Journal of Biomedical Informatics 66 (2017) 82–94

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

An unsupervised machine learning model for discovering latent infectious diseases using social media data

Sunghoon Lim [a], Conrad S. Tucker [b,a,*], Soundar Kumara [a]

[a] Department of Industrial and Manufacturing Engineering, The Pennsylvania State University, University Park, PA 16802, USA
[b] School of Engineering Design, Technology, and Professional Programs, The Pennsylvania State University, University Park, PA 16802, USA

ARTICLE INFO

Article history:
Received 4 August 2016
Revised 3 December 2016
Accepted 14 December 2016
Available online 26 December 2016

Keywords:
Latent infectious diseases
Information retrieval
Unsupervised machine learning
Sentiment analysis
Social media

ABSTRACT

Introduction: The authors of this work propose an unsupervised machine learning model that has the ability to identify real-world latent infectious diseases by mining social media data. In this study, a latent infectious disease is defined as a communicable disease that has not yet been formalized by national public health institutes and explicitly communicated to the general public. Most existing approaches to modeling infectious-disease-related knowledge discovery through social media networks are top-down approaches that are based on already known information, such as the names of diseases and their symptoms. In existing top-down approaches, necessary but unknown information, such as disease names and symptoms, is mostly unidentified in social media data until national public health institutes have formalized that disease. Most of the formalizing processes for latent infectious diseases are time consuming. Therefore, this study presents a bottom-up approach for latent infectious disease discovery in a given location without prior information, such as disease names and related symptoms.

Methods: Social media messages with user and temporal information are extracted during the data preprocessing stage. An unsupervised sentiment analysis model is then presented. Users' expressions about symptoms, body parts, and pain locations are also identified from social media data. Then, symptom weighting vectors for each individual and time period are created, based on their sentiment and social media expressions. Finally, latent-infectious-disease-related information is retrieved from individuals' symptom weighting vectors.

Datasets and results: *Twitter* data from August 2012 to May 2013 are used to validate this study. Real electronic medical records for 104 individuals, who were diagnosed with influenza in the same period, are used to serve as ground truth validation. The results are promising, with the highest precision, recall, and F_1 score values of 0.773, 0.680, and 0.724, respectively.

Conclusion: This work uses individuals' social media messages to identify latent infectious diseases, without prior information, quicker than when the disease(s) is formalized by national public health institutes. In particular, the unsupervised machine learning model using user, textual, and temporal information in social media data, along with sentiment analysis, identifies latent infectious diseases in a given location.

1. Introduction

While recent medical advancements have enabled biomedical professionals to diagnose many acute and chronic diseases with well-defined symptoms, many diseases still show heterogeneous manifestations of symptoms. Furthermore, a large number of individuals have difficulties diagnosing such heterogeneous symptoms, which creates a burden for the public health sector [1]. In order to address such problems, some existing studies comprehensively elucidate the relationships between symptoms and diseases [2]. However, this approach reveals the limitations of effectively discovering latent infectious diseases. In this work, a latent infectious disease is defined as a communicable disease that has not yet been formalized by national public health institutes and explicitly communicated to the general public. In many cases, it takes longer or is impractical to formalize the symptoms of latent infectious diseases. Some biomedical researchers have developed methodologies to detect infectious diseases using electronic medical records (EMRs) [3]. However, access to EMRs is limited and strictly regulated because of patient privacy and consent [4]. Furthermore, an infectious disease requires treatment before one can gain access to EMRs, because an infectious disease spreads in

* Corresponding author at: 213-N Hammond Building, The Pennsylvania State University, University Park, PA 16802, USA.
E-mail address: ctucker4@psu.edu (C.S. Tucker).

http://dx.doi.org/10.1016/j.jbi.2016.12.007
1532-0464/

a given population within a short period of time. Therefore, identifying infectious diseases is necessary for medical treatments in advance of the spread of diseases that results in an increased number of patients and excessive medical expenses [5].

Recently, social media (e.g., *Twitter*, *Facebook*, *Instagram*) has become especially significant as easy-to-access, real-time, and low-cost information sources in biomedical fields [6–10]. Social media data are utilized to communicate among patients and biomedical professionals and to monitor medical-related emergences or infectious diseases [11,12]. In addition, recent advances in social media analysis techniques have enabled researchers to transform social media data into infectious-disease-related knowledge [13,14].

Most existing studies on discovering infectious-disease-related information in social media networks use top-down approaches based on already known information, such as the names of diseases (e.g., Zika, Ebola) or their symptoms (e.g., fever, headache, diarrhea). However, top-down approaches are not appropriate for discovering "latent" infectious diseases in social media networks, because necessary information (e.g., the names of diseases and their symptoms) for top-down disease discovery is mostly unknown before national public health institutes, such as the Centers for Disease Control and Prevention (CDC), formalize latent infectious diseases for a given location and communicate that information to the public. For example, individuals did not use the term "Zika" on social media before CDC named the new virus "Zika". Even if the CDC had already named a disease "Zika", individuals may not use the term "Zika" on social media if they are unaware that Zika had spread to their region. Therefore, this research proposes a bottom-up approach instead of top-down approach for latent infectious disease discovery in a given location without prior information, such as disease names and related symptoms. This research is based on unsupervised machine learning algorithms, that use user, textual, and temporal information from social media networks, along with unsupervised sentiment analysis. A case study involving real EMRs and user, textual, and temporal information from *Twitter* data validates the proposed approach.

This model could prove useful for various disease-related research and applications through the use of easy-to-access, real-time, and low-cost social media data. In particular, this study can help biomedical professionals identify latent infectious diseases, in order to prevent a growing number of patients in a given location and excessive medical expenses.

The remainder of the paper is organized as follows: Section 2 outlines the literature related to this work. Section 3 presents the method based on unsupervised machine learning algorithms and sentiment analysis. Section 4 introduces the case study, and Section 5 presents the experimental results and discussion. Section 6 concludes the paper.

2. Literature review

The literature review describes literature related to disease-related information retrieval from social media networks (Section 2.1) and unsupervised machine learning algorithms using social media data (Section 2.2).

2.1. Disease-related information retrieval from social media networks

Disease-related information is substantial for disease monitoring, prevention, and control [13,15]. Traditional disease-related information retrieval systems from EMRs or biomedical professionals take time and have expensive processes [16]. Social media networks have recently been used for disease-related information retrieval as easy-to-access and real-time information sources. Merolli et al. review the studies on the effects of social media on chronic disease patients and explore the different ways that chronic disease patients use social media [8]. Paul and Dredze propose the Ailment Topic Aspect Model using supervised *tweet* filtering to mine general-public-health-related topics from *Twitter* data [17]. Heaivilin et al. show that social media data can be used as a potential source for dental surveillance [18].

Keyword-based methods and supervised-learning-based methods are two types of methodologies that identify disease-related textual information from social media data [19]. Keyword-based methods require a dictionary containing disease-related keywords as given information. A social media message is classified as "related" if it contains any keywords in the dictionary. Otherwise, it is classified as "non-related" [20]. For instance, several studies on flu-related-keywords are proposed to identify future influenza rates and influenza-like illness using *Twitter* data [21], *Google* search queries [22], and blog posts [23]. Polgreen et al. demonstrate the relationship between search queries for influenza and actual influenza occurrence with the keywords "influenza" and "flu" using the *Yahoo!* search engine [24]. Yang et al. introduce a method to detect the relationship between drugs and adverse drug reactions (ADRs) with related keywords [25]. Hamed et al. propose a network mining approach for linking and searching biomedical literature related to drug interaction and side-effects using *Twitter* hashtags [9]. Bhattacharya et al. present methodologies for surveillance of health beliefs on *Twitter* using probe statements, related to sickness, drugs, or diseases, that are selected manually [26] and automatically [27], respectively. Keyword-based methods are also applied to identify disease-related genes [14] and adverse drug events (ADEs) [28] from healthcare social media data.

Supervised-learning-based methods assume that researchers can use human labeled training data and classify the necessary information based on labeled training data [20]. Collier and Doan propose an algorithm to detect illness-related *tweets* based on naïve Bayes classifiers and support vector machines (SVMs) [29]. Aramaki et al. also use SVMs to train classifiers in order to detect flu-related *tweets* [30]. Huh et al. apply a binary classifier to *WebMD*'s online diabetes community data for assisting moderators in the community [10]. Bodnar et al. develop a supervised-learning-based system for disease detection at the individual level using a sample of professionally diagnosed individuals from social media data [31]. Tuarob et al. present an ensemble supervised-learning-based method that uses multiple classifiers in order to improve the performance of health-related social media message classification [19]. There are several studies on discovering information and evidence about ADRs based on social media data, such as *Twitter* data [12,32] and medical forum posts [33].

Table 1 illustrates a summary of previous studies and this work in relation to disease-related information retrieval contributions. Previous studies on disease-related information retrieval from social media networks are based on top-down approaches that use given information, such as predetermined disease-related keywords [9,14,18,22–28] or human-labeled training data [10,12,17,19,29–34]. However, given information for latent infectious disease discovery is not enough to select disease-related keywords, because latent infectious diseases are nameless, before

Table 1
Summary of previous studies and this work on disease-related information retrieval.

References	*Disease-related keywords*	*Human labeled training data*
[9,14,18,22–28]	Required	Not required
[10,12,17,19,29–34]	Not required	Required
Ours	Not required	Not required

national public health institutes formalize latent infectious diseases. Furthermore, their symptoms may be ambiguous. In addition, manual labeling in social media networks is an expensive process, and manually labeled training social media data are not available when trying to identify information about latent infectious diseases. In this research, a bottom-up method is presented in order to identify latent-infectious-disease-related content expressed in social media networks, without information such as disease-related keywords or human labeled training data.

2.2. Unsupervised machine learning algorithms using social media data

Unlike supervised learning algorithms, which train a learner based on manually labeled training data and then use the trained learner to classify unlabeled data, unsupervised machine learning algorithms train a machine to discover hidden structures and patterns from unlabeled data without target variables [35]. Several researchers have applied unsupervised machine learning algorithms to biomedical areas. For example, Zhang and Elhadad present a stepwise unsupervised method to recognize named entities from biomedical textual data [36]. Wiley et al. use association rule learning to examine pharmaceutical drug discussions on 10 different social media networks and discover that the characteristics of social media affect the content of discussions [37]. Huang et al. present a probabilistic risk stratification model based on topic modeling for clinical risk stratification [38]. Poole et al. propose an unsupervised learning method in order to learn laboratory test reference intervals using laboratory results and coded diagnoses [39].

