BIG DATA MINING AND ANALYTICS
ISSN 2096-0654 05/08 pp273–287
Volume 2, Number 4, December 2019
DOI: 10.26599/BDMA.2019.9020012

Tweetluenza: Predicting Flu Trends from Twitter Data

Balsam Alkouz, Zaher Al Aghbari*, and Jemal Hussien Abawajy

Abstract: Health authorities worldwide strive to detect Influenza prevalence as early as possible in order to prepare for it and minimize its impacts. To this end, we address the Influenza prevalence surveillance and prediction problem. In this paper, we develop a new Influenza prevalence prediction model, called Tweetluenza, to predict the spread of the Influenza in real time using cross-lingual data harvested from Twitter data streams with emphases on the United Arab Emirates (UAE). Based on the features of tweets, Tweetluenza filters the Influenza tweets and classifies them into two classes, reporting and non-reporting. To monitor the growth of Influenza, the reporting tweets were employed. Furthermore, a linear regression model leverages the reporting tweets to predict the Influenza-related hospital visits in the future. We evaluated Tweetluenza empirically to study its feasibility and compared the results with the actual hospital visits recorded by the UAE Ministry of Health. The results of our experiments demonstrate the practicality of Tweetluenza, which was verified by the high correlation between the Influenza-related Twitter data and hospital visits due to Influenza. Furthermore, the evaluation of the analysis and prediction of Influenza shows that combining English and Arabic tweets improves the correlation results.

Key words: Twitter data analysis; Influenza forecasting; prediction using social media; social media mining

1 Introduction

Seasonal Influenza is an acute respiratory infection caused by Influenza viruses capable of human to human transmission. It is one of the serious public health problems worldwide causing severe illnesses to people of any age group with estimated annual mortality rate of 250 000 to 500 000 worldwide[1]. The estimated 20–40 million causalities of the Influenza pandemic commonly known as "Spanish flu" and the recent variants of the Influenza virus such as "SARS" and "H1N1" that resulted in casualties numbering in millions highlight the impact of Influenza pandemic on the society. Apart from social impact, the economic impact on individuals and businesses (e.g., productivity losses due to absenteeism) and the costs associated with the interventions to treat Influenza are serious concerns for public health officials worldwide.

Early detection of Influenza prevalence is crucial to reduce the social and economic impacts of seasonal Influenza. This requires capabilities to monitor and predict the appearance and spread of seasonal Influenza in the population[2]. Traditional Influenza surveillance systems collect data from clinical diagnoses. Unfortunately, these systems are almost manual, resulting in long delays for clinical data acquisition. However, public health authorities need to forecast Influenza breakout as early as possible to ensure effective preventive measures, leading to an increasing need for efficient sources of data for prediction. To this end, accurate Influenza prevalence surveillance (i.e., to detect Influenza rates in real time) and prediction are very important for public health authorities to take preventive measures prior to the start of a seasonal Influenza. The conventional approaches that use historical Influenza-like illness datasets are

• Balsam Alkouz and Zaher Al Aghbari are with the Department of Computer Science, University of Sharjah, Sharjah 27272, UAE. E-mail: balkouz@sharjah.ac.ae; zaher@sharjah.ac.ae.
• Jemal Hussien Abawajy is with the Department of Science, Engineering and Built Environment, Deakin University, Melbourne 3125, Australia.
∗ To whom correspondence should be addressed.
Manuscript received: 2019-01-08; revised: 2019-05-12; accepted: 2019-05-22

susceptible to a high degree of error rates as they are often restricted by the time required to gather up-to-date and accurate datasets. The Google Flu Trends (GFT) system estimates an Influenza infection rates based on users search patterns. It has been shown that GFT improves the error rates associated with the conventional methods[3]. However, studies have shown that GFT still suffers from shortcomings in terms of accuracy[4]. This necessitates a better surveillance and prediction of Influenza prevalence than the exiting approaches.

Online social media platforms have the potential to improve the time lag in traditional Influenza surveillance systems. Twitter, a very widely spread microblogging service, allows registered users to freely express their feelings and share their health conditions with the wider audiences via short messages called tweets. It has been shown that Twitter streams are rich of immediate health related information[5–8]. For example, flu-related tweets such as "stuck at home with the flu" are common[9] and a recent study has correlated tweets with Influenza rates in the USA as well as with the traditional sentinel Influenza-like illness rates surveillance reports[10]. With an average of 328 million active users per month[8], Twitter users generate huge volume of Twitter streams that make it useful for tracking or even predicting different behaviors including diseases like Asthma[5] and Influenza[7]. The user-generated real-time nature of tweets makes them attractive as a tool for tracking a pandemic such as Influenza. It also enables the public health authorities to communicate the breakout of Influenza to their citizens at the earliest time to ensure effective preventive intervention. Hence, some of the exiting research works have analyzed the Twitter streams to infer the spread of a disease, like Influenza[6], or predict future status of a disease[7]. By facilitating the availability of millions of real-time user tweets that include geographical location and personal well-being information[9], Twitter has become a powerful tool in the battle to predict epidemics as early as possible. As a result, social media driven prediction of different diseases have received unprecedented attention among the research communities as a promising tool for tracking a pandemic.

In this paper, we develop a new cross-lingual Influenza prevalence prediction model to predict the spread of the Influenza in real time using cross-lingual data harvested from Twitter data streams with emphases on the United Arab Emirates (UAE). Health authorities in the UAE recommend that everyone should get vaccinated against Influenza during its season, especially those at high risk of Influenza-related complications, or those caring for people who are at high risk[11]. It is essential for local authorities to be able to detect and predict communicable and non-communicable illnesses such as Influenza. This will help local authorities in the UAE improve their preparedness to combat Influenza outbreaks. The problem of Influenza detection and prediction in countries such as the UAE brings two main challenges. First, the constantly changing weather conditions and high amount of dust. Second, with a large number of English speaking expatriates tweeting in English and a variety of Arabic dialect speaking residents tweeting differently, exploiting multilingual and multidialectal tweets poses serious challenges to harvest Twitter data streams to detect and forecast Influenza. The work in Ref. [12] contains preliminary results of analyzing only Arabic tweets. However, this paper proposes a system called Tweetluenza that supports multilingual and multidialectal tweets for the purpose of Influenza prediction. To the best of our knowledge, there is no previous research work that used multilingual and multidialectal tweets for the prediction of Influenza prevalence. Our contributions can be summarized as follows:

(1) A system called Tweetluenza that is able to predict the Influenza outbreak using Twitter data. Tweetluenza is multilingual and multidialectal system.

(2) An algorithm for the detection and prediction of early Influenza activity around the country. The algorithm reliably classifies Arabic and English tweets on Influenza related keywords. The classifier is context-aware, where words of the same root but with different meanings that do not represent Influenza are removed.

(3) Comprehensive analysis of the effects of different types of reporting tweets on the correlation between Twitter data and the hospital visits. Predicting the number of future Influenza-related hospital visits is of paramount importance to healthcare authorities, since Influenza disease can affect everyone.

(4) Real-data evaluation of the detection and prediction results of Tweetluenza against a ground truth data (records of the actual Influenza-related hospital visits of the UAE Ministry of Health).

The rest of the paper is organized as follows: the related work is presented in Section 2. The problem

overview is defined in Section 3. Section 4 discusses the proposed Tweetluenza system. In Section 5, we analyze the experimental results. Section 6 presents the concluding remarks of the paper.

2 Related Work

There are several studies showing that Twitter streams can yield useful public health information and can be harvested to monitor public health trends. As a result, Twitter-based Influenza detection and forecast have drawn increasing attention recently.

Twitter being a famous micro blogging service allows researchers to rely on the human sensors, i.e., Twitter users, to detect different kinds of events. Lots of work in the literature deal with general events detection, i.e., solutions that can be applied to different events. Guille and Favre[13] have discussed an approach to detect and track general events using user mentions. Detection of specific events like carnivals[14] and road traffic[15] were explored. Capdevila et al.[16] proposed a probabilistic model that facilitates the extraction of interesting events from Twitter datasets based on the geographic location, time, and tweets.

To improve the preparedness of emergency departments, systems that forecast the spread of illnesses were developed. Asthma-related number of visits to emergency departments of hospitals were predicted using Twitter data, Google search interests, and environmental sensor data[5].

