AN OVERVIEW ON THE USE OF DATA MINING AND LINGUISTICS TECHNIQUES FOR BUILDING MICROBLOG-BASED EARLY DETECTION SYSTEMS IN THE HEALTHCARE SECTOR

Haider M. Habeeb[1] and Nabeel Al-A'araji[2]

[1]Department of Information Networks, College of IT, University of Babylon, Iraq
[2]Ministry of Higher Education and Scientific Research, Iraq

ABSTRACT

The usage of Online Social Networks (OSN), such as Facebook and Twitter are becoming more and more popular in order to exchange and disseminate news and information in real-time. Twitter in particular allows the instant dissemination of short messages in the form of microblogs to followers. This Survey reviews literature to explore and examine the usage of how OSNs, such as the microblogging tool Twitter, can help in the detection of spreading epidemics. The paper highlights significant challenges in the field of Natural Language Processing (NLP) when using microblog based Early Disease Detection Systems. For instance, microblogging data is an unstructured collection of short messages (140 characters in Twitter), with noise and non-standard use of the English language. Hence, research is currently exploring the field of linguistics in order to determine the semantics of the text and uses data mining techniques in order to extract useful information for disease spread detection. Furthermore, the survey discusses applications and existing early disease detection systems based on OSNs and outlines directions for future research on improving such systems based on a combination of linguistics methods, data mining techniques and recommendation systems.

KEYWORDS

Data Mining, Social Networks, Healthcare

1. INTRODUCTION

Health authorities consider it to be very important to develop warning systems for flu, because of seasonal influenza epidemics are a major public health concern, causing tens of millions of respiratory illnesses and 250,000 to 500,000 deaths worldwide each year [1]. Health authorities treat people who visit GPs and health centers, however, in order to contain the pandemic it is desired to identify patients suffering from influenza by not visiting a GP. This could be facilitated by making use of information disseminated on OSNs by individuals.

Social Media disseminates different forms of media content that are widely available and created by end-users [2] such as text, images, videos, etc. Social Networks became very important for analyzing such media content [3]. Users on OSNs can connect to other users (i.e. friends or followers) and thus become aware of information disseminated in these networks such as news and events. Part of this information could be health status and wellbeing. Many applications and research projects have attempted to make use of OSN data for detecting topics and events [4]. The importance of OSN data becomes apparent by looking at the sheer amount of data generated, i.e.

there are more than 500 million users on twitter that publish more than 500 million tweets on a daily basis [5]. The Knowledge Discover from Data (KDD) process can be used to extract information from these huge amounts of data.

There are nearly 5000 clinics providing data to the Japanese Infection Disease Surveillance Center [6], also the U.S. Center for Disease Control and Prevention collects its own disease information data [6]. In spite of these collections of disease information, traditional surveillance systems have failed to report emerging diseases in real-time because of considerable delays caused by actively collecting data [7]. In order to close the gap between collecting data and analyzing it, researchers started looking into new technologies and data sources for collecting data in a timely fashion. OSNs, such as Twitter is just one, but an important platform that disseminates news and events in nearly real-time [8]. Twitter has become a very important medium of opinion expression and information propagation on varied topics [9]. Twitter users can write tweets and follow other users' tweets from anywhere in the world by using mobile devices or simply desktop computers. Usually, collecting data from web done by crawlers some of them retrieve content from publicly index-able Web which is called traditional crawlers. Hidden Web Exposer (HiWE), a prototype crawler built at Stanford which is designed for retrieve content from hidden Web [10].

Twitter users can be considered as “sensors” that capture aspects not yet measurable by any device-sensors. For example, it is impossible to measure how 'happy' or 'sad' iPhone users with the product by device-sensor. This sort of information can be gained by “human sensors” such as Twitter users [11]. There are many different types of information in Twitter that are potentially valuable to public health research about different diseases [12]. When early detection of a disease spreading is followed by a fast response, then health authorities can reduce the impact of both seasonal and pandemic diseases such as influenza[13][14][15]. For example, extracting influenza symptoms from tweets such as fever and cough are the best predictions for flue, since they have shown a good predictive accuracy of 80% according to the authors of [16].

This paper aims to show the state-of-the-art papers and concentrates on the challenges for the implementation of microblog based early detection systems in the healthcare sector. It will also explore data analysis techniques for the development of such systems, including Natural Language Processing (NLP) and Linguistics, but also data mining techniques, in order to extract meaning from such data. Furthermore, the paper will follow up on the discussion of the reviewed techniques and systems in order to propose a new methodology for early disease detection systems from microblogging data.

This paper has structured in four sections. The introduction is the first section followed by Data Analytics Techniques used for the Development of Microblog-based Early Detection Systems section which is branches into two sub-section “Natural Language Processing Techniques” and “Data Mining Techniques used for Microblog-based early detection systems”. The latter is divided into three sub-sections (Classification Techniques used in Healthcare and Medical Research, Clustering Techniques used in Healthcare and Medical Research, and Text Mining Techniques used in Healthcare Sector). Section three titled (Applications based on Early Detection Systems in the Healthcare Sector). Discussion and Conclusions is the fourth and final section.