Clustering is one of the traditional unsupervised machine learning algorithms. Clustering algorithms, such as the *k*-means algorithm, *k*-medoids algorithm, and hierarchical clustering algorithm, divide the entire unlabeled data into relatively homogeneous clusters in order to maximize data similarity within the cluster and data dissimilarity outside the cluster [40–42]. Unsupervised clustering algorithms find natural clusters without prior information, such as the predetermined number of clusters and specific characteristics of clusters [43]. Cluster algorithms have actively been used in biomedical research fields due to rapidly growing biological and medical data generation [44]. For instance, various clustering algorithms are applied to biomedical natural language processing and ontologies [45–49], medical image data analysis [50,51], cytometry data analysis [52,53], and physiological data analysis [54,55]. In particular, cluster algorithms are known as one of the most successful methods for genetic data analysis, such as gene expression data analysis [41,56–58], protein information analysis (e.g., analyzing protein structure, protein sequence, protein-protein interaction) [59,60], and genealogy reconstruction [61]. Clustering is already widely applied to disease-related information retrieval for outbreak detection [62,63], disease progression analysis [40,64,65], and disease clustering using EMRs [66] as well. Clustering algorithms using social media data have recently been applied to biomedical research. Text clustering algorithms using social media data are applied to discover health-related topics [67] and extract ADR-related postings [68]. Yang et al. use *k*-means algorithms for filtering ADR-related textual information from social media data [69].

Unsupervised (or partially supervised) sentiment analysis is used for biomedical studies as well. A partially supervised approach has been used to monitor content containing negative sentiments related to various drugs and medicines and to identify potential ADRs from web forums [70]. Cameron et al. also use sentiment extraction techniques that recognize sentiment-related formal or slang expressions and assess the topic-dependent polarity of each sentiment to discover sentiment clues from web-forum posts [11].

While several existing methodologies based on unsupervised machine learning algorithms and sentiment analysis have been applied to various biomedical areas, limited contributions have been made to identify latent infectious diseases in a given location. Existing methodologies above, based on predetermined attributes (e.g., gene expression or biochemical properties for genetic data analysis), are difficult to apply to latent infectious disease-related information retrieval, since disease attributes are not predetermined and such methodologies require prior information for diseases (e.g., the names of diseases or their symptoms) or EMRs. The main contribution of the proposed unsupervised machine learning model is to discover latent infectious diseases without using predetermined disease attributes.

3. Method

Fig. 1 outlines this research. First, social media messages with user and temporal information are extracted during the data preprocessing stage. Then, an unsupervised sentiment analysis model is presented. Users' expressions about symptoms, body parts, and pain locations are also identified from social media data. The method then creates symptom weighting vectors for each individual and time period, based on their sentiment and social media expressions. Finally, latent-infectious-disease-related information is retrieved from individuals' symptom weighting vectors.

3.1. Social media data acquisition and preprocessing

Social media messages, along with user, temporal, and geospatial information, are extracted. Users' geospatial information, extracted from their social media messages or profiles, are selected for this study. Social media application program interfaces (APIs) can be used for data extraction (e.g., *Twitter* API [71] for extracting *tweets*). Only User IDs, timestamps, geospatial information, and textual information of each message, are filtered and extracted. *t* is defined as a unit of time (e.g., one day, one week, one month), and each user's social media messages are subdivided based on *t*.

Data preprocessing is then implemented to remove noise and to enhance the quality of the results, because social media data are filled with noise that can produce unexpected results [72]. Specifically, *stop words* (e.g., "the", "an") are removed, which represent language-specific functional terms and frequently occurring words in the English dictionary that would be superfluous for disease-related information retrieval [73]. In addition, correcting misspellings and lowercasing are implemented, as well as stemming. Punctuation and hyperlinks are also removed. For example, an original *tweet* "The positive thing is that if its true we have a year to save up lol." is converted to "positive thing that true we have year save up" after preprocessing. Once data preprocessing has been established, it can be applied to any textual messages from different social media platforms without modification by domain experts.

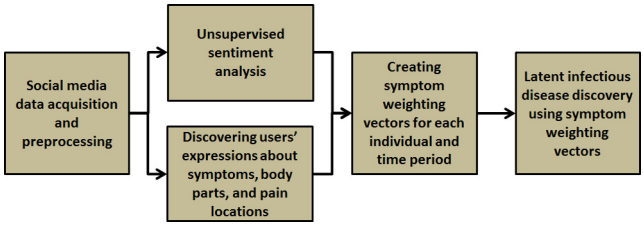

Fig. 1. Overview of this study.

3.2. Symptom discovery from social media data

Both (1) unsupervised sentiment analysis and (2) users' symptom, body part, and pain location expressions extracted from social media data are used to identify whether or not a social media message contains an individual's potential symptoms related to a latent infectious disease. If a message contains symptom, body part, or pain location expressions, but expresses positive sentiment, the user's potential symptoms cannot be identified from the message (A in Fig. 2), because the user's symptoms, such as ADRs, cannot be accompanied by positive sentiment [12]. For instance, the message "I had a headache the past two days, feeling better now because drugs, thanks mom!!!", which expresses positive sentiment, indicates that the user no longer has symptoms, even though the message contains a symptom expression (i.e., "headache"). If a message expresses negative or neutral sentiments, but has no symptom, body part, or pain location expressions, it cannot be classified as indicating a user's symptoms, since a symptom or disease is just one of the reasons for the negative or neutral sentiments (B in Fig. 2). For example, a message "I hate seeing bad parenting" expresses negative sentiment but is not related to the user's symptoms or diseases. Thus, only messages that express negative or neutral sentiments, along with symptom, body part, or pain location expressions (C in Fig. 2), are classified as containing a user's potential symptoms that relate to latent-infectious-disease-related information. These messages are identified through the method in order to discover latent infectious diseases.

3.2.1. Unsupervised sentiment analysis

Sentiment analysis uses natural language processing, text analysis, and computational linguistics to quantify subjective information (i.e., emotions) in a textual message. Since labeled training data is not used in this method, *SentiStrength*, developed by Thelwall et al. [74], is employed for unsupervised sentiment analysis that does not use labeled training data. A social media message is used as an input, and the output is a sentiment score that ranges from −5 to 5. Positive and negative numbers indicate positive sentiment (P) and negative sentiment (N), respectively, and 0 is neutral (−). Table 2 illustrates an example of unsupervised sentiment analysis for *tweets*.

3.2.2. Discovering users' symptom expressions, body part expressions, and pain location expressions from social media data

Among all social media messages written by all individuals, only the messages containing negative or neutral sentiments are considered for this section (e.g., the first, second, and fourth *tweets* in Table 2). A symptom list, a body part list, and a pain location list, along with their relationships, are used in this work for identifying individuals' potential symptoms based on the definitions and assumptions below.

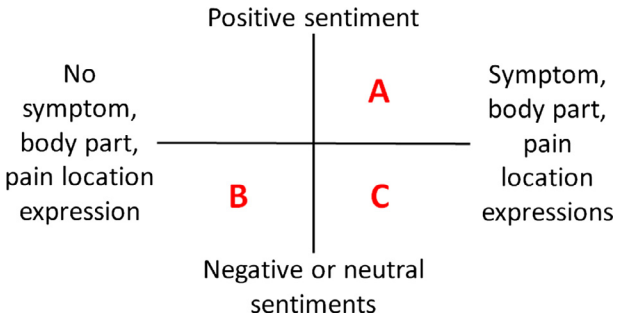

Fig. 2. Potential symptom discovery (A: messages that contain expressions about symptoms, body parts, or pain locations, but express positive sentiment, B: messages that express negative or neutral sentiments, but contain no expressions about symptoms, body parts, or pain locations, C: messages that contain expressions about symptoms, body parts, or pain locations and express negative or neutral sentiments).

Table 2
An example of unsupervised sentiment analysis (P: positive sentiment, N: negative sentiment, –: neutral).

Original tweet	*Sentiment Score*	*P/N/–*
I have the worst cough today. The people on the plane HATE me	−3	N
i am the feeling the fatigue	−2	N
what a beautiful day to buy a ton of new clothes	2	P
I have no opinion about anything at all	0	–
...	...	...

- A symptom is defined as subjective evidence of the disease observed by the individual [75].
- Identifying individuals' symptom expressions is necessary for discovering diseases, because a disease, even a latent infectious disease, can be characterized by different symptom combinations [2,76].
- It is assumed that listing all patient symptoms is available, because it is known that the number of symptoms that a patient expresses is finite [77].
- Identifying individuals' body part expressions (e.g., "mouth", "chest"), along with pain location expressions (e.g., "upper", "lower"), is also necessary, since some users express their condition using body part and pain location expressions instead of symptom expressions. For instance, a user can post on her *Twitter* account, "pain deep inside my head" instead of "I have a headache". Relationships between body parts, pain locations, and symptoms are required in order to predict individuals' symptoms based on their body part expressions and pain location expressions.
- It is assumed that listing all body parts and pain locations is possible, because it is known that the number of body parts and pain locations are finite [78].
- The names of existing diseases are not used in this research, since this study focuses on identifying latent infectious diseases, including nameless new diseases.
- A symptom list, a body part list, and a pain location list, along with their relationships, are therefore required for identifying individuals' potential symptoms from social media networks. These lists make it possible to distinguish different diseases with finite combinations of different symptoms.

In this study, symptom lists, body part lists, and pain location lists from *WebMD* [79], *Mayo Clinic* [80], and *MedlinePlus* [81] are used as data sources. Tables 3–5 show examples of the symptom list, the body part list, and the pain location list, respectively. These lists are used as a primary source in this study. The terms in a primary source can be used to identify biomedical terminologies from social media data. Fig. 3 shows the relationships between body parts (Table 4), pain locations (Table 5), and symptoms (Table 3). For example, Fig. 3 indicates that the body part "chest" can be

Table 3
An example of the symptom list.

Symptom ID	*Symptom expression*
1	cough
2	bleed
...	...
I	diarrhea

Table 4
An example of the body part list.