Detecting the outbreak of communicable diseases such as Influenza from social media data, e.g., tweets, was investigated by several researchers. Some researchers worked on the analysis of tweets and others used tweets to predict future visits to hospitals' emergency departments. The system in Ref. [6] classified the collected spatio-temporal data into four classes to be able to detect and forecast a disease. In Ref. [17], Influenza activities were detected by investigating the temporal and spatial features of the collected data steams. The work in Ref. [18] utilized National Language Processing (NLP) techniques to distinguish the Influenza reporting tweets from other tweets. Smith et al.[19] proposed a real-time surveillance system for disease awareness during the spread of Influenza epidemic. Broniatowski et al.[20] presented an analysis of the tweets collected during the 2012 – 2013 Influenza season. Byrd et al.[4] proposed a system that utilizes machine learning tools to predict the disease-related sentiment of tweets; however, the sentiments are based on the extracted keywords. The work in Ref. [21] proposed a mechanistic model that uses geo-localized data from Twitter to quantify relevant occurrences of Influenza. Zhang et al.[22] proposed a method to detect and analyze Influenza trends in China. They used a Support Vector Machine (SVM) classifier to detect Influenza posts from TencentWeibo, which is a popular Chinese microblogging social media similar to Twitter. Brennan et al.[23] used geo-tagged tweets of travelling Twitter users to explore how individuals contribute to the global spread of diseases such as Influenza by inferring properties of the flow of Twitter users between cities.

Predicting hospital visits due to Influenza was investigated using the patterns of search keywords of users that are found in Google Flu Trends[20]. Another approach utilized the hospital visits taken from the Centers for Disease Control (CDC) data along with social media data to forecast the future hospital visits by people suffering from Influenza like illness[24]. In Ref. [7], the authors represented each tweet with a 20-attribute feature vector and then these vectors are used to classify tweets into reporting and others. Machine learning tools, such as SVM, are then used to process the classified tweets to forecast Influenza-related hospital visits. Machine learning tools were also used by the system in Ref. [25] to predict number of Influenza-related hospital visits. Others extracted topics from tweets to enhance seasonal Influenza surveillance[26]. The system in Ref. [27] grouped the collected Influenza-related posts into three classes that depict the phases of Influenza activity. These classes are then used to forecast the expansion of the disease. Moreover, statistical models, such as auto-regression, on CDC data were used to forecast the spread of Influenza[28]. Detection and prediction of Influenza using tweets of a Korean language were explored by Refs. [29, 30]. Similarly, Lee et al.[31] proposed a multilayer perceptron with back propagation model to predict flu activities from Twitter streams. Ghosh et al.[32] presented a spatio-temporal model to predict rare disease outbreaks from online posts. Effective mining of patient generated wellness data from Twitter data streams was proposed by Akbari et al.[33] to provide actionable insight into the wellness of individuals.

All the above presented studies, which addressed the issue of Influenza surveillance and forecasting using Twitter streams, have mostly focused on the English language.

3 Problem Overview

In this paper, we address the problem of cross-lingual Influenza prevalence prediction in real-time with emphases on the UAE. There are many studies showing that Twitter streams can yield useful public health information and the tweets can be harvested to monitor public health trends. With UAE ranking fourth in the world where Internet users spend the most time on social media[34] and the third among Twitter users in the Arab countries between the years 2014 and 2017[35], it makes sense to harvest Twitter data streams to predict the spread of the Influenza in real time.

Our aim is to develop an Influenza detection and predication system based on Twitter streams that will complement the conventional epidemiological surveillance systems. In particular, we want to develop a system that can answer the following question:

- How can we use Twitter streams for Influenza surveillance and prediction in a cross-lingual and cross-dialect settings?

Twitter is often used by people to share information such as personal health like flu symptoms or recovery from flu. Although there are many research works that exploit Twitter streams for surveillance and prediction, these previous studies mostly focused on the English language. There is no work for assessing the potential usefulness of Twitter streams for Influenza surveillance and prediction in a cross-lingual and cross-dialect settings. Exploiting cross-lingual and cross-dialect Twitter streams requires quite different and involved data preprocessing, feature extraction and selection, and analysis and modeling than the English language only approaches. We want to close this gap by harnessing cross-lingual and cross-dialect Twitter streams for developing Influenza surveillance and forecasting system. Some challenges of processing Arabic text have pointed out in Refs. [36, 37].

Another important desired outcome of our research is how we can use Twitter streams to detect and predict Influenza to notify individuals and medical centers. Specifically, we want to answer the following question:

- How can we generate predictions of expected numbers of visits due to Influenza based on Twitter streams?

We need to reliably assess the accuracy levels of the real-time Influenza prediction model created within the UAE environment. This is important because predictions based on multilingual and multidialectal data can overestimate or underestimate the actual count of Influenza-related hospital visits. Fortunately, all hospitals within the UAE monitor and record patient visits due to Influenza. We will validate the accuracy of the developed model with the actual count of Influenza-related hospital visits within the UAE as a baseline.

Given the stream of English and Arabic Raw Tweets (RTs), from Twitter users in the UAE, our goal is to propose a system that filters and classifies the tweets into Influenza reporting and non-reporting tweets. The proposed system should utilize the reporting tweets to detect Influenza growth. Furthermore, the proposed system should be able to predict the Influenza-related counts of future hospital visits by learning from the patterns of previous weeks' visits and collected reporting tweets.

4 Tweetluenza System Architecture

The architecture of Tweetluenza is shown in Fig. 1. The main modules are data collection, data processing, tweets classification, and prediction. Note that each of these modules is process tweets of both languages, which is the main challenge in this work. These modules are explained in details below. Table 1 serves as a reference for the meanings of the main symbols used in this paper.

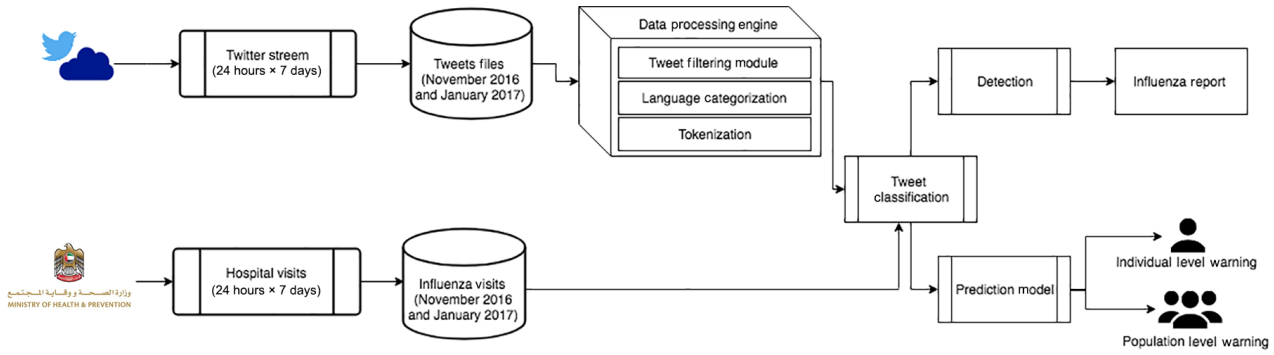

Fig. 1 Tweetluenza system architecture.

Table 1 Meaning of symbols.

Symbol	Meaning
RT	Raw tweet
CT	Cleaned tweet
CT_{jl}^{URL}	Tweet after filtering URLs
$CT_{jl}^{@}$	Tweet after filtering user mentions
CT_{jl}^{DP}	Tweet after filtering digits and punctuation
CT_{jl}^{E}	Tweet after filtering emojis
CT_{jl}^{SW}	Tweet after filtering the stop words
CT_{S}^{E}	Tweet after stemming
SR	Self reporting tweets
NSR	Nonself reporting tweets

The purpose of the system is shown as follows. First, collect tweets, called *RTs*, from the Twitter stream. Then, the *RTs* are processed in three steps: (1) The *RTs* are filtered to retrieve Influenza-related tweets from the UAE region, (2) language categorization, and (3) filtered tweets are tokenized into words. Next, each processed tweet is assigned an appropriate class label indicating whether a tweet is self-reporting, nonself-reporting, or non-reporting. The growth of Influenza can then be monitored by analyzing the reporting tweets. Then, to predict the Influenza-related count of future hospital visits, a linear regression model is applied on the reporting tweets. Below we discuss the tasks of each module of Fig. 1.