2. DATA ANALYTICS TECHNIQUES USED FOR THE DEVELOPMENT OF MICRO-BLOG-BASED EARLY DETECTION SYSTEMS

Data mining techniques are not sufficient for the development of microblog-based Early Disease Detection Systems. Therefor most research in the area considers tools from the linguistic domain as important to be taken into account for mining tweets [17] [18].

Knowledge Discovery from Data (KDD) process can be used to extract knowledge from data as displayed in figure 1. KDD consists of several steps and Data Mining is one but very important step of KDD [19]. In fact, Data Mining is considered one of the most important stages of the KDD process [20][21].

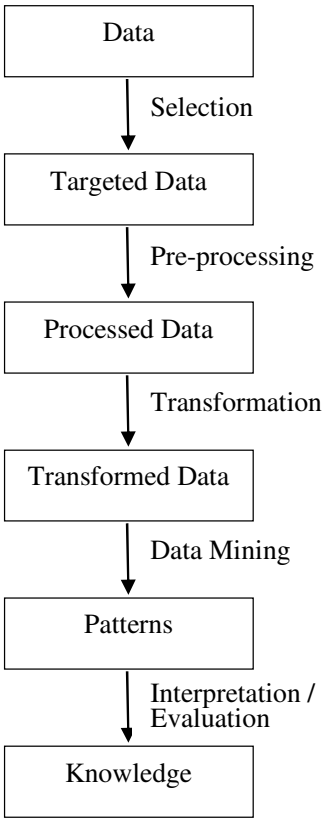

Figure 1. The Knowledge Discovery from Data process

The following steps explain the main activities of the KDD process:

1- Data is collected from different sources. A selection step is performed to extract target data.
2- This data is then pre-processed, i.e. resolving issues related to missing values, noise, etc.
3- Transformation step converts the data into a usable format for the data mining algorithms.
4- Data Mining uses intelligent methods for extracting meaningful patterns.
5- In the interpretation/evaluation step, the extracted patterns need to be interpreted in and evaluated in order to “extract knowledge”.

In the scope of this review, OSN (i.e. Twitter data) is the data source of KDD and thus the data is unstructured. This prompts investigators to apply a combination between linguistics (such as NLP) and data mining techniques on the transformed data which is then used for the remainder of the KDD process. Figure 2 illustrates how NLP is combined with Data Mining in the KDD progress.