Body part ID	*Body part expression*
1	mouth
2	chest
...	...
J	abdomen

Table 5
An example of the pain location list.

Pain location ID	*Pain location expression*
1	upper
2	lower
...	...
K	fore

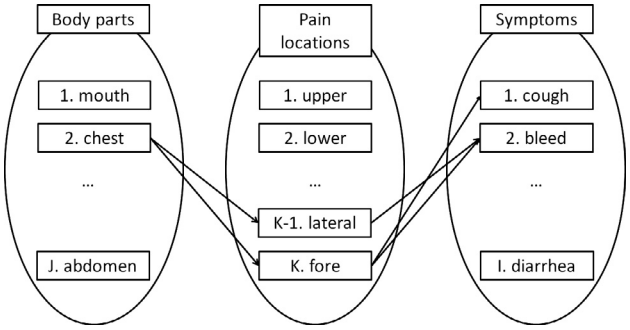

Fig. 3. An example of the relationships between body parts, pain locations, and symptoms for "lateral chest" and "fore chest".

subdivided into "fore chest" and "lateral chest". Fig. 3 also indicates that the symptom "bleed(ing)" can occur in the both the fore and lateral chest, but the symptom "cough" can only occur in the fore chest (not in the lateral chest). Nevertheless, it is assumed that only pain location expressions used with symptom expressions or body part expressions are considered in this research, because only those pain location expressions can be used to subdivide body parts in order to discover users' different potential symptoms. For instance, a message, "the upper middle class has more than doubled since 1979", contains the term "upper" listed in Table 5 but is not related to the user's symptoms or diseases. It is therefore possible to identify users' potential symptoms from their social media messages using not only symptom expressions but also body part and pain location expressions.

However, keyword filtering using a symptom list (Table 3), a body part list (Table 4), and a pain location list (Table 5) is not sufficient for identifying individuals' symptoms from social media messages, because social media messages contain nonstandard languages, such as jargon, due to the heterogeneity of writing formats and constraints placed by social media platforms, such as *Twitter*'s 140-character limit [20]. In addition, an individual who is not a biomedical professional rarely uses technical medical terms, especially for posting on her social media account [68]. For example, it may be common for a patient to post on her *Twitter* account "I have loose bowels" instead of "I have diarrhea". Nonstandard and nontechnical expressions in social media (e.g., synonyms from *WordNet* [82] or *Consumer Health Vocabulary* [83]) corresponding to the symptom, body part, and pain position lists in a primary source (Tables 3–5) are therefore used as a secondary source to minimize a false negative.

Table 6 shows symptom expressions, body part expressions, and pain location expressions from the primary and secondary sources. The primary and secondary sources can be used to identify

Table 6
Symptom expressions, body part expressions, pain location expressions, and sickness/medical expressions from the primary and secondary sources.

	Group	*Primary source*	*Secondary source*
Symptom expression	1	cough	bark
			...
	2	bleed	blood
			hemorrhage
			...
	...	...	...
	I	diarrhea	diarrhoea
			loose bowel
			...
Body part expression	1	mouth	oral
			...
	2	chest	breast
			thorax
			...
	...	...	...
	J	abdomen	stomach
			belly
			...
Pain location expression	1	upper	up
			upside
			...
	2	lower	down
			downside
			...
	...	...	...
	K	fore	forward
			forefront
			...
Sickness/medical expression			sick
			pain
			ill
			disease
			hospital
			clinic
			test
			...

not only biomedical terms (i.e., the primary source) but also nonstandard and nontechnical terms (i.e., the secondary source) from social media data. Sickness/medical expressions are listed in Table 6 as well, because some individuals (i.e., patients) express their sickness or conditions using only sickness or medical expressions (e.g., "sick", "pain") instead of symptom, body part, or pain location expressions (e.g., "I got sick yesterday"). Let I, J, and K be the number of groups for symptoms, body parts, and pain locations, respectively (see Tables 3–5). "Group" means the terms from the primary and secondary sources that indicate the same symptom, body part, or pain location. For example, the term "abdomen" from the primary source and the terms "stomach" and "belly" from the secondary source (i.e., synonyms of "abdomen") belong to the same group "Body part expression J". S_1, S_2, and S_3 are defined as Eqs. (1)–(3), respectively. S_4 is defined as the number of sickness/medical expressions used in this method (see Table 6). Once the primary and secondary sources have been provided in a general sense, they can be applied to any conditions (e.g., different regions, different social media networks) without modification by domain experts, since the proposed method is a bottom-up approach.

$$S_1 = max\{the\ number\ of\ synonyms\ for\ symptom\ i\}, \quad i \in 1,\ldots,I \tag{1}$$

$$S_2 = max\{the\ number\ of\ synonyms\ for\ body\ part\ j\}, \quad j \in 1,\ldots,J \tag{2}$$

$$S_3 = max\{the\ number\ of\ synonyms\ for\ pain\ location\ k\}, \quad k \in 1,\ldots,K \tag{3}$$

In addition, hidden expressions about diseases, symptoms, body parts, or pain locations can be considered. For example, the term "clap" usually refers to the act of striking together the palms of the hands (e.g., "they always clap for us"). However, the term "clap" can also be used instead of gonorrhea (e.g., "I will go out to get tested for the clap tomorrow") if "clap" is used with the term(s) contained in the primary or secondary sources (i.e., "test (ed)") in social media. Therefore, the top L frequent terms that are not stop words or already contained in the primary or secondary sources, are identified in all messages containing any symptom, body part, or sickness/medical expressions in Table 6. This identification discovers hidden expressions of diseases or symptoms without prior information, such as the disease names and related symptoms. Table 7 shows an example of the top L frequently used terms for discovering hidden expressions. Domain experts can set L differently to satisfy Eq. (4) based on the assumption that the optimal number of hidden expressions (i.e., synonyms of the expression indicating a latent infections disease) is not greater than the maximum number of synonyms for symptom, body part, or pain location expressions in Table 6. For instance, a relatively large value (e.g., $\max\{S_1, S_2, S_3\}$) less than 10 (e.g., The maximum number of synonyms for each symptom, body part, or pain location, that are identified through *WordNet* [82] and *Consumer Health Vocabulary* [83], is 8.) is used to set L when it is important to discover nameless new diseases from a given population. On the other hand, a small value (e.g., 0) is used to set L when it is necessary to decrease a false positive.

$$L \leqslant max\{S_1, S_2, S_3\} \tag{4}$$

Table 8 shows an example of how to identify potential symptoms from social media messages. Symptom weights are defined as the possibilities that a message contains information related to each potential symptom. The summation of symptom weights is set to 1 for each message if the message is considered to indicate user's potential symptom(s). Otherwise, it is set to 0. If more than one potential symptom is discovered in one message, symptom weights are evenly allocated to each potential symptom in this study, because it is assumed that each potential symptom has the same possibility only based on a social media message without symptom-related information. The user's symptom (i.e., cough) can be identified from the first *tweet* in Table 8, since the *tweet* contains the keyword "cough". Thus, a symptom weight 1 is allocated to the symptom "cough", since it is the only symptom that is discovered from the first *tweet*. While the second *tweet* in Table 8 does not contain any symptom keywords, five potential symptoms, including "cough" and "bleed", can be identified by the body part keyword "breast", the pain location keyword "fore", and their relationship (see Fig. 3). Therefore, a symptom weight 1/5 is allocated to five symptoms, including "cough" and "bleed", respectively. However, the fourth *tweet* is not considered to indicate potential symptoms, since pain location expressions without symptom or body part expressions cannot give disease-related or symptom-related information. Messages may not contain specific symptom or body part keywords, but if they contain an individual's sickness or medical expressions (e.g., the fifth *tweet* in Table 8) or any of L frequent terms (e.g., the sixth *tweet* in Table 8), they are not disregarded, since they can have potential symptom information for individuals. Because it is assumed that all potential symptoms can occur in the individual who wrote the message, symptom weight $1/I$ is allocated to all potential I symptoms in these cases.

Table 7
An example of the top L frequently used terms for discovering hidden expressions.

Number	*Frequently used term*
1	clap
2	drip
...	...
L	clam

However, the keyword "test" from Table 6 may be used as a symptom-related expression (e.g., the sixth *tweet* in Table 8), but it can be used as a non-symptom-related expression as well (e.g., "we test the microphone."). While the third *tweet* in Table 8 contains the body part keyword "chest" and symptom weights are allocated to the *tweet*, it is not actually related to the user's symptoms or diseases. Those cases above can increase false positives.

In this study, co-occurrence analysis is employed to reduce false positives without training data or prior information as follows. Table 9 shows an example of co-occurrence analysis. First, if the term "rhinorrhea" and first person singular pronouns (i.e., "I", "me", "my", "mine", "myself") co-occur in the same message (e.g., the fifth *tweet* in Table 9), it is assumed that the probability that the message indicates a user's symptoms is higher than the probability that other messages containing only the term "rhinorrhea" indicate a user's symptoms. This assumption occurs, since this research focuses on discovering users' symptoms instead of their friends' symptoms or general disease-related information, and first person singular pronouns are more frequently used in social media when the user is unstable [84,85]. Thus, a weighting factor α is assigned to the fifth *tweet* in Table 9. α is set to 1.802 as the default for *Twitter* data. Recent research has shown that there is an 80.2% higher probability that a social media message with first-person singular pronouns is written by a new mother with postpartum depression than a social media message that does not contain first-person singular pronouns, if it is assumed that *Twitter* is used for their social media [85,86].

In addition, if the terms "chest" and "pain" from Table 6 co-occur in the same message (e.g., the first *tweet* in Table 9), it is assumed that the probability that the message indicates a user's symptoms is higher than the probability that other messages containing only "chest" or "pain" indicate a user's symptoms, because a message containing more than one keyword is more informative than a message containing just one keyword [87]. Suppose that the third *tweet* in Table 9 contains only one term (i.e., "hospital") from Table 6 without any first person singular pronouns. A weighting factor can also be applied to the message, if the message and other messages that have a high probability of indicating a user's symptoms (i.e., the first or second *tweet* in Table 9) are written by the same user in the same period (e.g., the same month where t = one month). Thus, weighting factor β is assigned to the first, second, and third *tweets* in Table 9. β is set to 1.524 as the default based on Miller et al.'s [88] finding that considering term co-occurrence in the same sentence improve sense identifications for open-class words by 52.4%. It is therefore possible to classify all messages that have negative or neutral sentiments based on their potential symptom weights and co-occurrence analysis, as shown in Table 10. A symptom weighting vector is available for each individual for a certain period of time based on the message classification results in Table 10. Table 11 shows an example of how to create a symptom weighting vector that is normalized by the total number of *tweets* (i.e., 27) written by User 1 during Period 1 multiplied by $\alpha \cdot \beta$ (i.e., the maximum possible summation of symptom weights for each symptom).