4.1 Data collection

Using the Twitter Application Programming Interface (API) platform to collect streaming realtime tweets coming from the UAE, the collected tweets span the period of two months. In our implementation, we used the python library[38] (Tweepy Stream API) to collect the tweets. The geographical location of a tweet is represented by GeoJson coordinates format in the tweet's metadata. Therefore, to retrieve the tweets from the UAE region, we filtered the tweets based the UAE location coordinates. Although UAE is an Arab country, in which Arabic is the official language, about 88.5% of its residents are expats, who communicate mostly in English[39]. Therefore, Tweetluenza collected both Arabic and English tweets.

To serve as ground truth, the Influenza-related count of hospital visits to government hospitals and medical centers in the city of Sharjah, UAE, were obtained from the UAE Ministry of Health for the same two months. That is because the city of Sharjah hosts some of the largest government hospitals and represents the UAE general population well.

4.2 Data processing

The second module, data processing engine, is composed of three submodules, which are tweet filtering, language categorization, and tokenization. The collected raw tweets, *RTs*, are processed by these submodules to transform them to cleaned tweets, *CTs*. The sequence of cleaning the text of tweets, categorizing tweets based on their language, and extracting tokens from text of tweets is shown in Fig. 2.

Below we discuss the details of each submodule.

4.2.1 Tweet filtering

A raw tweet generally includes metadata such as tweet geolocation, time, and user ID. In addition to the text of a tweet, other elements may exist such as Uniform Resource Locators (URLs) and emojis. Therefore, RTs must undergo a filtering process to remove the non-text elements, which are (1) URLs, (2) user mentions, (3) digits and punctuations, and (4) emojis. This filtering process is language independent. The filtered tweets are fed into the language categorization submodule. Afterwards, the tweets undergo further language-dependent filtering.

(1) *URLs*: Links to websites, images, and videos are filtered (removed) from *RTs*. We denote the *j-th URL-filtered CT* as $CT_{j}^{URL} = \{CT_{j1}^{URL}, \ldots, CT_{jl}^{URL}, \ldots, CT_{jL_{j,f}}^{URL}\}$, where CT_{jl}^{URL} is the l-th token that is URL-filtered and $L_{j,f}$ is the total number of tokens after filtering URL.

(2) *User mentions*: Users may mention other users in order to communicate and link the mentioned user to their tweets. Quoted tweets also contain mentions of the original author. All user mentions are in

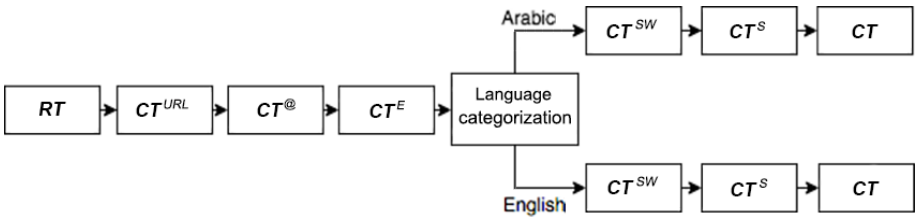

Fig. 2 Sequence of data processing steps

the format *@user* and are removed from *RTs*. We denote the j-th user-mention filtered tweet as $CT_j^{@} = \{CT_{j1}^{@}, \ldots, CT_{jl}^{@}, \ldots, CT_{jL_{j,u}}^{@}\}$, where $CT_{jl}^{@}$ is the l-th token that is *user-mention*-filtered and $L_{j,u}$ is the total number of tokens after filtering user-mentions.

(3) *Digits and punctuations*: Number characters and punctuation do not provide any useful information to the aim of Tweetluenza, and hence, they are removed. We denote the j-th digits/punctuation filtered tweet as $CT_j^{DP} = \{CT_{j1}^{DP}, \ldots, CT_{jl}^{DP}, \ldots, CT_{jL_{j,d}}^{DP}\}$, where CT_{jl}^{DP} is the l-th token from which digits/punctuations are removed and $L_{j,d}$ is the total number of tokens after filtering digits/punctuations.

(4) *Emojis*: The same emoji can be used for different purposes, and so their existence in the tweets adds no value to the analysis and predictions of Influenza; hence, they are filtered. We denote the j-th *emoji*-filtered as $CT_j^{E} = \{CT_{j1}^{E}, \ldots, CT_{jl}^{E}, \ldots, CT_{jL_{j,e}}^{E}\}$, where CT_{jl}^{E} is the l-th token that is *emoji*-filtered and $L_{j,e}$ is the total number of tokens after filtering emojis.

4.2.2 Language categorization

After filtering, tweets are categorized into four groups depending on their language: Arabic, English, combined language (English and Arabic), and others. This step was performed because Twitter's language attribute in the raw tweets is not accurate, as it considers a fully Arabic tweet as English if it contains a user mention or a link. To overcome this difficulty, we developed our own language categorizer tool for Tweetluenza.

4.2.3 Tokenization

The text of the Arabic and English tweets are separately parsed to remove the stop words of each language. Then, the words of each tweet are stemmed. As a result, the text of tweets is tokenized. The following steps are applied to both Arabic and English datasets.

(1) *Stop words removal*: Arabic (i.e., إذا، أن، هو) and English (i.e., and, the, of, etc.) stop words are removed from the filtered tweets, CT^{SW}, since their existence does not add a value to prediction. We denote the j-th stop word filtered tweet as $CT_j^{SW} = \{CT_{j1}^{SW}, \ldots, CT_{jl}^{SW}, \ldots, CT_{jL_{j,w}}^{SW}\}$, where CT_{jl}^{SW} is the l-th token that does not belong to the set of Arabic and English stop words and $L_{j,w}$ is the total number of tokens after filtering the stop words.

(2) *Stemming*: Words in tweets are stemmed so that different variations of a single word will have the same root word S. For example, the English variations *sneeze*, *sneezed*, and *sneezing* are stemmed to *sneez*; and the Arabic variations عطست، تعطس، يعطسون، يعطس are stemmed to عطس. The Natural Language Toolkit (NLTK) library[38] was used for stemming both Arabic and English tweets. We denote the j-th stemmed CT as $CT_j^{S} = \{CT_{j1}^{S}, \ldots, CT_{jl}^{S}, \ldots, CT_{jL_{j,s}}^{S}\}$, where CT_{jl}^{S} is the l-th token that has been stemmed with Arabic and English stemmers and $L_{j,s}$ is the total number of stemmed tokens. Examples of English and Arabic tweets before and after stemming are shown in Tables 2 and 3, respectively. The first left columns in Tables 2 and 3 show the text of a filtered tweet and the second left columns show the tweet after being tokenized. The right columns in Tables 2 and 3 are the class label of a tweet (see Section 4.3). In Table 3, the English translation of the Arabic tweet is included.

4.3 Classification of tweets

To classify the tweets, we developed a classifier, called *repTweets*. The classifier uses the keywords that represent the Influenza disease or the symptoms of Influenza for both Arabic and English. Examples of the language-specific keywords, *KWs*, are shown in Table 4. The set of cleaned tweets, *CTs*, are fed into the *repTweets* classifier to distinguish *reporting* tweets from *non-reporting* tweets. The reporting tweets are further classified into *self* and *nonself* reporting. Moreover, *repTweets* is context-aware and thus removes

Table 2 Examples of stemmed English tweets.

Tweet after filtering	Tweet after stemming	Class
I'm sorry to find out you got the swine flu	Im sorri to find out you got the swine flu	Self
i sneezed so hard i thought my brain was gonna explode	i sneez so hard i thought my brain wa gonna explod	Self
Body hurts, not sure if it's from the lack of exercise or I'm having a fever	Bodi hurts, not sure if it from the lack of exercis or Im have a fever.	Non-self
Weather changing fever and bad cold	Weather changing fever and bad cold	Non-self
tablets are the future they said They will kill desktops they said	tablet are the future they said They will kill desktops they said	Spam
Good morning Dubai Dubai Marina Yacht Bay	Good morn Dubai Dubai Marina Yacht Bay	Spam

Table 3 Examples of stemmed Arabic tweets.