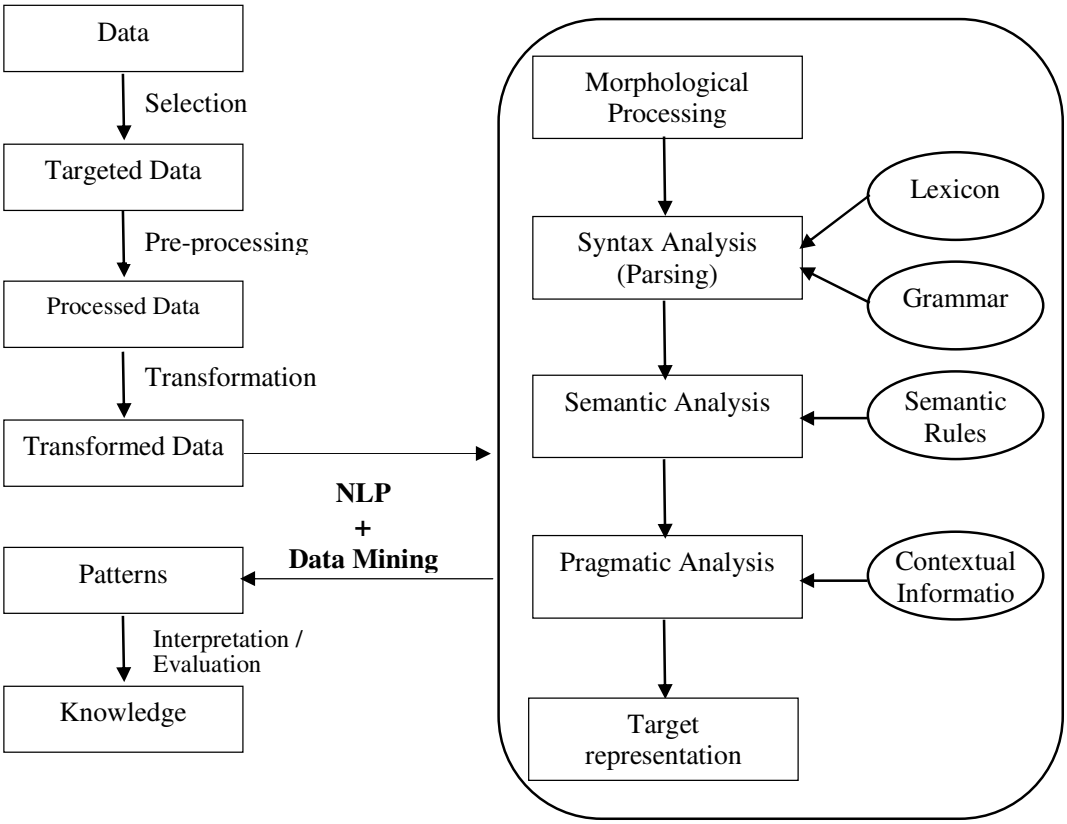

Figure 2. Combination of KDD and NLP

In the *morphological processing* stage, strings would be broken into set of tokens corresponding to discrete words, sub-words and punctuation forms. For example a word like "unhappily" can be broken into three sub-word tokens as: un-happy–ly. *Syntactic analyzer* uses a dictionary of word definitions (lexicon) and set of syntax rules (Grammar) to check if a string of words is well-formed and to break it up into a structure that shows the syntactic relationships between the different words. In order to carry out *semantic analysis* the lexicon must be expanded to include semantic definitions for each word it contains and the grammar must be extended to specify how the semantics of a phrase are formed from the semantics of its component parts. *Pragmatic analysis* simply fits actual objects/events that exist in a given context with object references obtained during semantic analysis. In other cases pragmatic analysis can disambiguate sentences, which cannot be fully disambiguated during the syntax and semantic analysis phases [22].

During analyzing the texts for finding flu related tweets, it is very important to differentiate between awareness and infection. This step can be conducted by further analysis using NLP [23]. However, with dynamic data like real-time messages from Twitter that have different slang in the expression, more techniques have involved for obtaining knowledge.

2.1. Natural Language Processing Techniques

The need to understand natural languages by computers lead to the emergence of NLP as a field of study to deal with human languages such as text or speech [24]. NLP applications are widely used on a daily basis [25]. There are many NLP techniques such as Machine Translation, Analysis of Discourse, Morphological Splitting, Generation and Understanding of Natural

Language, Identification of Named Entities, Marking Part of Speech, Optical Character Recognition, Recognizing Boundary of Sentences, Parsing of Text, Recognition of Speech, Analysis of Sentiments, Finding Words Boundary, and Word Sense Disambiguation [26]. NLP can be used to extract the meaning/knowledge from tweets expressed in natural language.

Lamb, et al.[23] attempted to use NLP in order to extract tweets that report a concrete flu infection rather than tweets that talk about the flu in general terms such as tweets about flu awareness.

They manually created a set of word class features (Table 1) that explain possessive words, flu related words, "self" words, "other" words, and "fear" related words.

Table 1. Set of Word Class Feature (Lamb, et al.)[23]

	Class Name	Words in Class
1	Infection	getting, got, recovered, have, having, had, has, catching, catch, cured, infected
2	Possession	bird, the flu, flu, sick, epidemic
3	Concern	afraid, worried, scared, fear, worry, nervous, dread, dreaded, terrified
4	Vaccination	vaccine, vaccines, shot, shots, mist, tamiflu, jab, nasal spray
5	Past Tense	was, did, had, got, were, or verb with the suffix "ed"
6	Present Tense	is, am, are, have, has, or verb with the suffix "ing"
7	Self	I, I've, I'd, I'm, im, my
8	Others	your, everyone, you, it, its, u, her, he, she, he's, she's, they, you're, she'll, he'll, husband, wife, brother, sister, people, kid, kids, children, son, daughter

They included tweets that match the features based on the word class features outlined in Table 1, matching specific sequences of words, word classes, and part of speech tags. They reported that their system had a correlation of 0.9887 during the weeks beginning 8/30/09–05/02/10.

One of the most fundamental elements in the linguistic domain is Part-of-Speech (POS) tagging that performs a syntactic analysis, which has numerous applications in NLP. Treebanks from the newswire domain (such as Wall Street Journal) have been widely used for training POS taggers [27]. However, tagging does not perform as well on texts that are outside the domain on which the Tagger was trained. In addition tweets have extra challenges due to the conversational nature of tweets, the lack of orthography and the 140 character limit of each tweet [27]. An English POS tagger designed especially for Twitter data with nearly to 90% accuracy has been created by Gimpel, et al.[27].

2.2. Data Mining Techniques used for Microblog-based early detection systems

Data Mining is considered one of the most widely used areas of academic research in order to discover useful information from massive amounts of data [19]. Data Mining has also become popular in the healthcare sector due to the need for analytics methods to extract previously unknown medical knowledge from data. Typically Healthcare related data is huge in size and complex, which makes it challenging to analyze for decision support [28]. Thus data mining has been used for the prediction of various diseases [29][30][31][32]. For better health, Association Rule (AR) applied to encourage the regularity in performing the mild exercise, just before sleep that may help deepening the sleeping patterns [33].

There are different Data Mining techniques, such as classification and clustering used in the healthcare sector in order to enhance the decision making process [28].

2.2.1 Classification Techniques used in Healthcare and Medical Research

Classification aims to automatically categorize data into target classes. For example, patients can be classified due to the risk as "high" or "low" according to their health condition using automatic data classification approaches. Classification is typically a supervised learning approach, where a classifier model in induced on training data with known classification before it is applied on data with unknown classification. Data may be categorized into binary (i.e. 'high risk' and 'low risk') or multiple classes (i.e. 'high risk', 'low risk' and 'medium risk'). In supervised learning the training data (data with known class labels) is divided into a training set used for inducing the classifier and a test set used for evaluating the classifiers predictive accuracy.

