chen, H. (2011). Managing emerging infectious diseases with information systems: Reconceptualizing outbreak management through the lens of loose coupling. *Information Systems Research, 22*(3), 447-468.

Choubey, D. K., Kumar, P., Tripathi, S., & Kumar, S. (2020). Performance evaluation of classification methods with pca and pso for diabetes. *Network Modeling Analysis in Health Informatics and Bioinformatics, 9*(1), 5-19.

Cioffi, I., Farella, M., Chiodini, P., Ammendola, L., Capuozzo, R., Klain, C., . . . Michelotti, A. (2017). Effect of weather on temporal pain patterns in patients with temporomandibular disorders and migraine. *Journal of Oral Rehabilitation, 44*(5), 333-339.

Clark, E. M., James, T., Jones, C. A., Alapati, A., Ukandu, P., Danforth, C. M., & Dodds, P. S. (2018). A sentiment analysis of breast cancer treatment experiences and healthcare perceptions across twitter. *arXiv preprint arXiv:1805.09959*.

Culotta, A. (2010). Towards detecting influenza epidemics by analyzing twitter messages. In *Proceedings of the first workshop on social media analytics* (pp. 115-122).

Dales, R. E., Cakmak, S., & Vidal, C. B. (2009). Air pollution and hospitalization for headache in chile. *American Journal of Epidemiology, 170*(8), 1057-1066.

Erraguntla, M., Zapletal, J., & Lawley, M. (2019). Framework for infectious disease analysis: A comprehensive and integrative multi-modeling approach to disease prediction and management. *Health Informatics Journal, 25*(4), 1170-1187.

Ettema, D., Friman, M., Olsson, L. E., & Gärling, T. (2017). Season and weather effects on travel-related mood and travel satisfaction. *Frontiers in Psychology, 8*, 140-163.

Fang, Z.-H., & Chen, C. C. (2016). A novel trend surveillance system using the information from web search engines. *Decision Support Systems, 88*, 85-97.

Harris, J. K., Hawkins, J. B., Nguyen, L., Nsoesie, E. O., Tuli, G., Mansour, R., & Brownstein, J. S. (2017). Research brief report: Using twitter to identify and respond to food poisoning: The food safety stl project. *Journal of Public Health Management and Practice, 23*(6), 577-592.

Hartley, D. M., Nelson, N. P., Arthur, R., Barboza, P., Collier, N., Lightfoot, N., . . . Madoff, L. (2013). An overview of internet biosurveillance. *Clinical Microbiology and Infection, 19*(11), 1006-1013.

Hoffmann, J., Schirra, T., Lo, H., Neeb, L., Reuter, U., & Martus, P. (2015). The influence of weather on migraine–are migraine attacks predictable? *Annals of Clinical and Translational Neurology, 2*(1), 22-28.

Holte, R. C. (1993). Very simple classification rules perform well on most commonly used datasets. *Machine Learning, 11*(1), 63-90.

Huang, Z., Dong, W., & Duan, H. (2015). A probabilistic topic model for clinical risk stratification from electronic health records. *Journal of Biomedical Informatics, 58*, 28-36.

Jehn, M., Appel, L. J., Sacks, F. M., & Miller, E. R. (2002). The effect of ambient temperature and barometric pressure on ambulatory blood pressure variability. *American Journal of Hypertension, 15*(11), 941-945.

Jordan, S. E., Hovet, S. E., Fung, I. C.-H., Liang, H., Fu, K.-W., & Tse, Z. T. H. (2019). Using twitter for public health surveillance from monitoring and prediction to public response. *Data, 4*(1), 6-16.

Joshi, A., Sparks, R., McHugh, J., Karimi, S., Paris, C., & MacIntyre, C. R. (2020). Harnessing tweets for early detection of an acute disease event. *Epidemiology, 31*(1), 90-97.

Kämpfer, S., & Mutz, M. (2013). On the sunny side of life: Sunshine effects on life satisfaction. *Social Indicators Research, 110*(2), 579-595.

Kamsu-Foguem, B., Rigal, F., & Mauget, F. (2013). Mining association rules for the quality improvement of the production process. *Expert Systems with Applications, 40*(4), 1034-1045.

Karami, A., Dahl, A. A., Turner-McGrievy, G., Kharrazi, H., & Shaw Jr, G. (2018). Characterizing diabetes, diet, exercise, and obesity comments on twitter. *International Journal of Information Management, 38*(1), 1-6.

Kim, K.-N., Lim, Y.-H., Bae, H. J., Kim, M., Jung, K., & Hong, Y.-C. (2016). Long-term fine particulate matter exposure and major depressive disorder in a community-based urban cohort. *Environmental Health Perspectives, 124*(10), 1547-1553.

Kitagawa, Y., Komachi, M., Aramaki, E., Okazaki, N., & Ishikawa, H. (2015). Disease event detection based on deep modality analysis. In *Proceedings of the ACL-IJCNLP 2015 Student Research Workshop* (pp. 28-34).