3.3. Latent infectious disease discovery

Biomedical professionals can investigate individuals who are predicted to have abnormal symptoms and latent infectious diseases based on their symptom weighting vectors. Symptom weighting vectors, which are created by social media data until the last time the individuals are diagnosed (i.e., when EMRs are

Table 8
An example of potential symptom identifications from social media messages.

Original tweets	*Symptom weight*				*Body part*				*Pain location*				*Sickness/ medical expression*	*Frequent term*
	1. cough	*2. bleed*	…	*I. diarrhea*	*1. mouth*	*2. chest*	…	*J. abdomen*	*1. upper*	*2. lower*	…	*K. fore*		
Been so sick the last few days I have such a bad **cough** ah it's no fun I hope you're all doing well!	1	0	…	0			…				…			
I just did some **sick** in my **front breast**	1/5	1/5	…	0		√	…		√		…		√	
When the gym is closed but it's **chest** day	1/7	1/7	…	0		√	…				…			
Lower prices, more jobs, more trade	0	0		0						√				
I got **sick** yesterday	1/*I*	1/*I*	…	1/*I*			…				…		√	
I will go out to get **tested** for the **clap** tomorrow	1/*I*	1/*I*	…	1/*I*			…				…		√	√
…	…	…	…	…			…		…	…	…	…		

Table 9
An example of co-occurrence analysis (α: weighting factor for the first person singular pronouns, β: weighting factor for term co-occurrence).

User	*Period*	*Original tweet*	*First person singular pronoun*	*Term co-occurrence*
1	1	it seems to have **chest pain**		β
2	2	**throat** currently feels like someone shoved sandpaper **down** it… **cough** syrup why you no help		β
2	2	have to go to the **hospital**		β
4	5	When the gym is closed but it's **chest** day		
5	4	**I** have **rhinorrhea** today	α	

Table 10
An example of a classification of all messages having negative or neutral sentiments based on their (potential) symptom expressions and co-occurrence analysis.

User	*Period*	*Total number of tweets*	*Original tweet*	*Symptom weight*				*First person singular pronoun*	*Co-occurrence*
				1. cough	*2. bleed*	…	*I. diarrhea*		
1	1	27	**I** got **sick** yesterday	1/*I*	1/*I*	…	1/*I*	α	β
			it seems to have **chest pain**	1/7	1/7	…	0		β
			…	…	…	…	…	…	…
	…	…	…	…	…	…	…	…	…
	11	34	so **sick** of wasting **my** time	1/*I*	1/*I*	…	1/*I*	α	
			I go to home after this busy/terrible day	0	0	…	0	α	
			…	…	…	…	…	…	…
…	…	…	…	…	…	…	…	…	…

Table 11
An example of how to create a symptom weighting vector for User 1 in Period 1.

User	*Period*	*Total number of tweets*	*Original tweet*	*Symptom weight*			
				1. cough	*2. bleed*	…	*I. diarrhea*
1	1	27	I got **sick** yesterday	$\alpha{\cdot}\beta$ /*I*	$\alpha{\cdot}\beta$ /*I*	…	$\alpha{\cdot}\beta$ /*I*
			it seems to have **chest pain**	β/7	β/7	…	0
			…	…	…	…	…
A normalized symptom weighting vector				$(\alpha{\cdot}\beta$ /*I* + β/7 + …)/$(27{\cdot}\alpha{\cdot}\beta)$	$(\alpha{\cdot}\beta$ /*I* + β/7 + …)/$(27{\cdot}\ \alpha{\cdot}\beta)$		$(\alpha{\cdot}\beta$ /*I* + 0 + …)/$(27{\cdot}\alpha{\cdot}\beta)$

available: P1 in Fig. 4) and validated by real medical records, can be used as clustering symptom weighting vectors that indicate the same existing disease (e.g., influenza). For methodological convenience, the "absence of disease" can also be considered as just one of the diseases in this step. Those symptom weighting vectors can be applied to not only individuals who have been diagnosed during P1, but also individuals who have not been diagnosed during P1. This application is possible, because a disease can be characterized by different symptom combinations for different individuals [2,76], and symptom weighting vectors are already normalized in the previous step (see Table 11). In addition, the symptom weighting vectors that indicate existing diseases are not labeled training data, since this research aims to discover latent infectious diseases instead of existing diseases. Prior information (e.g., EMRs for latent infectious diseases, the names of diseases, related symptoms) for labeling training data is unavailable in this

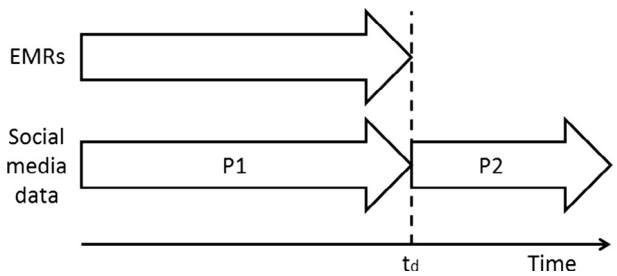

Fig. 4. Time periods for social media data based on EMR availability (P1: EMRs available, P2: EMRs not available, t_d: the last time the individuals are diagnosed.).

study. If the similarity between (1) a new symptom weighting vector v, which is created using social media data after the last time the individuals are diagnosed (i.e., when EMRs are not available:

P2 in Fig. 4) and (2) a cluster C, which contains symptom weighting vectors (created during P1 in Fig. 4) indicating an existing disease D, is less than a similarity criterion δ (i.e., greater than a dissimilarity criterion $1-\delta$), biomedical professionals should investigate the individual who corresponds to the new symptom weighting vector v in order to diagnose potential latent infectious diseases. The average linkage clustering is used as the distance between a cluster C and a new symptom weighting vector v ($D(C, v)$) as Eq. (5), since it is assumed that the centroid of the cluster C represents the symptom weighting vector for an existing infectious disease D.

$$D(C, v) = \frac{1}{|C||v|}\sum_{i\in C}\sum_{j=v} d_{ij} = \frac{1}{|C|}\sum_{i=C} d_{iv} \quad (5)$$

Based on the recent text mining research, δ is set to 0.8 as the default with a cosine similarity (i.e., the cosine similarity value 0.8 used for clustering topic vectors) [89].

Table 12 illustrates an example of the individuals' weighting vectors, which are subdivided based on P1 and P2 in Fig. 4. Fig. 5 shows an example of clusters that indicate existing diseases for P1 (i.e., "influenza" and "absence of disease") and new clustering objects (i.e., new symptom weighting vectors for P2). Objects 1, 2, and 3 indicate symptom weighting vectors for User 1 in Period 1, User 2 in Period 2, and User 10 in Period 10, respectively, in Table 12 (P1). In addition, Objects 4, 5, and 6 indicate vectors for User 1 in Period 11, User 13 in Period 13, and User 15 in Period 13, respectively, in Table 12 (P2). For instance, biomedical professionals should examine User 15 in order to determine whether or not she had a disease(s) with symptoms that include "cough" and "bleed(ing)". Biomedical professionals should also determine if the similarity between her weighting vector for Period 13 (present) and any weighting vector indicating an existing disease(s), including the "absence of disease," is less than a similarity criterion δ (i.e., greater than a dissimilarity criterion $1-\delta$ (Fig. 5)). Figs. 6 and 7 illustrate a process example of the overall method and show how to a create symptom weighting vector for User 1 in Period 1 (i.e., the magnified dotted box in Fig. 6), respectively.

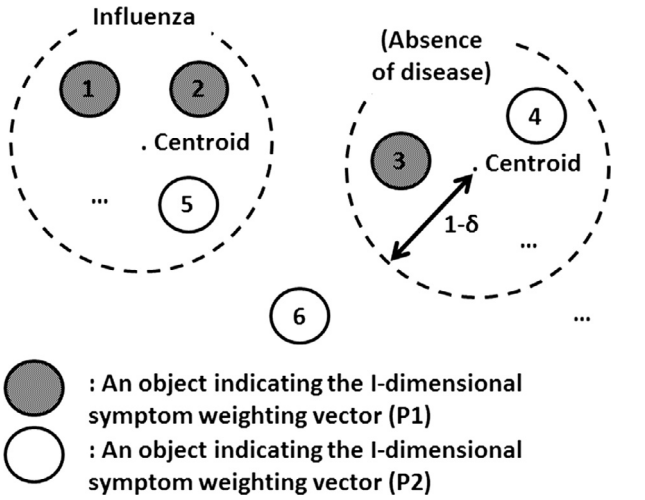

Fig. 5. An example of clusters indicating existing diseases for P1 and new symptom weighting vectors for P2 (Object 1: a vector for User 1 in Period 1, Object 2: a vector for User 2 in Period 2, Object 3: a vector for User 10 in Period 10, Object 4: a vector for User 1 in Period 11, Object 5: a vector for User 13 in Period 13, Object 6: a vector for User 15 in Period 13).

4. Application

This section provides a case study involving real EMRs and social media data to verify this work. Experiments are conducted on an i5-750 processor with 4.00 GB RAM using *Python* 2.7.12. The case study identifies one infectious disease (i.e., influenza) in a given location (i.e., Centre County, Pennsylvania). Therefore, the number of clusters are two (i.e., "influenza" and "absence of disease"), in this case study (see Fig. 5). The Fox stop list [90] and the Porter stemming algorithm [91] are used for removing stop words and stemming, respectively. The default values are used for setting α, β, and δ in the case study (α = 1.802, β = 1.524, δ = 0.8).