The healthcare sector uses classification techniques such as K-Nearest Neighbor (KNN), Decision Tree (DT), Support Vector Machine (SVM), Neural Network (NN), and Bayesian methods [34]. For example KNN and Linear Discriminate Analysis (LDA) have been used as classification techniques for chronic disease in generating early warning system [34]. Also a KNN classifier for analyzing the patients suffering from heart disease has been used in [35]. A universal Hybrid Decision Tree classifier (which incorporates elements of SVMs and Naive Bayes) has been proposed for classifying the activity of patient having chronic disease in the works of [36]. Decision trees illustrated the patterns of smoking in adults for better understanding the health condition, distress, demographic and alcohol[37]. In [38] a genetic Support Vector Machine (SVM) classifier was proposed for analyzing heart valve disease. It extracts the important features and classifies the signal obtained from the ultrasound of heart valve into "normal" and "abnormal". Furthermore NNs were used as a model for analyzing chest diseases in [39]. In addition in order to develop effective decision support system for diagnosing heart disease, the use of an ensemble NNs has proposed in [40]. Also a decision support system using Bayesian Belief Networks (BNN) has been developed for analyzing risks that are associated with health effects [41].

2.2.2 Clustering Techniques used in Healthcare and Medical Research

Clustering is a data mining technique which separates data into small subgroups according to some measure of similarity. The aim is to group similar data items in the same group and dissimilar items in different groups [42]. It is different from classification because it has no predefined class [43].

Partitioned, hierarchical, and density based Clustering are different kinds of methods used in clustering. K-means and K-medoids are an example of partitioned clustering where K-means partitions data into k clusters (k needs to be defined by the user in advance), whereas K-medoids uses medoids (medoids are similar in concept to means or centroids, but they are always members of the data set) instead of mean for grouping the cluster[19]. K-means clustering has been used to classify the Alzheimer's disease data feature into pathologic and non-pathologic groups which were used for early detection of Alzheimer's disease [44]. In [45],[46] and [47] more uses of partitioned clustering in healthcare sector can be found.

Hierarchical clustering does not need to define the number of clusters in advance. It separates data into hierarchical clusters either using bottom up (agglomerative) approach or top down (divisive) approach. The result often represented in a tree like structure called dendrogram as depicted in Figure 3.

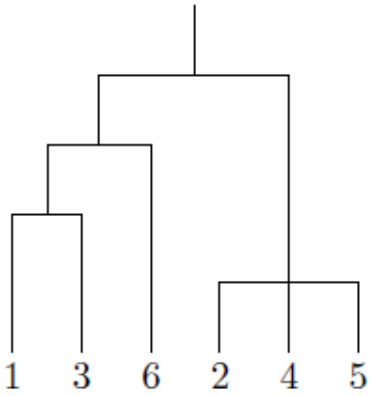

Figure 3. Example of Dendrogram

Agglomerative clustering initially considers each data point as a separate cluster, the two closest clusters are then repeatedly merged until all data items are contained in one cluster or a termination condition has been reached. Divisive clustering assumes that all data items are initially contained in once cluster which is then repeatedly divided into smaller clusters until each cluster contains only one data item or a termination condition is reached [19]. In [48] hierarchical clustering approach has been used for grouping patients according to their length of stay in the hospital in order to enhance the capability of hospital resource management.

Density clustering approach can handle outliers and arbitrary shaped clusters. DBSCAN and OPTICS are two approaches of Density based clustering, they build clusters according to density connectivity analysis [19]. DENCLUE is another approach of density based clustering that forms the grouping of data according to distribution value analysis of a density function[19]. Density clustering has been used to separate unhealthy or wound skin from healthy skin and determines the sub regions of varied colour or spotted parts inside the unhealthy skin, which is useful for classification [49].

2.2.3 Text Mining Techniques used in Healthcare Sector

Data mining tools are typically designed to deal with structured data from databases, while text mining techniques work with unstructured or semi-structured data [50]. Often unstructured data is transformed (Figure 1) into structured data in order to apply data mining techniques on it. As this survey is about unstructured micro-blog data analysis in the healthcare sector, some research on text mining will be highlighted in this Section.

Several models have been investigated for analyzing tweets in order to predict rates of influenza like illnesses.

The authors of [51] analyzed about 500 million tweets during an eight months period and found that it is possible to forecast future influenza rates through tracking the number of flu related tweets [51]. The authors concluded that supplying the classifier with more sophisticated linguistic features, such as n-grams and synonyms, would improve the accuracy. Loosely speaking, they concluded that a more careful pre-processing stage will improve the quality of analysis [51].
Aramaki et al.[52] addressed the issue of detecting influenza epidemics at a large scale and in real-time. They filtered out tweets that contained the word “influenza” which resulted in 300 million tweets extracted from Twitter within a two year period. They then classified the tweets into negative and positive labels depending if the author of the tweet, or a person close to the

author were infected with the flu. The classification process has been done using SVM classifier. The experimental results showed the practicality of the proposed approach (0.89 correlation to the ground truth). Their proposed method shows high correlation especially during the outbreak and early spread of the flu. They used manually classified a training and test sample of the whole corpus. A subset of this corpus is shown in Table 2.