Tweet after filtering	Tweet after stemming	English translation	Class
اللوز كبرها بتطلع حلقي وربي	لوز كبر طلع حلق ورب	I feel my throat is coming out (expressing pain)	Self
ماشاءالله هم يعطسون عطوهم اجازه	ماشاءالله هم عطس عطو جزه	They sneeze and they get a holiday	Self
أقل درجة حرارة سجلت درجة جبل جيس	اقل درجة حرارة سجل جبل درج جيس سعة صبح	The minimum temperature registered today is in Jees mountain	Non-self
العطس يطلع صوت	عطس طلع صوت	Sneeze produces a loud voice	Non-self
انا فهمت بس يالسه استغبى	انا فهم بس يلس غبى	I understood but I am acting dumb	Spam
استغفرك ربي وأتوب إليك	غفر ربي أتب الك	Forgive me God	Spam

Table 4 Examples of Influenza related keywords.

	Keyword
English	Flu, Influenza, cold, sneeze, fever, sore, dry, throat, fatigue, virus, germs, runny nose, respiration, mucus, sinusitis
Arabic	انفلونزا، زكام، سخونة، عطس، حلق، لوز، فايروس جراثيم، سيلان، رشح، احتقان، حرارة، التهاب، جيوب

expressions that contain the keywords but are non-representative of Influenza.

The *repTweets* classifier excludes expressions that contain the keywords but are non-representative of Influenza. Examples of these non-representative Influenza expressions are shown in Tables 4 and 5, such as cold hearted or anti-virus. As shown in the *repTweet* algorithm (Algorithm 1), a cleaned tweet, CT_i, is considered *self-reporting,* if it satisfies the two conditions. (1) After removing all non-representative keywords, NKW_j (see examples in Tables 6 and 7), the tweet still contains an Influenza-related keyword, KW_j, from the set shown in Tables 4 and 5, and (2) it contains a personal pronoun (Table 8). However, if the cleaned tweet CT_i satisfies condition (1), but not condition (2), then it is considered as a *nonself-reporting* tweet. But,

Table 5 Examples of Influenza keywords in the UAE dialect.

UAE dialect	Standard Arabic	English translation
مزجم	مزكم	Have a flu
زجام	زكام	Flu

Algorithm 1 repTweet

Input: *CTs*
Result: classified tweets
1 foreach $CT_i \in CTs$, i=1,...,n **do**
2 *removeNKW(CT_i)*
3 **if** CT_i *contains* $KW_j \mid KW_j \in$ *set of keywords* **then**
4 **if** *a personal pronoun exists in* CT_i **then**
5 CT_i assigned to *self* reporting class;
6 **else**
7 CT_i assigned to *nonself* reporting class
8 **end**
9 **else**
10 CT_i assigned to *non-reporting* class
11 **end**
12 **end**

Table 6 English words with multiple meaning based on contexts.

Root word	Different forms
Flu	Influence, influential, flute, flutes, fluency, flurry, mellifluous, flutter, fluent, fluorine, fluorescent
Virus	Anti-virus
Fever	Fever media
Cold	Cold hearted, cold blooded, cold shoulder, cold fish, coldplay/cold play (band name)

Table 7 Arabic words with multiple meanings based on context.

Root word	Different forms	English translation
رشح	ترشيح مُرشح	Nominate/Nominee
عرق	عراق	Iraq
برد	الجو بارد	Cold weather
حلق	حلقة، حلقات	Episode/Episodes

Table 8 Personal pronouns in English and Arabic.

	Personal pronoun
English	I, me, my, mine, you, your, yours, he, she, it, him, her, his, hers, its, we, us, our, ours, you, your, yours, they, them, their, theirs
Arabic	أنا، نحن، أنت، أنتم ، أنتم ، أنتن، هو، هي، هما، هم، هن، هؤلاء، هذا، هذه، هذان

if the cleaned tweet CT_i does not satisfy condition (1), it is considered as *non-reporting* tweet.

Examples of the Influenza-related standard Arabic keywords are shown in Table 4. However, the UAE dialect is somehow different from the standard Arabic. Examples of Influenza-related word differences between UAE dialect and standard Arabic are shown in Table 5. Therefore, *repTweet* classifier was designed to detect the Influenza-related local UAE dialect words in the captured tweets. The Influenza-related UAE dialect words are included to the set of keywords used for tweet classification.

In the Arabic language, some words have the same root (stem) as some Influenza keywords, but have different contextual meanings. For example, حلق ـ حلقة can mean throat or a series episode. Such word, or expressions (see Table 7) pose a challenge to accurate keyword detection from tweets, which usually are misclassified as Influenza reporting tweets. Similarly, some English words or expressions after being stemmed have the same root as the Influenza keywords, but have different contextual meanings. Tables 6 and 7 show examples of these words and expressions in English and Arabic, respectively. Therefore, we have added the list of such words and expressions (Tables 6 and 7) to our classifier to be overlooked during classification.

Influenza-related tweets are assigned to one of the following classes:

• **Reporting**: It contains tweets that report Influenza occurrences, i.e., tweets that contain Influenza-related keywords. Then, the tweets labeled as reporting undergo further classification into *self* and *nonself* reporting subclasses:

(1) Self: This subclass contains tweets that are reporting an Influenza disease experience of the user. Tweets are classified as *self* reporting based on the following criteria: (a) it contains at least one Influenza-related keyword (not a non-representative Influenza word), and (b) it contains a personal pronoun. An example of a *self* reporting English tweet is *"I'm sorry to find out I got the swine flu"*. In this tweet, the keyword is *flu* and the personal pronoun is *I*. An example of Arabic tweet is ماشاءالله هم يعطسون عطوهم اجازة. In this Arabic tweet, the keyword is يعطسون that means *are sneezing* and it has a personal pronoun هم that means *they*. Table 8 shows the personal pronouns used in English and Arabic.

(2) Nonself: This subclass contains tweets reporting general news about Influenza. A tweet is classified as *nonself* reporting if (a) it contains at least one Influenza-related keyword (not a non-representative Influenza word), and (b) it does not contain a personal pronoun. For example, *"Weather changing, fever and bad cold"* is classified into the *nonself* reporting class. This tweet contains the keywords *fever* and *cold*, but no personal pronoun. Another example, أقل درجة حرارة سجلت درجة جبل جيس is classified into the *nonself* reporting class. This Arabic tweet does not contain a personal pronoun and it has the keyword حرارة, which means *temperature*.

• **Non-reporting**: A tweet is classified as non-reporting if it matches one of the following conditions: (1) it does not contain any Influenza-related keyword, or (2) it contains words or expression, that are not representative of Influenza or its symptoms, but whose roots match the Influenza-related keywords. For example, the tweet *"You are a cold hearted person"* has the *cold hearted* expression that has nothing to do with Influenza although the individual word *cold* is related to Influenza. Another example, استغفرك ربي وأتوب إليك does not have any keyword and is considered *non-reporting*.

For the purpose of measuring the accuracy of *repTweet* classifier, three human subjects were employed to manually annotate the tweets of each language. That is, each annotator assigned class labels to all tweets in both languages. These annotated tweets served as ground truth and were compared with the class assignments of *repTweet*. Table 9 shows the comparison results. Note that the analysis of the classification results will be discussed in the next section.

We conducted a k-fold cross validations, where $k = 10$, to obtain the classification results and remove any bias of the results. That is, about 50% are *non-reporting*

Table 9 Accuracy of classification by repTweets.

		Accuracy (%)		
English	Self	**81.0**	14.3	4.7
	Nonself	23.1	**76.9**	0
	Non-reporting	17.0	18.9	**64.1**
Arabic	self	**93.8**	6.2	0
	Nonself	36.0	**64.0**	0
	Non-reporting	3.4	25.4	**71.2**

tweets and the rest is divided between *self* and *nonself* tweets. We considered the tweets that are classified as true positives to compute the classification accuracy of the classes (shown in bold in Table 9). For the English tweets, the accuracy for *self* and *nonself* are about 81% and 77%, respectively, while the accuracy of *non-reporting* is about 64%. Similarly, the accuracies of the Arabic *self* and *nonself* are about 94% and 64%, respectively, while the accuracy of *non-reporting* is about 71%.