Symptom lists, body part lists, and pain location lists provided by *WebMD* [79], *Mayo Clinic* [80], and *MedlinePlus* [81] are used as a primary source for the symptom list, body part list, and pain location list for this case study. The numbers of symptoms (I), body parts (J), and pain locations (K) are 87, 37, and 9, respectively. Synonyms of terms on a symptom list, a body part list, and a pain location list obtained from *WordNet* [82] and *Consumer Health Vocabulary* [83], along with 17 sickness/medical expressions from *Consumer Health Vocabulary* [83] (i.e., "sick", "pain", "ill", "disease", "hospital", "clinic", "test", "ache", "damage", "dysfunction", "chronic", "disorder", "injury", "discomfort", "abnormal", "health", "medical"), are used as a secondary source.

EMRs from August 2012 to May 2013 (10 months) for 104 individuals who were diagnosed with influenza from the Penn State's Health Services and had used their *Twitter* accounts in the same period serve as voluntary participants in the case study [31]. All 104 participants were residents of Centre County, Pennsylvania. Data collection was approved through Penn State's IRB (approval #41345). EMRs only indicate which month an individual was diagnosed with influenza. Table 13 illustrates EMRs used in this case study and "√" indicates that the individual was diagnosed with influenza during the period. EMRs for influenza (instead of EMRs

Table 12
An example of the individuals' weighting vectors that are subdivided based on P1 and P2, where t_d: the time between Period 10 and Period 11.

User	*Period*		*A normalized symptom weighting vector*				*EMRs*
			1. cough	*2. bleed*	...	*I. diarrhea*	
1	1	P1	0.35	0.01		0.14	Influenza
...	...		...	...		...	...
2	2		0.37	0.00		0.19	Influenza
...	...		...	...		...	...
10	10		0.00	0.01		0.00	(Absence of disease)
1	11	P2	0.00	0.01	...	0.00	Not available
...	...		...	...	...	...	...
13	13 (present)		0.40	0.03	...	0.18	Not available
...	...		...	...	...	...	...
15	13 (present)		0.33	0.54	...	0.01	Not available
...	...		...	...	...	...	...

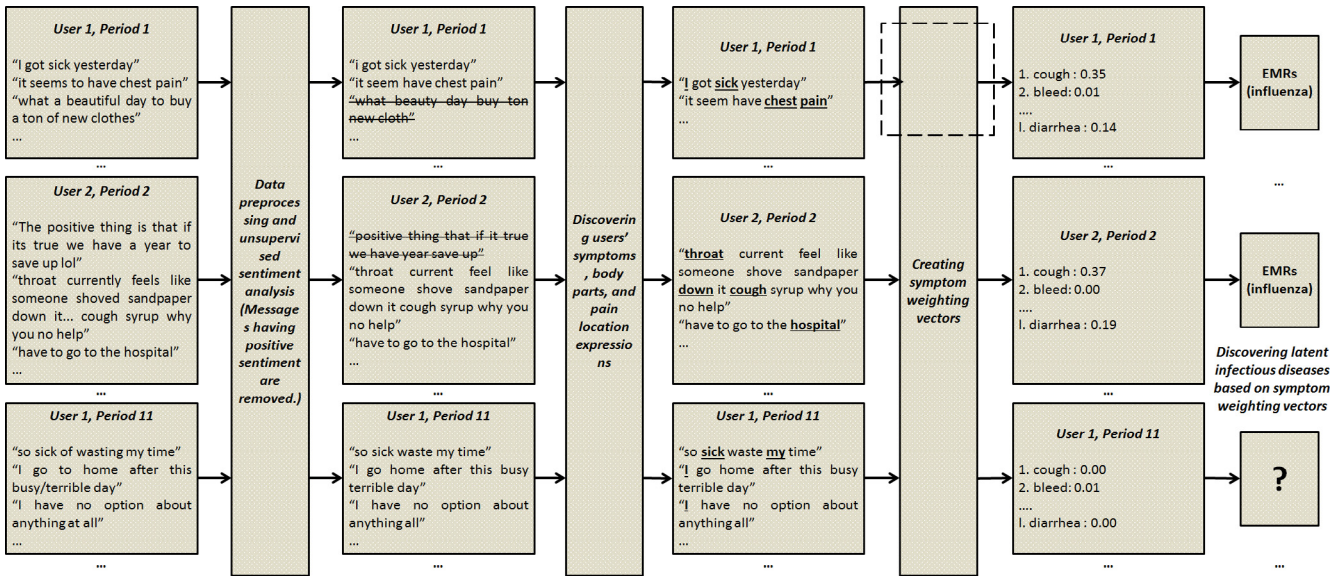

Fig. 6. A process example of this study.

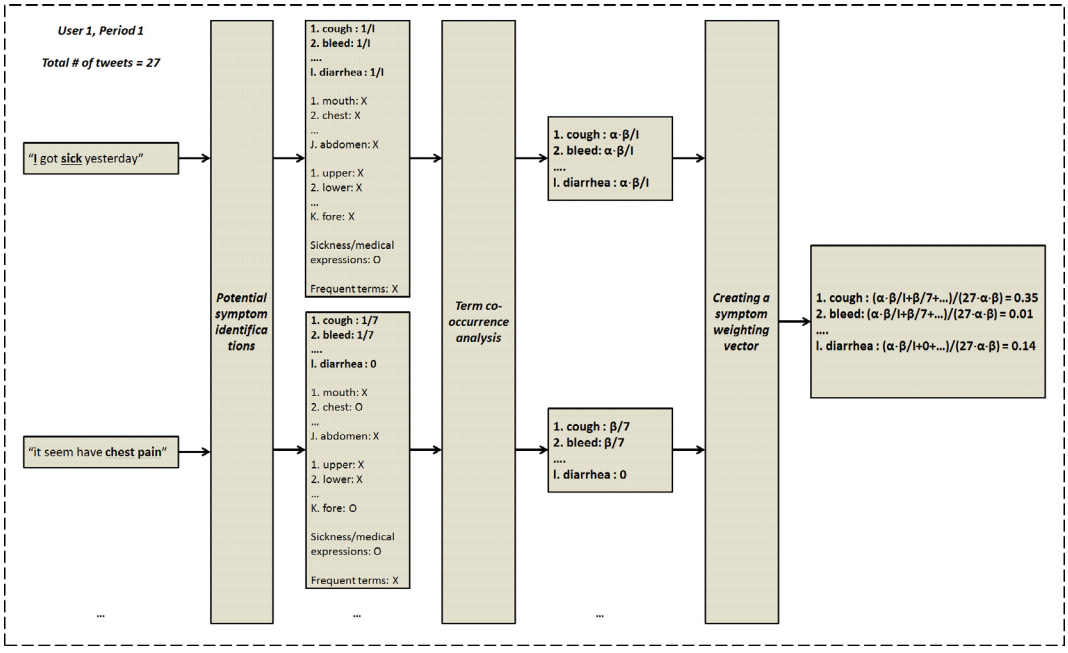

Fig. 7. A process example of how to a create symptom weighting vector for User 1 in Period 1 (i.e., the magnified dotted box in Fig. 7).

for latent infectious diseases that are not available) are only used to serve as ground truth validation. An actual implementation of the method would not require real EMRs, since this research aims to identify latent infectious diseases much earlier than waiting until EMR access.

Twitter data (i.e., all 104 participants' *tweets*) from August 2012 to May 2013 are used in the case study as well [31]. The *Twitter* API is used for data extraction from all 104 *Twitter* accounts. A filter limits the *Twitter* data acquisition process to the most recent 3,000 *tweets* for each user [71]. Only timestamps and textual information of each *tweet* are extracted using user IDs as a further filter. The extracted information is stored in compressed text files, yielding the total size of 31.8 MB (37,599 *tweets* with user and temporal information). In order to compare EMRs based on the same unit of time, t is set to one month. In this case study, *Twitter* data from August 2012 to January 2013 (i.e., P1 in Fig. 4) are used to create symptom weighting vectors for influenza, based on symptom weighting vectors of individuals who were diagnosed with influenza. On the other hand, *Twitter* data from February 2013 to May 2013 (i.e., P2 in Fig. 4) are used for validating this study (i.e., unseen data for validation).

5. Experiments and results

Among a total of 37,599 *tweets,* 14,501 *tweets* containing positive sentiments are removed through unsupervised sentiment analysis. Among 23,098 *tweets* containing neutral or negative sentiments, only 8,877 *tweets* are considered to have potential symptom-related information based on the primary and secondary sources.

Table 13
EMRs used in this case study (√: The individual was diagnosed with influenza.).

Individual	*Aug. 2012*	*Sep. 2012*	*Oct. 2012*	...	*May. 2013*
1				...	
...	...	...	...	...	...
10		√	√	...	
...	...	...	...	...	...
104				...	

Table 14
The top eight most frequently used terms.

Rank	*Term*	*Rank*	*Term*
1	year	5	people
2	game	6	college
3	time	7	today
4	day	8	watch

An F_1 score, which is often used in the field of information retrieval, along with precision and recall, is used for validating this study using *Twitter* data and ground truth data (i.e., EMRs), because the presence of a disease (i.e., influenza in this case study) is considered more important than its absence (i.e., asymmetric), and an F_1 score is not affected by the value of true negatives [92]. Both precision (i.e., a positive predictive value (PPV)) and recall (i.e., a true positive rate (TPR)) are important in latent infectious disease discovery, because low precision (i.e., a high false positive rate) can cause excessive medical expenses and low recall (i.e., a high false negative rate) can cause growing number of patients due to infectiousness of the diseases. A negative predictive value (NPV) is also used to validate the effects of true negatives that are not used in an F_1 score, precision, and recall. If the cosine similarity between the centroid of the cluster containing the symptom weighting vectors for influenza (created during P1 in Fig. 4) and each individual's symptom weighting vector during P2 in Fig. 4 is greater than δ, the individual is predicted to have influenza during that period. A default value is used to set δ (i.e., 0.8). Different values of L, ranging from 0 to 8, are used, because the maximum value of L (i.e., max $\{S_1, S_2, S_3\}$) is 8 (i.e., the number of synonyms of the symptom "swelling" ("swell", "dropsi", "hydrop", "oedema", "lump", "edema", "bulg", and "tumefact" after applying stemming)) in this study. Table 14 indicates the top eight most frequently used terms. According to Table 14, these terms are related to time (i.e., "year", "time", "day", "today") or daily life (i.e., "game", "people", "college"). In this case study, it is postulated that individual's symptom or sickness expressions often accompany temporal or daily life expressions.