Table 2. Corpus (Tweets with a Positive or Negative Label)[52]

+ / -	Tweets
+	A bad influenza is going around in our lab.
+	I caught the flu. I was burning up.
+	I think I'm coming down with the flu.
+	It's the flu season. I had it and now he does.
+	Don't give me the flu.
+	My flu is worse than it was yesterday.
-	In the normal flu season, 80 percent of deaths occur in people over 65
-	Influenza is now raging throughout Japan.
-	His wife also contracted the bird flu, but has recovered.
-	You might have the flu. Has anyone around you had it?
-	Bird flu damage is spreading in Japan.

Achrekar et al. [53] made use of tweets in order to predict the activity of Influenza Like Illnesses (ILI) in a population. They collected tweets that contained phrases like “I got flu” or “down with swine flu” as early indicators for influenza activity. They developed the Social Network Enabled Flu Trends (SNEFT) architecture as depicted in Figure 4 (adapted from [57]) for collecting relevant tweets.

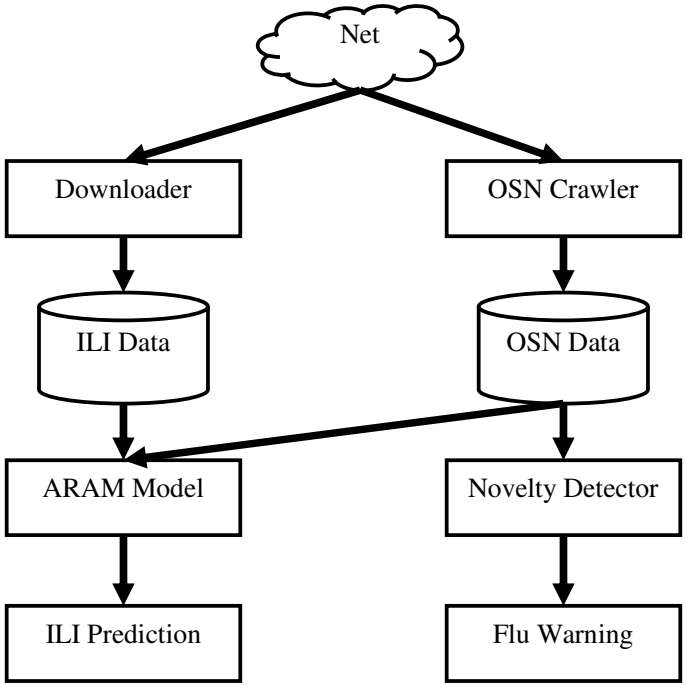

Figure 4. The system architecture of SNEFT.[53]

The results have shown a high correlation between related tweets and ILI activity in the Center for Disease Control (CDC) data with Pearson correlation coefficient of 0.9846. They build auto-regression models to predict number of ILI cases in a population as percentage of visits to physicians in successive weeks. They tested their regressive models with the historic CDC data and verified that Twitter data substantially improves the model's accuracy in predicting ILI cases. In view of the lag inherent in CDC's ILI reports, Twitter data provides near real-time assessment of influenza activity and can be used to predict current ILI activity levels [53].

3. APPLICATIONS BASED ON EARLY DETECTION SYSTEMS IN THE HEALTHCARE SECTOR

This section introduces applications and projects that could be used to track diseases in the early stages of an outbreak. These kinds of applications are usually called application based early detection systems.

Denecke et al.[3] developed the M-Eco project which uses an event-based approach to the early detection of emerging health threats. Health threats can take a long time to become visible and it is very important to reduce this time [3]. The M-Eco project has reported to improve the abilities for disease surveillance by technologies that allow the monitoring of social and multimedia data.[3]

In addition to the sources of traditional disease monitoring systems M-Eco exploits social media and multimedia data for detecting indicators of health threats in order to inform health authorities. M-Eco addresses limitations of current systems for Epidemic Intelligence by exploiting more sophisticated event-detection technologies such as unsupervised and supervised methods. It monitors additional resources like Web 2.0 data and multimedia and enabling access to additional information related to disease outbreaks collected from multiple sources. It can personalize and filter these results.

M-Eco focuses on creating user-defined signal searches. The user (epidemiologist, public health official or decision maker) has to specify a signal search, i.e. he/she selects symptoms or disease names together with a location or time span of the user's interest. The system shows related alerts according to a given such signal definition. The M-Eco system consists of four major components:

- Content collection and preprocessing,
- Event detection,
- Signal generation,
- Recommendation and user modeling.

The classification of tweets has been evaluated in comparison with human annotation. For this 4,000 documents in English and 500 in German have been labeled for training the classifier. For testing 1,000 annotated documents in English and 200 in German have been used. Precision and recall of the system has been calculated and the results of the experiments are stated Table 3.