Kööts, L., Realo, A., & Allik, J. (2011). The influence of the weather on affective experience. *Journal of Individual Differences 32*, 74-84

Lamb, A., Paul, M. J., & Dredze, M. (2013). Separating fact from fear: Tracking flu infections on twitter. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 789-795).

Lantéri-Minet, M., Duru, G., Mudge, M., & Cottrell, S. (2011). Quality of life impairment, disability and economic burden associated with chronic daily headache, focusing on chronic migraine with or without medication overuse: A systematic review. *Cephalalgia, 31*(7), 837-850.

Lim, S., Tucker, C. S., & Kumara, S. (2017). An unsupervised machine learning model for discovering latent infectious diseases using social media data. *Journal of Biomedical Informatics, 66*, 82-94.

Liu, F., Weng, C., & Yu, H. (2019). Advancing clinical research through natural language processing on electronic health records: Traditional machine learning meets deep learning. In *Clinical Research Informatics* (pp. 357-378). Springer, Cham.

Makris, G. D., Reutfors, J., Larsson, R., Isacsson, G., Ösby, U., Ekbom, A., . . . Papadopoulos, F. C. (2016). Serotonergic medication enhances the association between suicide and sunshine. *Journal of Affective Disorders, 189*, 276-281.

Mannix, S., Skalicky, A., Buse, D. C., Desai, P., Sapra, S., Ortmeier, B., . . . Hareendran, A. (2016). Measuring the impact of migraine for evaluating outcomes of preventive treatments for migraine headaches. *Health and Quality of Life Outcomes, 14*(1), 143-168.

Molaei, S., Khansari, M., Veisi, H., & Salehi, M. (2019). Predicting the spread of influenza epidemics by analyzing twitter messages. *Health and Technology*, *1*, 1-16.

Naslund, J. A., Aschbrenner, K. A., McHugo, G. J., Unützer, J., Marsch, L. A., & Bartels, S. J. (2019). Exploring opportunities to support mental health care using social media: A survey of social media users with mental illness. *Early Intervention in Psychiatry, 13*(3), 405-413.

Nejad, M. Y., Delghandi, M. S., Bali, A. O., & Hosseinzadeh, M. (2020). Using twitter to raise the profile of childhood cancer awareness month. *Network Modeling Analysis in Health Informatics and Bioinformatics, 9*(1), 3-18.

Noelke, C., McGovern, M., Corsi, D. J., Jimenez, M. P., Stern, A., Wing, I. S., & Berkman, L. (2016). Increasing ambient temperature reduces emotional well-being. *Environmental Research, 151*, 124-129.

Patowary, P., Sarmah, R., & Bhattacharyya, D. K. (2020). Developing an effective biclustering technique using an enhanced proximity measure. *Network Modeling Analysis in Health Informatics and Bioinformatics, 9*(1), 1-17.

Paul, M. J., Sarker, A., Brownstein, J. S., Nikfarjam, A., Scotch, M., Smith, K. L., & Gonzalez, G. (2016). Social media mining for public health monitoring and surveillance. In *Biocomputing 2016: Proceedings of the Pacific symposium* (pp. 468-479).

Peres, M. F. P., Mercante, J. P., Tobo, P. R., Kamei, H., & Bigal, M. E. (2017). Anxiety and depression symptoms and migraine: A symptom-based approach research. *The Journal of Headache and Pain, 18*(1), 37-52.

Petridou, E., Papadopoulos, F. C., Frangakis, C. E., Skalkidou, A., & Trichopoulos, D. (2002). A role of sunshine in the triggering of suicide. *Epidemiology, 13*(1), 106-109.

Plattt, J. (1998). Fast training of support vector machines using sequential minimal optimization. *Advances In Kernel Methods-Support Vector Learning, In Schlkopf B., Burges, C., and Smola, A., editors, Advances in Kernel Methods-Support Vector Learning*. MIT Press.

Poole, S., Schroeder, L. F., & Shah, N. (2016). An unsupervised learning method to identify reference intervals from a clinical database. *Journal of Biomedical Informatics, 59*, 276-284.

Power, M. C., Kioumourtzoglou, M.-A., Hart, J. E., Okereke, O. I., Laden, F., & Weisskopf, M. G. (2015). The relation between past exposure to fine particulate air pollution and prevalent anxiety: Observational cohort study. Br Med J*, 350*, h1111. doi.org/10.1136/bmj.h1111

Priyadarshi, A., & Saha, S. K. (2020). Web information extraction for finding remedy based on a patient-authored text: A study on homeopathy. *Network Modeling Analysis in Health Informatics and Bioinformatics, 9*(1), 1-12.

Ravat, S., Chaudhari, S., & Chafekar, N. (2019). Clinical profile of primary headaches and awareness of trigger factors in migraine patients. *MVP Journal of Medical Sciences, 5*(2), 145-150.