Table 15
The precision, recall, negative predictive value (NPV), and F_1 score results for each case (cases: sorted in descending order by the average F_1 score, underlined values: the highest values).

Case	*L*	*Precision*	*Recall*	*NPV*	*F_1 score*
(4) α = 1.802 (default), β = 1.524 (default)	0	0.773	0.680	0.098	0.724
	1	0.773	0.680	0.098	0.724
	2	0.773	0.680	0.098	0.724
	3	0.739	0.680	0.099	0.708
	4	0.739	0.680	0.099	0.708
	5	0.708	0.680	0.100	0.694
	6	0.708	0.680	0.100	0.694
	7	0.708	0.680	0.100	0.694
	8	0.708	0.680	0.100	0.694
	Average	0.725	0.680	0.099	0.702
(2) α = 1.802 (default), β = 1	0	0.625	0.600	0.125	0.612
	1	0.625	0.600	0.125	0.612
	2	0.625	0.600	0.125	0.612
	3	0.625	0.600	0.125	0.612
	4	0.625	0.600	0.125	0.612
	5	0.625	0.600	0.125	0.612
	6	0.625	0.600	0.125	0.612
	7	0.625	0.600	0.125	0.612
	8	0.625	0.600	0.125	0.612
	Average	0.625	0.600	0.125	0.612
(3) α = 1, β = 1.524 (default)	0	0.636	0.560	0.134	0.596
	1	0.609	0.560	0.136	0.583
	2	0.609	0.560	0.136	0.583
	3	0.609	0.560	0.136	0.583
	4	0.609	0.560	0.136	0.583
	5	0.583	0.560	0.138	0.571
	6	0.583	0.560	0.138	0.571
	7	0.583	0.560	0.138	0.571
	8	0.583	0.560	0.138	0.571
	Average	0.600	0.560	0.137	0.579
(1) α = 1, β = 1	0	0.565	0.520	0.148	0.542
	1	0.565	0.520	0.148	0.542
	2	0.565	0.520	0.148	0.542
	3	0.565	0.520	0.148	0.542
	4	0.565	0.520	0.148	0.542
	5	0.565	0.520	0.148	0.542
	6	0.565	0.520	0.148	0.542
	7	0.565	0.520	0.148	0.542
	8	0.565	0.520	0.148	0.542
	Average	0.565	0.520	0.148	0.542

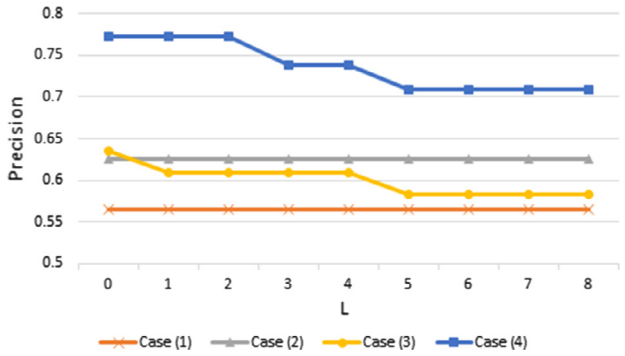

Fig. 8. The precision results for each case.

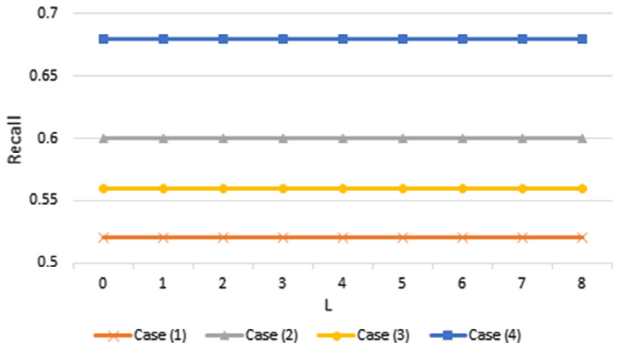

Fig. 9. The recall results for each case.

Sensitivity analysis is implemented to validate the performance of this study with real EMRs, instead of comparing the results with other existing methods, since the proposed method is a bottom-up approach to identify latent infectious diseases from social media data without prior information (e.g., predetermined disease-related keywords or human-labeled training data). In order to quantify the effects of both first person singular pronouns and term co-occurrence, four cases, which are (1) not considering both first person singular pronouns and term co-occurrence (i.e., $\alpha = 1$ and $\beta = 1$), (2) only considering first person singular pronouns (i.e., $\alpha = 1.802$ and $\beta = 1$), (3) only considering term co-occurrence (i.e., $\alpha = 1$ and $\beta = 1.524$), and (4) considering both first person singular pronouns and term co-occurrence (i.e., $\alpha = 1.802$ and $\beta = 1.524$), are run with a different L (from 0 to 8) in the case study. The running time for each run is less than one minute. Table 15 indicates the precision, recall, NPV, and F_1 score results for each case. Figs. 8–11 show the precision, recall, NPV, and F_1 score results, respectively.

According to Table 15 and Figs. 8–11, the proposed method that considers both first person singular pronouns and term co-occurrence (i.e., Case (4)) provides better precision, recall, NPV, and F_1 score values than Case (1), Case (2), and Case (3), regardless of the value of L. Table 15 also shows that the highest value of F_1 score is 0.724 (Case (4) where $L = 0$, 1, or 2). Table 16 illustrates F_1 scores of previous studies and this work in relation to disease-related information retrieval from *Twitter* data, in order to evaluate the performance of this work on a qualitative basis.

Table 16 shows that this study is useful to identify latent infectious diseases in early stages without (1) disease-related keywords (i.e., the term "influenza" and its synonyms in this case study) or (2) human labeled training data in comparison with F_1 scores from previous studies that use disease-related keywords or human labeled training data. The effects of both first person singular pronouns and term co-occurrence are not negligible, since they can be used for improving identification performance (i.e., precision, recall, and F_1 score values in this case study). Future work will investigate possible weighting factors, other than first person singular pronouns (α) and term co-occurrence (β) used in this research, based on social media semantic analysis that give better values of precision, recall, and F_1 score than the current results (Table 15).

Table 15 indicates that the value of L does not affect the precision, recall, NPV, and F_1 score values when term co-occurrence is not considered (i.e., Case (1) and Case (2)), since the top L frequently used terms can only be used when they co-occur with terms in the primary or secondary source (see Table 6) by definition in Section 3.2.2. Table 15 and Fig. 9 show that the value of L does not affect the values of recall for all cases. In addition, according to Table 15 and Figs. 8 and 11, the precision and F_1 score values decrease as the value of L increases when term co-occurrence is considered (i.e., Case (3) and Case (4)). In the same manner, Table 15 and Fig. 10 indicate that the NPV values increase as the value of L increases when term co-occurrence is considered. This means that the L top frequently used terms in this case study are not beneficial to discover potential diseases (i.e., influenza, instead of nameless new diseases, in this case study). It is postulated that all of the top eight most frequently used terms in this case study (e.g., "year", "game") do not directly relate to an individual's sickness or condition, so they can increase false positives, along with a term co-occurrence weighting factor β. Nevertheless, the top most frequently used terms can be useful for biomedical professionals as references when investigating hidden expressions for nameless new diseases.

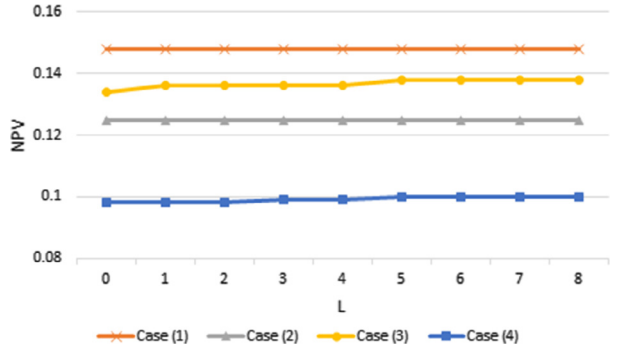

Fig. 10. The NPV results for each case.

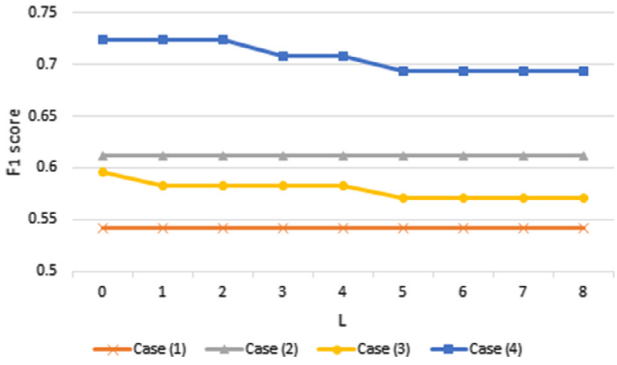

Fig. 11. The F_1 score results for each case.

According to Table 15, the recall values are relatively less than the precision values, even for the highest values that are underlined in Table 15. Based on *Twitter* data used in this case study, it is postulated that (1) some users rarely use their social media accounts, (2) some users only use their social media account for

Table 16
F_1 scores of previous studies and this work on disease-related information retrieval from *Twitter* data.

Reference	*Disease-related keywords*	*Human labeled training data*	*(Highest) F_1 score*
[21]	Required	Required	0.902
[26]	Required	Not required	0.823
[19]	Not required	Required	0.635
[32]	Not required	Required	0.770
[68]	Not required	Required	0.721
Ours	Not required	Not required	0.724

sharing news or information (e.g., *retweets*), or (3) some users do not tend to share their sickness or condition with others. Future work will present how to increase the recall values when considering social media user tendencies, since the recall values are related to the patient's growth rate due to infectiousness of the diseases. A symptom allocation method that creates weighting vectors will be also proposed to allocate symptoms when considering symptom incidence rates instead of the equal allocation used in this method (e.g., the second, third, fifth, and sixth *tweets* in Table 8), in order to improve the precision values, since the precision values are related to medical expenses (i.e., the cost of misdiagnosis).

6. Conclusion and future work

The authors present a method to discover latent infectious diseases without given information, such as the name of diseases and their symptoms. The proposed unsupervised machine learning model identifies latent infectious diseases in a given location using user, textual, and temporal information in social media data.