4.4 Analysis

For each language (Arabic and English), the Influenza-related tweets of each class are counted for every week. Also, the Influenza-related tweet counts per week of the combination of Arabic and English tweets for each class are computed. To investigate the correlation between the Twitter data and the actual number of Influenza-related hospital visits, we compared the weekly counts of each class with the actual weekly Influenza-related counts of hospital visits represented in Figs. 3 – 11 as time series. These time series are normalized to a mean of zero and a standard variation of one before comparing and plotting them on a graph.

To compare the behavior of the tweet counts with and the actual hospital visits counts, we computed correlation between the two time series using

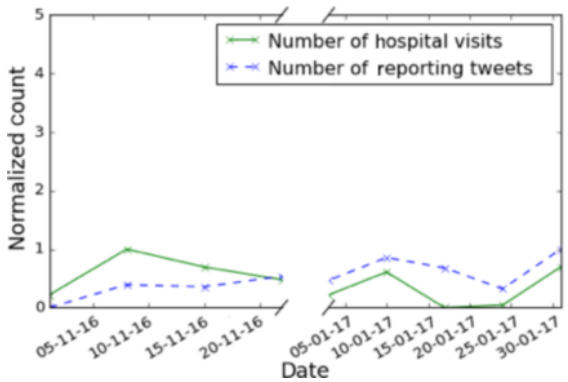

Fig. 3 Count of English tweets (classified as self) versus count of actual hospital visits.

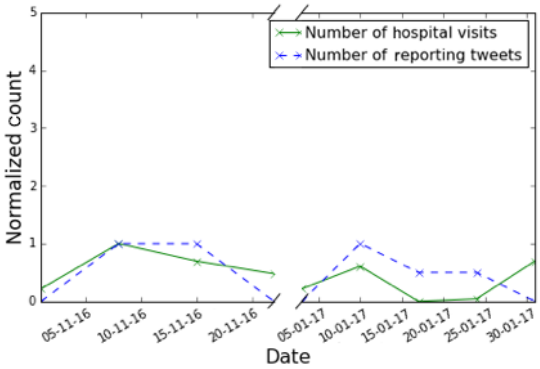

Fig. 4 Count of Arabic tweets (classified as self) versus count of actual hospital visits.

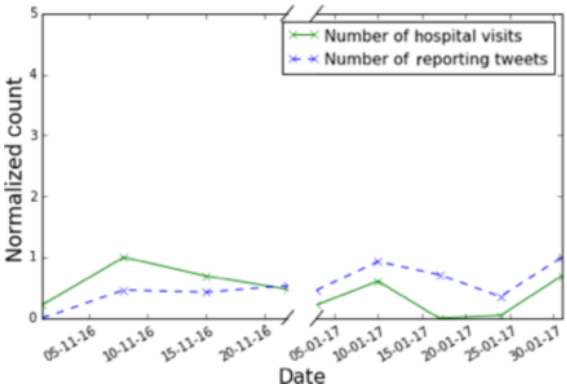

Fig. 5 Count of combined English and Arabic tweets (classified as self) versus count of actual hospital visits.

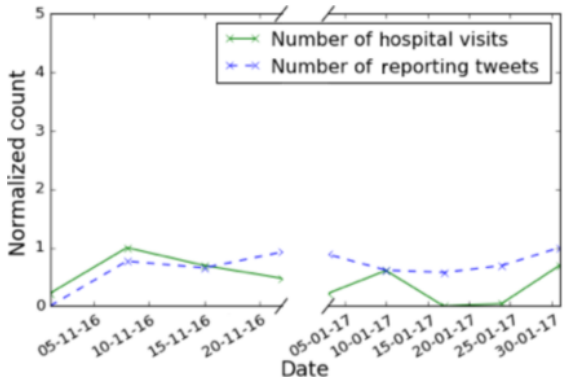

Fig. 6 Count of English tweets (classified as nonself) versus count of actual hospital visits.

Normalized Cross Correlation (NCC) as shown in Eq. (1).

$$\mathrm{NCC} = \frac{1}{n} \sum_{x,y} \frac{1}{\sigma_f \sigma_t} (f\,(x,y) - \overline{f})\,(t\,(x,y) - \overline{t}) \quad (1)$$

In Eq. (1), the number of weeks is represented by n, the weekly count of tweets is represented by x, and the actual weekly count of hospital visits is represented by y. Thus, the function of tweets is denoted by $f(x, y)$

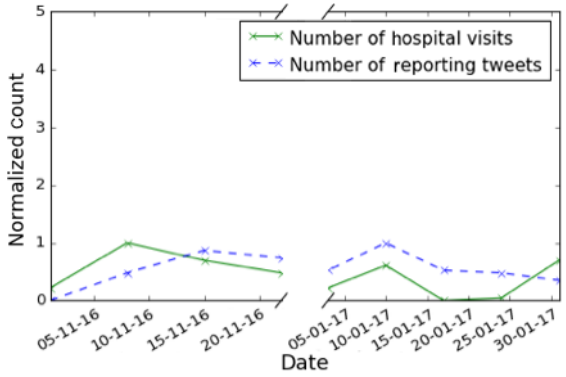

Fig. 7 Count of Arabic tweets (classified as nonself) versus count of actual hospital visits.

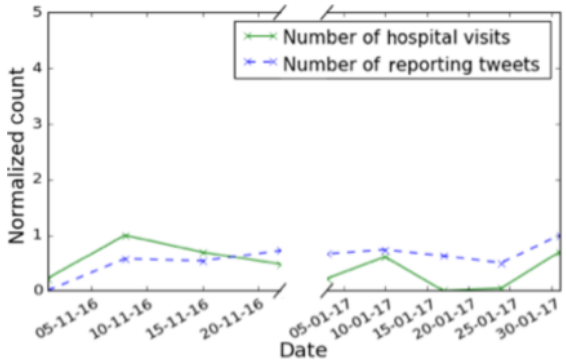

Fig. 8 Count of combined English and Arabic tweets (classified as nonself) versus count of actual hospital visits.

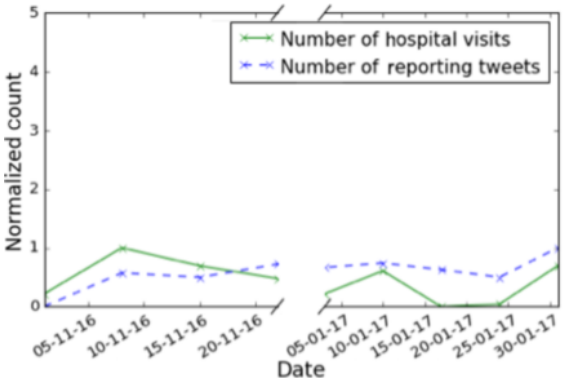

Fig. 9 Count of English tweets (classified as self and nonself) versus count of actual hospital visits.

and the function of hospital visits is denoted by $t(x, y)$. Here, we use $\overline{f}$ to represent the mean of weekly count of tweets and $\overline{t}$ to represent the mean of weekly count of hospital visits. The standard deviations of the weekly count of tweets and weekly count of hospital visits are depicted by σ_f and σ_t, respectively.

4.5 Prediction

To improve the readiness of emergency departments, prediction of Influenza growth is performed. We used

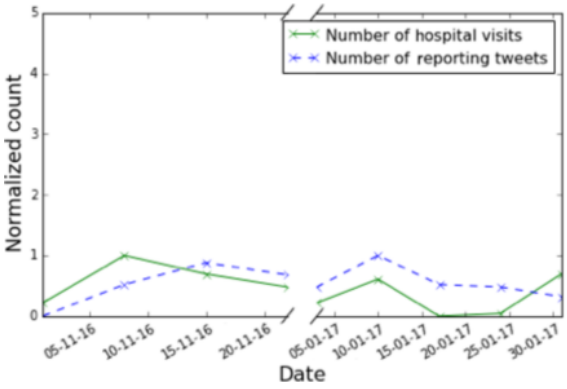

Fig. 10 Count of Arabic tweets (classified as self and nonself) versus count of actual hospital visits.