Richter, A. N., & Khoshgoftaar, T. M. (2020). Sample size determination for biomedical big data with limited labels. *Network Modeling Analysis in Health Informatics and Bioinformatics, 9*(1), 12.

Ring, M., & Eskofier, B. M. (2016). An approximation of the gaussian rbf kernel for efficient classification with svms. *Pattern Recognition Letters, 84*, 107-113.

Sabatovych, I. (2019a). Do social media create revolutions? Using twitter sentiment analysis for predicting the maidan revolution in ukraine. *Global Media and Communication, 15*(3), 275-283.

Sabatovych, I. (2019b). Use of sentiment analysis for predicting public opinion on referendum: A feasibility study. *The Reference Librarian, 60*(3), 202-211.
Salzberg, S. L. (1994). C4. 5: Programs for machine learning by j. Ross quinlan. Morgan kaufmann publishers, inc., 1993. *Machine Learning, 16*(3), 235-240.
Santillana, M., Nguyen, A., Louie, T., Zink, A., Gray, J., Sung, I., & Brownstein, J. S. (2016). Cloud-based electronic health records for real-time, region-specific influenza surveillance. *Scientific reports, 6*, 25732.
Sarsam, S. M., & Al-Samarraie, H. (2018). A first look at the effectiveness of personality dimensions in promoting users' satisfaction with the system. *SAGE Open, 8*(2), 2158244018769125.
Sarsam, S. M., Al-Samarraie, H., & Omar, B. (2019). Geo-spatial-based emotions: A mechanism for event detection in microblogs. In *Proceedings of the 2019 8th International Conference on Software and Computer Applications* (pp. 1-5).
Savini, L., Tora, S., Di Lorenzo, A., Cioci, D., Monaco, F., Polci, A., . . . Conte, A. (2018). A web geographic information system to share data and explorative analysis tools: The application to west nile disease in the mediterranean basin. *PloS One, 13*(6), e0196429.
Schneider, A., Schuh, A., Maetzel, F.-K., Rückerl, R., Breitner, S., & Peters, A. (2008). Weather-induced ischemia and arrhythmia in patients undergoing cardiac rehabilitation: Another difference between men and women. *International Journal of Biometeorology, 52*(6), 535-547.
Schroeder, R. A., Brandes, J., Buse, D. C., Calhoun, A., Eikermann-Haerter, K., Golden, K., . . . Moriarty, M. (2018). Sex and gender differences in migraine—evaluating knowledge gaps. *Journal of Women's Health, 27*(8), 965-973.
Șerban, O., Thapen, N., Maginnis, B., Hankin, C., & Foot, V. (2019). Real-time processing of social media with sentinel: A syndromic surveillance system incorporating deep learning for health classification. *Information Processing & Management, 56*(3), 1166-1184.
Shin, J., Park, J. Y., & Choi, J. (2018). Long-term exposure to ambient air pollutants and mental health status: A nationwide population-based cross-sectional study. *PloS One, 13*(4), e0195607.
Singh, Ram, & Sodhi. (2013). Improving efficiency of apriori algorithm using transaction reduction. *International Journal of Scientific and Research Publications, 3*(1), 1-4.
Spasova, Z. (2012). The effect of weather and its changes on emotional state–individual characteristics that make us vulnerable. *Advances in Science and Research, 6*(1), 281-290.
Tian, J., Zhang, Y., & Zhang, C. (2018). Predicting consumer variety-seeking through weather data analytics. *Electronic Commerce Research and Applications, 28*, 194-207.
Vioulès, M. J., Moulahi, B., Azé, J., & Bringay, S. (2018). Detection of suicide-related posts in twitter data streams. *IBM Journal of Research and Development, 62*(1), 1-7.
Wang, Y., Xu, K., Kang, Y., Wang, H., Wang, F., & Avram, A. (2020). Regional influenza prediction with sampling twitter data and pde model. *International Journal of Environmental Research and Public Health, 17*(3), 678.
Weng, J., & Lee, B.-S. (2011). Event detection in twitter. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (*pp.401-408).
Yang, A. C., Fuh, J.-L., Huang, N. E., Shia, B.-C., & Wang, S.-J. (2015). Patients with migraine are right about their perception of temperature as a trigger: Time series analysis of headache diary data. *The Journal of Headache and Pain, 16*(1), 49-71.
Zadeh, A. H., Zolbanin, H. M., Sharda, R., & Delen, D. (2019). Social media for nowcasting flu activity: Spatio-temporal big data analysis. *Information Systems Frontiers*, *1,* 1-18.
Zaeem, R. N., Liau, D., & Barber, K. S. (2018). Predicting disease outbreaks using social media: Finding trustworthy users. In *Proceedings of the Future Technologies Conference* (pp. 369-384). Springer, Cham.
Zebenholzer, K., Rudel, E., Frantal, S., Brannath, W., Schmidt, K., Wöber-Bingöl, Ç., & Wöber, C. (2011). Migraine and weather: A prospective diary-based analysis. *Cephalalgia, 31*(4), 391-400.