The proposed research is comprised of four main steps. First, user, textual, and temporal information are extracted from social media data, and data preprocessing is exploited. Unsupervised sentiment analysis is then implemented, and users' expressions about symptoms, body parts, and pain locations are discovered from social media data. An unsupervised-based method is then presented to create symptom weighting vectors for each individual and time period based on an individual's sentiments and expressions. Finally, latent-infectious-disease-related information is identified from individuals' symptom weighting vectors.

A case study involving EMRs and *Twitter* data is used to validate this work. It is concluded that this research, which uses social media data, can identify latent infectious diseases without prior information in a short period of time (e.g., less than one minute, in this case study). Domain experts (i.e., biomedical professionals or biomedical data analysts) can set α, β, and δ differently based on different social media networks or similarity measures, instead of the default values (i.e., $\alpha = 1.802$, $\beta = 1.524$, $\delta = 0.8$) used in this case study.

Future work will include theoretical approaches on how to improve the performance (i.e., precision, recall, NPV, and F_1 score) of the proposed method which identifies latent infectious diseases using social media data. The authors will also present a method to improve the accuracy of identifying latent infectious diseases when considering social media user information (e.g., gender, age, posting frequency). A research expansion to identify latent infectious diseases where individuals in a given population cannot use social media due to their symptoms (e.g., serious eye or hand damage) will be considered in future research as well.

Acknowledgements

The authors acknowledge the NSF I/UCRC Center for Healthcare Organization Transformation (CHOT), NSF I/UCRC grant #1067885 and Penn State's Global Engagement Network (GEN) grant for funding this research. Any opinions, findings, or conclusions found in this paper are those of the authors and do not necessarily reflect the views of the sponsors. The authors would like to acknowledge Haojun Sui for the programming aspects of this work.

References

[1] C.B. Forrest, R.J. Bartek, Y. Rubinstein, S.C. Groft, The case for a global rare-diseases registry, Lancet 377 (9771) (2011) 1057–1059.
[2] X. Zhou, J. Menche, A.-L. Barabási, A. Sharma, Human symptoms–disease network, Nat. Commun. 5 (2014).
[3] H.J. Murff, V.L. Patel, G. Hripcsak, D.W. Bates, Detecting adverse events for patient safety research: a review of current methodologies, J. Biomed. Inform. 36 (1–2) (2003) 131–143.
[4] P.B. Jensen, L.J. Jensen, S. Brunak, Mining electronic health records: towards better research applications and clinical care, Nat. Rev. Genet. 13 (6) (2012) 395–405.
[5] Centers for Disease Control, Principles of epidemiology. An introduction to applied epidemiology and biostatistics, Atlanta, 1992.
[6] P. Chira, L. Nugent, K. Miller, T. Park, S. Donahue, A. Soni, D. Nugent, C. Sandborg, Living profiles: design of a health media platform for teens with special healthcare needs, J. Biomed. Inform. 43 (5) (2010) S9–S12.
[7] P.F. Brennan, S. Downs, G. Casper, Project HealthDesign: rethinking the power and potential of personal health records, J. Biomed. Inform. 43 (5) (2010) S3–S5.
[8] M. Merolli, K. Gray, F. Martin-Sanchez, Health outcomes and related effects of using social media in chronic disease management: a literature review and analysis of affordances, J. Biomed. Inform. 46 (6) (2013) 957–969.
[9] A.A. Hamed, X. Wu, R. Erickson, T. Fandy, Twitter K-H networks in action: advancing biomedical literature for drug search, J. Biomed. Inform. 56 (2015) 157–168.
[10] J. Huh, M. Yetisgen-Yildiz, W. Pratt, Text classification for assisting moderators in online health communities, J. Biomed. Inform. 46 (6) (2013) 998–1005.
[11] D. Cameron, G.A. Smith, R. Daniulaityte, A.P. Sheth, D. Dave, L. Chen, G. Anand, R. Carlson, K.Z. Watkins, R. Falck, PREDOSE: a semantic web platform for drug abuse epidemiology using social media, J. Biomed. Inform. 46 (6) (2013) 985–997.
[12] A. Sarker, G. Gonzalez, Portable automatic text classification for adverse drug reaction detection via multi-corpus training, J. Biomed. Inform. 53 (2015) 196–207.
[13] S. Ayvaz, J. Horn, O. Hassanzadeh, Q. Zhu, J. Stan, N.P. Tatonetti, S. Vilar, M. Brochhausen, M. Samwald, M. Rastegar-Mojarad, M. Dumontier, R.D. Boyce, Toward a complete dataset of drug–drug interaction information from publicly available sources, J. Biomed. Inform. 55 (2015) 206–217.
[14] J. Kim, H. Kim, Y. Yoon, S. Park, LGscore: a method to identify disease-related genes using biological literature and Google data, J. Biomed. Inform. 54 (2015) 270–282.
[15] L.N. Carroll, A.P. Au, L.T. Detwiler, T. Fu, I.S. Painter, N.F. Abernethy, Visualization and analytics tools for infectious disease epidemiology: a systematic review, J. Biomed. Inform. 51 (2014) 287–298.
[16] N.G. Weiskopf, G. Hripcsak, S. Swaminathan, C. Weng, Defining and measuring completeness of electronic health records for secondary use, J. Biomed. Inform. 46 (5) (2013) 830–836.
[17] M.J. Paul, M. Dredze, Discovering health topics in social media using topic models, PLoS One 9 (8) (2014) e103408.
[18] N. Heaivilin, B. Gerbert, J.E. Page, J.L. Gibbs, Public health surveillance of dental pain via Twitter, J. Dent. Res. 90 (9) (2011) 1047–1051.
[19] S. Tuarob, C.S. Tucker, M. Salathe, N. Ram, An ensemble heterogeneous classification methodology for discovering health-related knowledge in social media messages, J. Biomed. Inform. 49 (2014) 255–268.
[20] S. Lim, C.S. Tucker, A Bayesian sampling method for product feature extraction from large scale textual data, J. Mech. Des. 138 (6) (2016) 061403.
[21] A. Culotta, Towards detecting influenza epidemics by analyzing Twitter messages, in: Proceedings of the First Workshop on Social Media Analytics, New York, NY, USA, 2010, pp. 115–122.
[22] J. Ginsberg, M.H. Mohebbi, R.S. Patel, L. Brammer, M.S. Smolinski, L. Brilliant, Detecting influenza epidemics using search engine query data, Nature 457 (7232) (2009) 1012–1014.
[23] C.D. Corley, D.J. Cook, A.R. Mikler, K.P. Singh, Text and structural data mining of influenza mentions in web and social media, Int. J. Environ. Res. Public Heal. 7 (2) (2010) 596–615.
[24] P.M. Polgreen, Y. Chen, D.M. Pennock, F.D. Nelson, R.A. Weinstein, Using Internet searches for influenza surveillance, Clin. Infect. Dis. 47 (11) (2008) 1443–1448.
[25] C.C. Yang, H. Yang, L. Jiang, M. Zhang, Social media mining for drug safety signal detection, in: Proceedings of the 2012 International Workshop on Smart Health and Wellbeing, New York, NY, USA, 2012, pp. 33–40.
[26] S. Bhattacharya, H. Tran, P. Srinivasan, J. Suls, Belief Surveillance with Twitter, in: Proceedings of the 4th Annual ACM Web Science Conference, 2012, pp. 43–46.
[27] S. Bhattacharya, H. Tran, P. Srinivasan, Discovering health beliefs in Twitter, in: AAAI Fall Symposium: Information Retrieval and Knowledge Discovery in Biomedical Text, 2012.