Table 3. M-Eco Results of relevance classification. [3]

	Precision	Recall	F-Measure (F1)
English Tweets	83.3%	74.4%	78.59%
German Tweets	83%	46%	59.2%

Talvis et al. [54] developed *Flutrack.org*, an open source platform for monitoring influenza epidemics in Twitter. Flutrack.org is a real-time application that tracks the spread of the flu. It gathers flu related tweets from the entire world based on search terms that are synonym of influenza or flu symptoms. The tags being tracked are: Influenza, flu, chills, headache, sore throat, runny nose, sneezing, fever, and dry cough. For every tweet extracted, additional metadata is extracted too.

Tags and hashtags have been removed automatically from tweets. Also Flutrack only saves tweets that have a geolocation for visualization purposes, more than 5 characters and no non ASCII characters. Profile location is also automatically filtered from "suspicious" words such as "home", "heaven" etc.), in consideration of avoiding false or non-existing location coordinates.

The tweets in Flutrack are visualized on a world map and updated every 20 minutes. New tweets are pinned in a worldwide map. In terms of city level, the Flutrack user can navigate through Google maps in order to examine where the tweets were exactly posted. Visuals are displayed on a styled Google map, free of unnecessary information and centered on mirroring the accurate position of tweets. The system displays new tweets and tweets that are up to seven days old, with a deflection of 20 meters in coordinates. Figure 5 is a snapshot of Flutrack application.

Figure 5. Snapshot of Flutrack application

The corresponding coefficient of determination was estimated to be 0.79 thus showing a high degree of correlation between Google Flu Trends and the Flutrack platform. Application Programming Interface (API) enables the user to access a website's data without going near its databases. Flutrack's data are provided to every user via a simple JSON call. JavaScript Object Notation (JSON) is a lightweight, easy and popular way to exchange data.

4. DISCUSSION AND CONCLUSION

This review shades some light about the research field of microblogging early detection systems to show that more investigations are required in order to improve results. Relying on keywords or query logs only is not enough to proof event's prediction [18]. Although many people talk about disease in their tweets, this does mean that they are not sick. Consequently, researches applied further steps in extracting flu related tweets by classifying them into "author" and "other" using NLP techniques [19]. Some improvements were shown by applying these new techniques comparison to keywords method. For instance, Lamb, et al. [23] succeeds to classify the tweets. However, this classification was depending on their class features that were created manually.

Based on the informal language that is usually used in OSN, class features cannot be determined. They remain flexible and depend on the accuracy of results.

The classifier that was introduced by Aramaki et al.[52] seems more rational in classifying tweets because a training and test sample of the whole corpus were classified manually. Hence, it can be concluded that their method depends on the OSN language because they have dealt with the same environment. On the other hand, this approach still suffers from many issues such as the size of corpus and the need for manual classification to get more accuracy.

In the system that was developed by Talvis et al. [54] (*Flutrack.org*), a good corpus of flu related tweets was used. The system considered the classification of new tweets. This clearly presents a suitable application by visualizing flu related tweets on the map.

In order to overcome the above discussed shortcomings, our web-based recommender system has been proposed to determine geolocation of potential emerge epidemic depends on symptoms by exploring OSN data such as Twitter. The main features of the proposed system are digging deeply to determine the location instead of ignoring it as well as adding new factor (feature) to weigh the correctness of author's location depends on identifying from which device a tweet comes through. This approach will enhance the accuracy of determining the location of emerging potential epidemic.

To conclude, KDD and NLP have to be integrated together in order to explore informal languages such as Twitter. This is because of its unstructured nature. Additionally, tweets need more investigations by combining text mining, NLP techniques and recommendation systems techniques. Such integration will promote the accuracy of the obtained results.

REFERENCES

[1] K.Thursky, "Working towards a simple case definition for influenza surveillance . PubMed Commons," J. Clin. Virol., vol. 27, no. 2, pp. 170–179, 2003.

[2] A.Kaplan and M. Haenlein, "Users of the world, unite! The challenges and opportunities of Social Media," Bus. Horiz., vol. 53, no. 1, pp. 59–68, Jan. 2010.

[3] K.Denecke, P. Dolog, and P. Smrz, "Making use of social media data in public health," Proc. 21st Int. Conf. companion World Wide Web - WWW '12 Companion, p. 243, 2012.

[4] H.Achrekar, A. Gandhe, R. Lazarus, S. Yu, and B. Liu, "Twitter Improves Seasonal Influenza Prediction.," in Fifth Annual International Conference on Health Informatics, 2012.

[5] I.Ifrim, G., Shi, B., & Brigadir, "Event Detection in Twitter using Aggressive Filtering and Hierarchical Tweet Clustering.," SNOW-DC@ WWW, pp. 33–40, 2014.

[6] E.Aramaki, S. Maskawa, and M. Morita, "Influenza Patients Are Invisible in the Web: Traditional Model Still Improves the State of the Art Web Based Influenza Surveillance.," AAAI Spring Symp. Self-Tracking Collect. Intell. Pers. Wellness, no. 15, pp. 5–8, 2012.

[7] Y.Xie, Z. Chen, Y. Cheng, and K. Zhang, "Detecting and tracking disease outbreaks by mining social media data," Proc. Twenty-Third Int. Jt. Conf. Artif. Intell. AAAI Press, pp. 2958–2960, 2013.