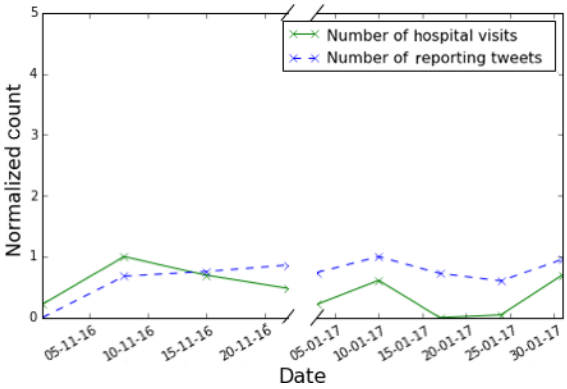

Fig. 11 Count of combined English and Arabic tweets (classified as self and nonself) versus count of actual hospital visits.

a Linear Regression model to predict the Influenza-related count of future hospital visits. These predictions can benefit individuals as well as organizations.

- Individual level: Predictions can be used to alert individuals specially those sensitive to Influenza from going to certain regions that is forecasted to have Influenza outbreak in the near future.
- Organization level: Public health organizations, such as hospitals, can benefit from predictions by improving their preparedness for future Influenza outbreaks. That is, health organizations can prepare the required equipment and necessary vaccines beforehand.

To predict the Influenza-related count of future hospital visits, a linear regression function is fed with two types of input data: (1) count of Influenza reporting tweets, specifically self and nonself reporting tweets of current and previous weeks, and (2) count of Influenza-related hospital visits.

We evaluated the prediction of Tweetluenza by computing the Root Mean Square Error (*R*MSE) between Tweetluenza predictions and the actual data

of hospital visits using Eq. (2). The number of *self* and *nonself* reporting tweets is represented by N. In addition, the Influenza-related count of actual hospital visits is represented by u and the predicted count by Tweetluenza is represented by v.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(u_i - v)^2} \quad (2)$$

5 Expiremental Results

We evaluated the performance of Tweetluenza in this section. All experiments were conducted on a workstation equipped with 8 cores intel xeon CPU e5-2630 v3@2.40 GHz Processor. The workstation has 31.3 GB memory and a 13.8 TB disk. In addition, Tweetluenza was developed over Ubuntu 12.04 OS.

The total number of tweets collected from November 2016 (from 1 November until 20 November) and January 2017 (from 1 January until 30 January) is 2 588 570. About 1 089 684 tweets, which were written either in Arabic or English, are considered in our experiment. The rest of the tweets were discarded because they were written in other languages or do not contain textual information. The Arabic tweets represent 48% (518 558 tweets) and the English tweets represent 52% (571 126 tweets) of the considered tweets. In the Arabic tweets, the percentage of *self* to *nonself* is 6% to 94%, whereas in the English tweets the percentage of *self* to *nonself* is 46% to 54%.

5.1 Correlation analysis

In this section, we analyze the correlation (NCC value) between the counts of Influenza-related tweets with the counts of actual hospital visits due to Influenza. In these experiments, tweets of the self, non-self, and their combination are compared separately against the counts of hospital visits. The correlation is given by Eq. (1). Additionally, the experiments were performed separately for Arabic tweets, English tweets, and the combination of both languages. Below we discuss the correlation results.

(1) Self class: Figures 3 – 5 show the Influenza-related counts of tweets, which are classified as *self*, and the counts of actual hospital visits. These counts are represented as normalized time series. Figures 3 – 5 depict the comparison of the English, Arabic, and combined languages, respectively. The objective of these figures is to show the correlation between Influenza-related tweet counts with counts of actual hospital visits in different weeks. We notice from Figs. 3 – 5 that the two time series, which represent the number of hospital visits and number of reporting tweets, have similar behavior that is confirmed by the correlation values between the time series shown in Tables 10 – 12. Generally, the correlation values are greater than or equal to 0.5. Note that the NCC values for English tweets are slightly better than those of Arabic tweets since the processing of Arabic tweets are usually more complex than that of the English tweets. However, the combination of tweets of both languages enhanced the NCC values as compared to the NCC values of either individual language (see Table 12). Note that the NCC values of Arabic tweets are lower than the NCC values of the English tweets. This is due to the difficulty of distinguishing Arabic personal pronouns, which are in enclitic form (pronouns connected to the word before it). The above result indicates that healthcare authorities in the UAE could rely on the analysis of the combination of Arabic and English tweets to detect Influenza outbreaks rather than on only tweets of either individual language.

(2) Nonself class: The counts of the Influenza-related tweets, which are classified as *nonself*, are represented as normalized time series in Figs. 6 – 8. These figures compare the counts of tweets with the counts of actual hospital visits. Figures 6 – 8 show the comparisons of the English, Arabic, and combined languages, respectively. These figures show correlation between the Influenza-related tweet counts with the counts of actual hospital visits. Note that the time series in these figures are correlated, which are confirmed

Table 10 NCC value between the English tweets (classified as self) and count of actual hospital visits.

Month	NCC value
Nov.	0.57
Jan.	0.67

Table 11 NCC value between the Arabic tweets (classified as self) and count of actual hospital visits.

Month	NCC value
Nov.	0.47
Jan.	0.57

Table 12 NCC value between combined English and Arabic tweets (classified as self) and count of actual hospital visits.

Month	NCC value
Nov.	0.68
Jan.	0.68

by the correlation values of the two time series shown by Tables 13 – 15. When the combined languages are used, the NCC value is greater than or equal to 0.5. Also, note that the NCC values for Arabic tweets are on the average slightly higher than the values of English tweets; this is because the *repTweets* algorithm requires only the keywords to distinguish *nonself-reporting* tweets. That is, unlike the *self-reporting* class, the extraction of the complex personal pronouns in enclitic form is not required. However, the processing of Arabic tweets is usually more complex than that of the English tweets. The analysis results of the *nonself-reporting* class support the findings of the self-reporting class in that the UAE healthcare sector can reliably use the analysis of the combination of Arabic and English tweets to detect Influenza outbreaks. That is because combining Arabic and English tweets enhances the detection of Influenza in the UAE.

(3) Reporting class: The *reporting* class is the combination of both of the *self* and *nonself* classes. Figures 9 – 11 show the counts of the Influenza-related tweets, which are classified as *reporting*. The English, Arabic, and combined languages tweets are depicted in Figs. 9 – 11, respectively. These figures compare the counts of reporting tweets with the counts of actual hospital visits. The objective of these figures is to show whether the counts of the Influenza-related reporting tweets are correlated with the actual hospital visits. As can be seen from the behavior of time series, they have similar behavior, which is verified by the correlation values shown in Tables 16 – 18. Generally, both months show a correlation value that is greater than or equal to 0.5. We notice that the *reporting* class of the combined languages resulted in higher NCC values. This is because the performance of English is better in *self-reporting* and Arabic performed better in *nonself-reporting*. The combination of tweets of both languages enhanced the NCC values as compared to the NCC values of either individual language (see Table 18). Therefore, Influenza outbreaks in the UAE can be reliably detected by the healthcare sector through analyzing the combined Arabic and English tweets rather than the tweets of either individual language.

Table 13 NCC value between the English tweets (classified as nonself) and count of actual hospital visits.

Month	NCC value
Nov.	0.65
Jan.	0.54

Table 14 NCC value between the Arabic tweets (classified as nonself) and count of actual hospital visits.

Month	NCC value
Nov.	0.49
Jan.	0.95

Table 15 NCC value between combined English and Arabic tweets (classified as nonself) and count of actual hospital visits.

Month	NCC value
Nov.	0.63
Jan.	0.79

Table 16 NCC value between the English tweets (classified as self and nonself) and count of actual hospital visits.

Month	NCC value
Nov.	0.62
Jan.	0.79

Table 17 NCC value between the Arabic tweets (classified as self and nonself) and count of actual hospital visits.

Month	NCC value
Nov.	0.57
Jan.	0.92

Table 18 NCC value between combined English and Arabic tweets (classified as self and nonself) and count of actual hospital visits.

Month	NCC value
Nov.	0.63
Jan.	0.93

The above discussed results of the *self*, *nonself*, and *reporting* classes illustrate that the counts of tweets and the counts of actual hospital visits are correlated, and therefore the stream of tweets can be used for analysis and prediction of Influenza.