[28] R. Winnenburg, A. Sorbello, A. Ripple, R. Harpaz, J. Tonning, A. Szarfman, H. Francis, O. Bodenreider, Leveraging MEDLINE indexing for pharmacovigilance – inherent limitations and mitigation strategies, J. Biomed. Inform. 57 (2015) 425–435.
[29] N. Collier, S. Doan, Syndromic Classification of Twitter Messages arXiv:1110.3094[cs]Oct. 2011.
[30] E. Aramaki, S. Maskawa, M. Morita, Twitter catches the flu: detecting influenza epidemics using Twitter, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing, Stroudsburg, PA, USA, 2011, pp. 1568–1576.
[31] T. Bodnar, V.C. Barclay, N. Ram, C.S. Tucker, M. Salathé, On the ground validation of online diagnosis with Twitter and medical records, in: Proceedings of the Companion Publication of the 23rd International Conference on World Wide Web Companion, 2014, pp. 651–656.
[32] N. Alvaro, M. Conway, S. Doan, C. Lofi, J. Overington, N. Collier, Crowdsourcing Twitter annotations to identify first-hand experiences of prescription drug use, J. Biomed. Inform. 58 (2015) 280–287.
[33] N. Slonim, N. Tishby, The power of word clusters for text classification, 23rd European Colloquium on Information Retrieval Research, vol. 1, 2001, p. 200.
[34] R. Artstein, M. Poesio, Inter-coder agreement for computational linguistics, Comput. Linguist. 34 (4) (2008) 555–596.
[35] S. Bashir, U. Qamar, F.H. Khan, IntelliHealth: a medical decision support application using a novel weighted multi-layer classifier ensemble framework, J. Biomed. Inform. 59 (2016) 185–200.
[36] S. Zhang, N. Elhadad, Unsupervised biomedical named entity recognition: experiments with clinical and biological texts, J. Biomed. Inform. 46 (6) (Dec. 2013) 1088–1098.
[37] M.T. Wiley, C. Jin, V. Hristidis, K.M. Esterling, Pharmaceutical drugs chatter on Online Social Networks, J. Biomed. Inform. 49 (2014) 245–254.
[38] Z. Huang, W. Dong, H. Duan, A probabilistic topic model for clinical risk stratification from electronic health records, J. Biomed. Inform. 58 (2015) 28–36.
[39] S. Poole, L.F. Schroeder, N. Shah, An unsupervised learning method to identify reference intervals from a clinical database, J. Biomed. Inform. 59 (2016) 276–284.
[40] M. Paoletti, G. Camiciottoli, E. Meoni, F. Bigazzi, L. Cestelli, M. Pistolesi, C. Marchesi, Explorative data analysis techniques and unsupervised clustering methods to support clinical assessment of Chronic Obstructive Pulmonary Disease (COPD) phenotypes, J. Biomed. Inform. 42 (6) (2009) 1013–1021.
[41] M. Brameier, C. Wiuf, Co-clustering and visualization of gene expression data and gene ontology terms for *Saccharomyces cerevisiae* using self-organizing maps, J. Biomed. Inform. 40 (2) (2007) 160–173.
[42] Z. Luo, M. Yetisgen-Yildiz, C. Weng, Dynamic categorization of clinical research eligibility criteria by hierarchical clustering, J. Biomed. Inform. 44 (6) (2011) 927–935.
[43] L. Cure, J. Zayas-Castro, P. Fabri, Clustering-based methodology for analyzing near-miss reports and identifying risks in healthcare delivery, J. Biomed. Inform. 44 (5) (2011) 738–748.
[44] S.J. Fodeh, C. Brandt, T.B. Luong, A. Haddad, M. Schultz, T. Murphy, M. Krauthammer, Complementary ensemble clustering of biomedical data, J. Biomed. Inform. 46 (3) (2013) 436–443.
[45] M. Dupuch, N. Grabar, Semantic distance-based creation of clusters of pharmacovigilance terms and their evaluation, J. Biomed. Inform. 54 (2015) 174–185.
[46] H.-S. Oh, Y. Jung, Cluster-based query expansion using external collections in medical information retrieval, J. Biomed. Inform. 58 (2015) 70–79.
[47] H.-T. Zheng, C. Borchert, H.-G. Kim, GOClonto: an ontological clustering approach for conceptualizing PubMed abstracts, J. Biomed. Inform. 43 (1) (2010) 31–40.
[48] T. Hao, A. Rusanov, M.R. Boland, C. Weng, Clustering clinical trials with similar eligibility criteria features, J. Biomed. Inform. 52 (2014) 112–120.
[49] K.R. Gøeg, R. Cornet, S.K. Andersen, Clustering clinical models from local electronic health records based on semantic similarity, J. Biomed. Inform. 54 (2015) 294–304.
[50] A. Wismüller, A. Meyer-Bäse, O. Lange, D. Auer, M.F. Reiser, D. Sumners, Model-free functional MRI analysis based on unsupervised clustering, J. Biomed. Inform. 37 (1) (2004) 10–18.
[51] S. Istephan, M.-R. Siadat, Unstructured medical image query using big data – An epilepsy case study, J. Biomed. Inform. 59 (2016) 218–226.
[52] F. Yang, T. Jiang, Cell image segmentation with kernel-based dynamic clustering and an ellipsoidal cell shape model, J. Biomed. Inform. 34 (2) (2001) 67–73.
[53] J. Lakoumentas, J. Drakos, M. Karakantza, G.C. Nikiforidis, G.C. Sakellaropoulos, Bayesian clustering of flow cytometry data for the diagnosis of B-Chronic Lymphocytic Leukemia, J. Biomed. Inform. 42 (2) (2009) 251–261.
[54] M. Korürek, A. Nizam, A new arrhythmia clustering technique based on ant colony optimization, J. Biomed. Inform. 41 (6) (2008) 874–881.
[55] J.A. Lara, D. Lizcano, A. Pérez, J.P. Valente, A general framework for time series data mining based on event analysis: application to the medical domains of electroencephalography and stabilometry, J. Biomed. Inform. 51 (2014) 219–241.
[56] J.H. Kim, I.S. Kohane, L. Ohno-Machado, Visualization and evaluation of clusters for exploratory analysis of gene expression data, J. Biomed. Inform. 35 (1) (2002) 25–36.
[57] K.Y. Yip, D.W. Cheung, M.K. Ng, K.-H. Cheung, Identifying projected clusters from gene expression profiles, J. Biomed. Inform. 37 (5) (2004) 345–357.
[58] B. Pontes, R. Giráldez, J.S. Aguilar-Ruiz, Biclustering on expression data: a review, J. Biomed. Inform. 57 (2015) 163–180.
[59] P. Radivojac, N.V. Chawla, A.K. Dunker, Z. Obradovic, Classification and knowledge discovery in protein databases, J. Biomed. Inform. 37 (4) (2004) 224–239.
[60] Q.T. Zeng, J.P. Pratt, J. Pak, D. Ravnic, H. Huss, S.J. Mentzer, Feature-guided clustering of multi-dimensional flow cytometry datasets, J. Biomed. Inform. 40 (3) (2007) 325–331.
[61] G. Milani, C. Masciullo, C. Sala, R. Bellazzi, I. Buetti, G. Pistis, M. Traglia, D. Toniolo, C. Larizza, Computer-based genealogy reconstruction in founder populations, J. Biomed. Inform. 44 (6) (2011) 997–1003.
[62] X. Wang, D. Zeng, H. Seale, S. Li, H. Cheng, R. Luan, X. He, X. Pang, X. Dou, Q. Wang, Comparing early outbreak detection algorithms based on their optimized parameter values, J. Biomed. Inform. 43 (1) (2010) 97–103.
[63] D.L. Buckeridge, H. Burkom, M. Campbell, W.R. Hogan, A.W. Moore, Algorithms for rapid outbreak detection: a research synthesis, J. Biomed. Inform. 38 (2) (2005) 99–113.
[64] Y. Li, S. Swift, A. Tucker, Modelling and analysing the dynamics of disease progression from cross-sectional studies, J. Biomed. Inform. 46 (2) (2013) 266–274.
[65] A.V. Carreiro, P.M.T. Amaral, S. Pinto, P. Tomás, M. de Carvalho, S.C. Madeira, Prognostic models based on patient snapshots and time windows: predicting disease progression to assisted ventilation in Amyotrophic Lateral Sclerosis, J. Biomed. Inform. 58 (2015) 133–144.
[66] T. Tran, T.D. Nguyen, D. Phung, S. Venkatesh, Learning vector representation of medical objects via EMR-driven nonnegative restricted Boltzmann machines (eNRBM), J. Biomed. Inform. 54 (2015) 96–105.
[67] Y. Lu, P. Zhang, J. Liu, J. Li, S. Deng, Health-related hot topic detection in online communities using text clustering, PLoS One 8 (2) (2013) e56221.
[68] A. Nikfarjam, A. Sarker, K. O'Connor, R. Ginn, G. Gonzalez, Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features, J. Am. Med. Inform. Assoc. 22 (3) (2015) 671–681.
[69] M. Yang, M. Kiang, W. Shang, Filtering big data from social media – building an early warning system for adverse drug reactions, J. Biomed. Inform. 54 (2015) 230–240.
[70] M. Yang, X. Wang, M.Y. Kiang, Identification of consumer adverse drug reaction messages on social media, in: PACIS, 2013, p. 193.
[71] API Overview | Twitter Developers, Twitter. [Online]. <https://dev.twitter.com/overview/api> (accessed: 24-Jan-2016).
[72] M.A. Russell, Mining the Social Web: Data Mining Facebook, Twitter, LinkedIn, Google+, GitHub, and More, O'Reilly Media Inc., 2013.
[73] J.A. Danowski, Network analysis of message content, Prog. Commun. Sci. 12 (1993) 198–221.
[74] M. Thelwall, K. Buckley, G. Paltoglou, D. Cai, A. Kappas, Sentiment strength detection in short informal text, J. Am. Soc. Inform. Sci. Technol. 61 (12) (2010) 2544–2558.
[75] Medical Dictionary, MedlinePlus Merriam-Webster. [Online]. <http://c.merriam-webster.com/medlineplus/symptom> (accessed: 24-Jan-2016).
[76] K.R. Emmett, Nonspecific and atypical presentation of disease in the older patient, Geriatrics 53 (2) (1998) 50–52.
[77] DSM-5 American Psychiatric Association, A Bayesian biosurveillance method that models unknown outbreak diseases, in: Intelligence and Security Informatics: Biosurveillance, American Psychiatric Publishing, Arlington, 2013.
[78] R. Engelbrecht, second ed., Connecting Medical Informatics and Bio-informatics: Proceedings of MIE2005: The XIXth International Congress of the European Federation for Medical Informatics, vol. 116, IOS Press, 2005.
[79] WebMD - Better information. Better health, WebMD. [Online]. <http://www.webmd.com/default.htm> [accessed: 14-Jul-2015).
[80] Symptoms, Mayo Clinic. [Online]. <http://www.mayoclinic.org/symptoms> (accessed: 24-Mar-2016).
[81] Symptoms, MedlinePlus. [Online]. <https://www.nlm.nih.gov/medlineplus/symptoms.html> (accessed: 24-Mar-2016).
[82] WordNet, WordNet. [Online]. <http://wordnet.princeton.edu/> (accessed: 24-Jan-2016).
[83] Consumer Health Vocabulary (CHV), CHV Wiki. [Online]. <http://consumerhealthvocab.chpc.utah.edu/CHVwiki/> (accessed: 24-Jan-2016).
[84] S. Rude, E.-M. Gortner, J. Pennebaker, Language use of depressed and depression-vulnerable college students, Cogn. Emot. 18 (8) (2004) 1121–1133.
[85] M. De Choudhury, S. Counts, E.J. Horvitz, A. Hoff, Characterizing and predicting postpartum depression from shared facebook data, 2014, pp. 626–638.
[86] RT this: OUP Dictionary Team monitors Twitterer's tweets, OUPblog.
[87] B. O'Connor, M. Krieger, D. Ahn, TweetMotif: exploratory search and topic summarization for Twitter, in: ICWSM, 2010.
[88] G.A. Miller, M. Chodorow, S. Landes, C. Leacock, R.G. Thomas, Using a semantic concordance for sense identification, in: Proceedings of the Workshop on Human Language Technology, 1994, pp. 240–243.
[89] K. Tian, M. Revelle, D. Poshyvanyk, Using latent dirichlet allocation for automatic categorization of software, in: Mining Software Repositories, 2009. MSR'09. 6th IEEE International Working Conference on, 2009, pp. 163–166.
[90] C. Fox, A stop list for general text, ACM SIGIR Forum, vol. 24, 1989, pp. 19–21.
[91] M.F. Porter, An algorithm for suffix stripping, Program 14 (3) (1980) 130–137.
[92] P.-N. Tan, S. Michael, V. Kumar, Chapter 6: association analysis: basic concepts and algorithms, in: Introduction to Data Mining, Addison-Wesley, 2005, pp. 327–414.