[8] A.Java, X. Song, T. Finin, and B. Tseng, "Why We Twitter : Understanding Microblogging," in Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis, 2007, pp. 56–65.

[9] M.Adedoyin-Olowe, M. M. Gaber, and F. Stahl, "TRCM: A Methodology for Temporal Analysis of Evolving Concepts in Twitter," in Artificial Intelligence and Soft Computing, no. 7895, 2013, pp. 135–145.

[10] R.Shandilya, S. Sharma, and S. Qamar, "A Domain Specific Indexing Technique for Hidden Web Documents," vol. 2, no. 2, pp. 37–41, 2012.

[11] V.Singh, M. Gao, and R. Jain, "Event Analytics on Microblogs," in Proceedings of the WebSci10: Extending the Frontiers of Society On-Line, April 26-27th, Raleigh, NC: US., 2010, pp. 1–4.

[12] M.Paul and M. Dredze, "You are what you Tweet: Analyzing Twitter for public health.," in In Proc. of the 5th International AAAI Conference on Weblogs and Social Media (ICWSM)., 2011.

[13] J.K.Taubenberger and D. M. Morens, "Influenza : The Once and Future Pandemic," Public Heal. Rep., vol. 125, no. Suppl 3, pp. 16–26, 2010.
[14] Beveridge W, "The chronicle of influenza epidemics . PubMed Commons," Hist. Philos. Life Sci., pp. 223–234, 1991.
[15] Belshe RB, "An introduction to influenza : lessons from the past in epidemiology , prevention , and treatment . PubMed Commons," Manag. Care, pp. 2–7, 2008.
[16] R.Eccles, "Understanding the symptoms of the common cold and influenza," Lancet Infect. Dis., vol. 5, no. 11, pp. 718–725, 2005.
[17] S.Doan, "Enhancing twitter data analysis with simple semantic filtering: Example in tracking influenza-like illnesses," … Informatics, Imaging …, no. 1, 2012.
[18] M.Sokolova, S. Matwin, and D. Schramm, "How Joe and Jane Tweet about Their Health : Mining for Personal Health Information on Twitter," in Proceeding of Recent Advances in Natural Language Processing, 2013, no. September, pp. 626–632.
[19] J.Han and M. Kamber, Data Mining: Concepts and Techniques. 2006.
[20] F.Stahl and J. Ivan, "An overview on the use of neural networks for data mining tasks," WIREs Data Min. Knowl. Discov., pp. 193–208, 2012.
[21] U.Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "The KDD process for extracting useful knowledge from volumes of data," Commun. ACM, vol. 39, no. 11, pp. 27–34, 1996.
[22] "Lkit: A Toolkit for Natuaral Language Interface Construction." [Online]. Available: https://www.scm.tees.ac.uk/isg/website/downloads/lkit/overview (SCL-MSc).pdf. [Accessed: 15-Dec-2014].
[23] A.Lamb, M. Paul, and M. Dredze, "Separating Fact from Fear: Tracking Flu Infections on Twitter.," HLT-NAACL, 2013.
[24] K.B. Cohen and L. Hunter, Natural Language Processing and Systems Biology. 2004.
[25] B.Sujatha, V. Raju, and H. Shaziya, "A Survey of Natural Language Interface to Database Management System," Int. J. Sci. Adv. Technol., vol. 2, no. 6, pp. 56–61, 2012.
[26] P.M. Nadkarni, L. Ohno-Machado, and W. W. Chapman, "Survey of natural language processing," J. Am. Med. Inform. Assoc., vol. 18, no. 5, pp. 544–551, 2011.
[27] K.Gimpel, N. Schneider, and B. O'Connor, "Part-of-speech tagging for twitter: Annotation, features, and experiments," Proc. 49th …, no. 2, 2011.
[28] T.Divya and S. Agarwal, "A survey on Data Mining approaches for Healthcare," Int. J. Bio-Science Bio-Technology, vol. 5, no. 5, pp. 241–266, 2013.
[29] M.Kumari and S. Godara, "Comparative Study of Data Mining Classification Methods in Cardiovascular Disease Prediction," Int. J. Comput. Sci. Technol., vol. 2, no. 2, pp. 304–308, 2011.
[30] J.Soni, U. Ansari, D. Sharma, and S. Soni, "Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction," Int. Conf. Ind. Autom. Comput., vol. 17, no. April, 2014.
[31] D.Chaitrali and A. Sulbha, "Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques," Int. J. Comput. Appl., vol. 47, no. 10, pp. 44–48, 2012.
[32] S.Gupta, D. Kumar, and A. Sharma, "Data Mining Classification Techniques Applied for Breast Cancer Diagnosis and Prognosis," Indian J. Comput. Sci. Eng., vol. 2, no. 2, pp. 188–195, 2011.
[33] S.Sharma, U. S. Tim, M. Payton, H. Cohly, S. Gadia, J. Wong, and S. Karakala, "Contextual motivation in physical activity by means of association rule mining," Egypt. Informatics J., 2015.
[34] C.-H. Jen, C.-C. Wang, B. C. Jiang, Y.-H. Chu, and M.-S. Chen, "Application of classification techniques on development an early-warning system for chronic illnesses," Expert Syst. Appl., vol. 39, no. 10, pp. 8852–8858, 2012.
[35] M.Shouman, T. Turner, and R. Stocker, "Applying k-Nearest Neighbour in Diagnosing Heart Disease Patients," Int. J. Inf. Educ. Technol., vol. 2, no. 3, pp. 220–223, 2012.
[36] C.Chien and G. Pottie, "A Universal Hybrid Decision Tree Classifier Design for Human Activity Classification," in Engineering in Medicine and Biology Society (EMBC), Annual International Conference of the IEEE, 2012, pp. 1065 – 1068.
[37] S.S. Moon, S.-Y. Kang, W. Jitpitaklert, and S. B. Kim, "Decision tree models for characterizing smoking patterns of older adults," Expert Syst. Appl., vol. 39, no. 1, pp. 445–451, 2012.
[38] E.Avci, "A new intelligent diagnosis system for the heart valve diseases by using genetic-SVM classifier," Expert Syst. Appl., vol. 36, no. 7, pp. 10618–10626, 2009.
[39] O.Er, N. Yumusak, and F. Temurtas, "Chest diseases diagnosis using artificial neural networks," Expert Syst. Appl., vol. 37, no. 12, pp. 7648–7655, 2010.
[40] R.Dasa, I. Turkoglu, and A. Sengur, "Effective diagnosis of heart disease through neural networks ensembles," Expert Syst. Appl., vol. 36, no. 4, pp. 7675–7680, 2009.