5.2 Prediction analysis

To investigate the impact of Twitter data on the prediction of Influenza, we conducted several experiment on the dataset of tweets. We applied the k-fold cross validations on the dataset. That is, we split the dataset into k groups, where k is equal to 10 in our experiments. Then, each group is taken as a training data set and the remaining groups as testing data sets. We applied a linear regression model on this dataset to forecast the hospital visits in future weeks. The accuracy of prediction was measured by the *RMSE*. The lower the *RMSE* values, the higher the prediction accuracy.

Table 19 shows the prediction results of the

Table 19 Prediction evaluation using counts of reporting tweets and hospital visits.

		Training *RMSE*	Test *RMSE*
Exp. 1	Nov.	0.076	0.532
	Jan.	0.087	0.616
Exp. 2	Nov.	0.002	0.303
	Jan.	0.004	0.344
Exp. 3	Nov.	0.001	0.183
	Jan.	0.002	0.140

conducted experiments. In these experiments, the inputs to the prediction model are the *reporting* tweets, depicted as *rt-week*$_i$, and the hospital visits, depicted as *hv-week*$_i$. In the first experiment, the inputs are (1) counts of reporting tweets of the current and previous weeks (*rt-week*$_i$ and *rt-week*$_{i-1}$) and (2) count of the current week's hospital visits, *hv-week*$_i$. These inputs are fed into the prediction model (linear regression) to predict the following week's count of hospital visits, *hv-week*$_{i+1}$. The accuracy of the predicted results is investigated by means of the *RMSE* as shown in Table 19. Similarly, the second and third experiments are conducted but with more historical data. In the second experiment, the inputs are counts of *reporting* tweets for three weeks (*rt-week*$_i$, *rt-week*$_{i-1}$, and *rt-week*$_{i-2}$) and counts of actual hospital visits for two weeks (*hv-week*$_i$ and *hv-week*$_{i-1}$). In the third experiment, the inputs are the counts of four weeks of *reporting* tweets (*rt-week*$_i$, *rt-week*$_{i-1}$, *rt-week*$_{i-2}$, and *rt-week*$_{i-3}$) and (2) counts of three weeks of hospital visits (*hv-week*$_i$, *hv-week*$_{i-1}$, and *hv-week*$_{i-2}$). The objective of the above experiments is to predict the following week's count of hospital visits, *hv-week*$_{i+1}$.

We notice from the prediction results shown in Table 19 that the *RMSE* values of the first experiment are higher than the other experiments since the prediction is based on only two weeks of tweet counts (*rt-week*$_i$ and *rt-week*$_{i-1}$) and one week of hospital visits (*hv-week*$_i$). Then, for the second experiment, we added input data for one more week of tweet counts and one more week of hospital visits. As a result, the prediction improved as shown by the lower *RMSE* values. Similarly, for the third experiment, we added one more week of tweet counts and one more week of hospital visits. The prediction of Tweetluenza improved even more. The above result indicates that the healthcare authorities in the UAE can use Twitter data to predict Influenza-related future hospital visits.

Note that the prediction of hospital visits in the future was improved by the use of the inputs: counts of tweets and counts of actual visits. To ensure that the prediction result is not dominated by either input of the linear regression model, the regression weights are estimated to have equal contribution to the prediction. To show that the combination of tweet counts and hospital visits counts improves the prediction of future hospital visits, we repeated the second experiment twice to predict the number of hospital visits: once using only the counts of tweets and once using only the counts of hospital visits. From Table 20, we note that the prediction by using either tweet counts only or the hospital visits counts only produces higher *RMSE* error than using them combined as compared to the result of the second experiment.

6 Conclusion

In this paper, we proposed Tweetluenza system that uses Twitter streams for Influenza surveillance and forecasting in a cross-lingual and cross-dialect settings. Tweetluenza processes the streams of tweets and groups them in *reporting* (*self* and *nonself*) and *non-reporting*. Tweetluenza is context-aware, where words of the same root but with different contextual meanings that do not represent Influenza are removed. The system supports two languages: English and Arabic tweets. Furthermore, Tweetluenza shows that combining tweets of multiple languages (e.g., English and Arabic) in countries were residence speak different languages, such as the UAE, can improve the surveillance and prediction of Influenza. This conclusion was confirmed by the high NCC value (0.78 on average) between Twitter data and the actual hospital visits. Our findings suggest that the healthcare authorities in the UAE can reliably use Twitter data to predict Influenza-related counts of future hospital visits.

Acknowledgment

The authors would like to express their sincere thanks to the UAE Ministry of Health specially their Statistics and Research Center for providing us the actual counts of Influenza-related hospital visits to be used in our research.

Table 20 Prediction evaluation using either tweet counts or counts of hospital visits.

		Number of tweets only		Number of hospital visits only	
		Training *RMSE*	Test *RMSE*	Training *RMSE*	Test *RMSE*
Exp. 2	Nov.	0.006	0.576	0.002	0.745
	Jan.	0.002	0.362	0.061	0.516

References

[1] H. Achrekar, A. Gandhe, R. Lazarus, S. H. Yu, and B. Y. Liu, Online social networks flu trend tracker: A novel sensory approach to predict flu trends, in *Proc. 5th Int. Joint Conf. Biomedical Engineering Systems and Technologies*, Berlin, Germany, 2012, pp. 353–368.

[2] M. J. Paul, M. Dredze, and D. Broniatowski, Twitter improves Influenza forecasting, *PLoS Curr.*, doi: 10.1371/currents.outbreaks.90b9ed0f59bae4ccaa683a39865d9117.

[3] A. F. Dugas, M. Jalalpour, Y. Gel, S. Levin, F. Torcaso, T. Igusa, and R. E. Rothman, Influenza forecasting with Google flu trends, *PLoS One*, vol. 8, no. 2, p. e56176, 2013.

[4] K. Byrd, A. Mansurov, and O. Baysal, Mining Twitter data for Influenza detection and surveillance, in *Proc. 2016 IEEE/ACM Int. Workshop on Software Engineering in Healthcare Systems*, Austin, TX, USA, 2016, pp. 43–49.

[5] S. Ram, W. L. Zhang, M. Williams, and Y. Pengetnze, Predicting asthma-related emergency department visits using big data, *IEEE J. Biomed. Health Inform.*, vol. 19, no. 4, pp. 1216–1223, 2015.

[6] J. W. Li and C. Cardie, Early stage Influenza detection from Twitter, arXiv preprint arXiv: 1309.7340, 2013.

[7] M. Shah, Disease propagation in social networks: A novel study of infection genesis and spread on Twitter, in *Proc. 5th Workshop on Big Data, Streams and Heterogeneous Source Mining: Algorithms, Systems, Programming Models and Applications*, San Francisco, CA, USA, 2016, pp. 85–102.

[8] Statista, Number of monthly active Twitter users worldwide from 1st quarter 2010 to 1st quarter 2019 (in millions), https://www.statista.com/statistics/282087/number-of-monthly-active-Twitter-users/, 2019.

[9] M. A. Al-garadi, M. S. Khan, K. D. Varathan, G. Mujtaba, and A. M. Al-Kabsi, Using online social networks to track a pandemic: A systematic review, *J. Biomed. Inform.*, vol. 62, pp. 1–11, 2016.

[10] A. A. Aslam, M. H. Tsou, B. H. Spitzberg, L. An, J. M. Gawron, D. K. Gupta, K. M. Peddecord, A. C. Nagel, C. Allen, J. A. Yang, et al., The reliability of tweets as a supplementary method of seasonal Influenza surveillance, *J. Med. Internet Res.*, vol. 16, no. 11, p. e250, 2014.

[11] The National, Do you need to take the flu vaccination this winter? https://www.thenational.ae/lifestyle/wellbeing/do-you-need-to-take-the-flu-vaccination-this-winter-1.197490, 2016.

[12] B. Alkouz and Z. Al Aghbari, Analysis and prediction of Influenza in the UAE based on Arabic tweets, in *Proc. 3rd Int. Conf. Big Data Analysis*, Shanghai, China, 2018, pp. 61–66.