[41] K.Lu, F. R. Liu, and C. F., "BBN-Based Decision Support for Health Risk Analysis," in Fifth International Joint Conference on INC, IMS and IDC, 2009.
[42] D.Everitt, Brian S., Landau, Sabine, Leese, Morven, Stahl, Cluster Analysis, 5th ed. Wiley, 2011.
[43] M.Silver, T. Sakara, H. C. Su, C. Herman, S. B. Dolins, and M. J. O'shea, "Case study: how to apply data mining techniques in a healthcare data warehouse," Heal. Inf. Manag., vol. 12, no. 2, pp. 155–164, 2001.
[44] J.Escudero, J. P. Zajicek, and E. Ifeachor, "Early Detection and Characterization of Alzheimer's Disease in Clinical Scenarios Using Bioprofile Concepts and K-Means," in 33rd Annual International Conference of the IEEE EMBS Boston, Massachusetts USA, 2011.
[45] L.Lenert, A. Lin, R. Olshen, and C. Sugar, "Clustering in the Service of the Public ' s Health," http://statweb.stanford.edu/~olshen/manuscripts/helsinki.PDF. .
[46] S.Belciug, F. Gorunescu, A. Salem, and M. Gorunescu, "Clustering-based approach for detecting breast cancer recurrence," in 10th International Conference on Intelligent Systems Design and Applications, 2011.
[47] T.Balasubramanian and R. Umarani, "An Analysis on the Impact of Fluoride in Human Health (Dental) using Clustering Data mining Technique," in Proceedings of the International Conference on Pattern Recognition, Informatics and Medical Engineering, 2012.
[48] S.Belciug, "Patients length of stay grouping using the hierarchical clustering algorithm," Annals of University of Craiova, Math. Comp. Sci. Ser., vol. 36, no. 2, pp. 79–84, 2009.
[49] M.E. Celebi, Y. A. Aslandogan, and R. P. Bergstresser, "Mining Biomedical Images with Density-based Clustering," in Proceedings of the International Conference on Information Technology: Coding and Computing (ITCC'05), 2005.
[50] V.Gupta and G. Lehal, "A survey of text mining techniques and applications," J. Emerg. Technol. Web Intell., vol. 1, no. 1, pp. 60–76, 2009.
[51] A.Culotta, "Towards detecting influenza epidemics by analyzing Twitter messages," Proc. First Work. Soc. Media Anal. - SOMA '10, pp. 115–122, 2010.
[52] E.Aramaki, S. Maskawa, and M. Morita, "Twitter Catches The Flu : Detecting Influenza Epidemics using Twitter The University of Tokyo The University of Tokyo National Institute of," Conf. Empir. Methods Nat. Lang. Process. EMNLP, pp. 1568–1576, 2011.
[53] H.Achrekar and A. Gandhe, "Predicting flu trends using twitter data," Comput. Commun. Work. (INFOCOM WKSHPS), pp. 702–707, 2011.
[54] K.Talvis, K. Chorianopoulos, and K. L. Kermanidis, "Real-Time Monitoring of Flu Epidemics through Linguistic and Statistical Analysis of Twitter Messages," 2014 9th Int. Work. Semant. Soc. Media Adapt. Pers., pp. 83–87, Nov. 2014.

AUTHORS

Haider M. Habeeb Al-Khumais
Department of Information Networks,
College of Information Technology,
University of Babylon.
IRAQ

Prof. Nabeel Hashim Al-A'araji
Head of Supervision and Scientific Evaluation Apparatus,
Ministry of Higher Education and Scientific Research. IRAQ.