[13] A. Guille and C. Favre, Event detection, tracking, and visualization in Twitter: A mention-anomaly-based approach, *Social Network Anal. Min.*, vol. 5, no. 1, p. 18, 2015.

[14] M. Musleh, Spatio-temporal visual analysis for event-specific tweets, in *Proc. 2014 ACM SIGMOD Int. Conf. Management of Data*, Snowbird, UT, USA, 2014, pp. 1611–1612.

[15] E. D'Andrea, P. Ducange, B. Lazzerini, and F. Marcelloni, Real-time detection of traffic from Twitter stream analysis, *IEEE Trans. Intell. Trans. Syst.*, vol. 16, no. 4, pp. 2269–2283, 2015.

[16] J. Capdevila, J. Cerquides, and J. Torres, Recognizing warblers: A probabilistic model for event detection in Twitter, in *Proc. 2016 Anomaly Detection Workshop in the Int. Conf. Machine Learning*, New York, NY, USA, 2016.

[17] K. Lee, A. Agrawal, and A. Choudhary, Real-time disease surveillance using Twitter data: Demonstration on flu and cancer, in *Proc. 19th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, Chicago, IL, USA, 2013, pp. 1474–1477.

[18] E. Aramaki, S. Maskawa, and M. Morita, Twitter catches the flu: Detecting Influenza epidemics using Twitter, in *Proc. Conf. Empirical Methods in Natural Language Processing*, Edinburgh, UK, 2011, pp. 1568–1576.

[19] M. C. Smith, D. A. Broniatowski, M. J. Paul, and M. Dredze, Towards real-time measurement of public epidemic awareness: Monitoring Influenza awareness through Twitter, in *AAAI Spring Symposium on Observational Studies through Social Media and Other Human-Generated Content*, Stanford, CA, USA, 2016.

[20] D. A. Broniatowski, M. J. Paul, and M. Dredze, National and local Influenza surveillance through Twitter: An analysis of the 2012–2013 Influenza epidemic, *PLoS One*, vol. 8, no. 12, p. e83672, 2013.

[21] Q. Zhang, N. Perra, D. Perrotta, M. Tizzoni, D. Paolotti, and A. Vespignani, Forecasting seasonal Influenza fusing digital indicators and a mechanistic disease model, in *Proc. 26th Int. Conf. World Wide Web*, Perth, Australia, 2017, pp. 311–319.

[22] F. Zhang, J. Luo, C. Li, X. Wang, and Z. Y. Zhao, Detecting and analyzing Influenza epidemics with social media in China, in *Proc. 18th Pacific-Asia Conf. Knowledge Discovery and Data Mining*, Tainan, China, 2014, pp. 90–101.

[23] S. Brennan, A. Sadilek, and H. Kautz, Towards understanding global spread of disease from everyday interpersonal interactions, in *Proc. 23rd Int. Joint Conf. Artificial Intelligence*, Beijing, China, 2013.

[24] H. Achrekar, A. Gandhe, R. Lazarus, S. H. Yu, and B. Y. Liu, Twitter improves seasonal Influenza prediction, in *Proc. Int. Conf. Health Informatics*, Vilamoura, Portugal, 2012, pp. 61–70.

[25] A. Signorini, A. M. Segre, and P. M. Polgreen, The use of Twitter to track levels of disease activity and public concern in the U.S. during the Influenza A H1N1 pandemic, *PLoS One*, vol. 6, no. 5, p. e19467, 2011.

[26] I. Kagashe, Z. Yan, and I. Suheryani, Enhancing seasonal Influenza surveillance: Topic analysis of widely used medicinal drugs using Twitter data, *J. Med. Internet Res.*, vol. 19, no. 9, p. e315, 2017.

[27] S. Grover and G. S. Aujla, Prediction model for Influenza epidemic based on Twitter data, *Int. J. Adv. Res. Comput. Commun. Eng.*, vol. 3, no. 7, pp. 7541–7545, 2014.

[28] H. Achrekar, A. Gandhe, R. Lazarus, S. H. Yu, and B. Y. Liu, Predicting flu trends using Twitter data, in *Proc. 2011 IEEE Conf. Computer Communications Workshops*, Shanghai, China, 2011, pp. 702–707.

[29] E. K. Kim, J. H. Seok, J. S. Oh, H. W. Lee, and K. H. Kim, Use of Hangeul Twitter to track and predict human Influenza infection, *PLoS One*, vol. 8, no. 7, p. e69305, 2013.

[30] H. Woo, H. S. Cho, E. Shim, J. K. Lee, K. Lee, G. Song, and Y. Cho, Identification of keywords from Twitter and web blog posts to detect Influenza epidemics in Korea, *Disaster Med. Public Health Prep.*, vol. 12, no. 3, pp. 352–359, 2018.

[31] K. Lee, A. Agrawal, and A. Choudhary, Forecasting Influenza levels using real-time social media streams, in *Proc. 2017 IEEE Int. Conf. Healthcare Informatics*, Park City, UT, USA, 2017, pp. 409–414.

[32] S. Ghosh, T. Rekatsinas, S. R. Mekaru, E. O. Nsoesie, J. S. Brownstein, L. Getoor, and N. Ramakrishnan, Forecasting rare disease outbreaks with spatio-temporal topic models, in *NIPS 2013 Workshop on Topic Models*, Lake Tahoe, NV, USA, 2013.

[33] M. Akbari, X. Hu, F. Wang, and T. S. Chua, Wellness representation of users in social media: Towards joint modelling of heterogeneity and temporality, *IEEE Trans. Knowl. Data Eng.*, vol. 29, no. 10, pp. 2360–2373, 2017.

[34] Gulfbusiness, Top 10 nations where people spend most time on social media, http://gulfbusiness.com/top-10-nations-people-spend-time-social-media/, 2016.

[35] Weedoo, Twitter arab world — Statistics Feb 2017, https://weedoo.tech/Twitter-arab-world-statistics-feb-2017/, 2017.

[36] L. Dinges, A. Al-Hamadi, M. Elzobi, Z. Al Aghbari, and H. Mustafa, Offline automatic segmentation based recognition of handwritten Arabic words, *Int. J. Signal Process., Image Process. Pattern Recognit.*, vol. 4, no. 4, pp. 131–143, 2011.

[37] M. Elzobi, A. Al-Hamadi, Z. Al Aghbari, L. Dings, and A. Saeed, Gabor wavelet recognition approach for off-line handwritten Arabic using explicit segmentation, in *Image Processing and Communications Challenges 5*, Heidelberg, Germany, 2014, pp. 245–254.

[38] NLTK Python Library, Natural Language Toolkit, http://www.nltk.org/, 2019.

[39] Go-Gulf, Expats in middle east — Statistics and trends, https://www.go-gulf.ae/blog/expats-middle-east/, 2019.

Balsam Alkouz received the MS degree from the University of Sharjah, Sharjah, United Arab Emirates (UAE) in 2018. She also works as a research assistant in the Data Mining and Multimedia Research Group under the Research Institute of Sciences and Engineering at University of Sharjah. Her interests focus on data mining and social media analysis as well as augmented and virtual reality technologies. She is a Google Developer Group manager and a Women Techmaker leader in Sharjah, UAE.

Jemal Hussien Abawajy is a full professor at Deakin University, Australia. He has delivered more than 60 keynote and seminars worldwide and has been involved in the organization of more than 300 international conferences in various capacity including chair and general co-chair. He has also served on the editorial-board of numerous international journals including *IEEE Transaction on Cloud Computing*. He is the author/co-author of more than 300 refereed articles and supervised numerous PhD students to completion.

Zaher Al Aghbari is a professor and chairman of the Department of Computer Science at University of Sharjah, UAE. He received the BSc degree from the Florida Institute of Technology, Melbourne, FL, USA, in 1987, and the MSc and PhD degrees from Kyushu University, Fukuoka, Japan, in 1998 and 2001, respectively. He was with the Department of Intelligent Systems, Kyushu University, Japan, from 2001 to 2003. Since 2003, he has been with the Department of Computer Science, University of Sharjah, UAE. Currently, he is a professor of databases and data mining in the Department of Computer Science, College of Sciences, University of Sharjah. His research interests include data mining, multimedia databases, big data, social networks, wireless sensor networks, data streams, and Arabic handwritten text retrieval.