From Social Media to Public Health Surveillance: Word Embedding based Clustering Method for Twitter Classification

Xiangfeng Dai, *Member, IEEE*, Marwan Bikdash, *Member, IEEE*, Bradley Meyer, *Member, IEEE*
Department of Computational Science and Engineering
North Carolina A&T State University
Greensboro, USA

***Abstract*—Social media provide a low-cost alternative source for public health surveillance and health-related classification plays an important role to identify useful information. In this paper, we summarized the recent classification methods using social media in public health. These methods rely on bag-of-words (BOW) model and have difficulty grasping the semantic meaning of texts. Unlike these methods, we present a word embedding based clustering method. Word embedding is one of the strongest trends in Natural Language Processing (NLP) at this moment. It learns the optimal vectors from surrounding words and the vectors can represent the semantic information of words. A tweet can be represented as a few vectors and divided into clusters of similar words. According to similarity measures of all the clusters, the tweet can then be classified as related or unrelated to a topic (e.g., influenza). Our simulations show a good performance and the best accuracy achieved was 87.1%. Moreover, the proposed method is unsupervised. It does not require labor to label training data and can be readily extended to other classification problems or other diseases.**

***Keywords*—Unsupervised Classification, Public Health, Twitter, Social Network, Big data, Surveillance, Word Embeddings, Word2Vec, Machine learning, Natural Language Processing, Clustering Process, Similarity Measure**

I. Introduction

Disease monitoring and tracking is of tremendous value, not only for containing the spread of contagious diseases but also for avoiding unnecessary public concerns and even panic [13]. Traditional public health surveillance is often limited by the time required to collect data. For example, the influenza surveillance system of the Center for Disease Control and Prevention (CDC) collects data from health departments, healthcare providers, clinics, professionals, health laboratories, and emergency departments [5]. This introduces about one-to-two week reporting delay [12] [20].

Social media platforms (such as Twitter, Facebook, Reddit, Tumblr, Pinterest and Instagram) have seen unprecedented growth in the era of big data. For example, Twitter, one of the most popular social network websites, which has been growing at a very fast pace. It has 284 million monthly active users, and 500 million tweets are sent per day [51]. Users often share their feelings, thoughts, activities, opinions and random details of their lives on social networks. Several studies have been demonstrated using social media as a low-cost alternative source for public health surveillance and health-related classification plays an important role to identify useful information [11].

TABLE I
Recent Classification Approaches in Public Health

Author	Year	Health Topic	Approach
Lampos et al.[30]	2010	influenza	I:KS
Scanfeld et al.[44]	2010	antibiotics	I
Achrekar et al.[1]	2011	influenza	I:KS
Collier et al.[7]	2011	influenza	II:SVM/NB
Aramaki et al.[15]	2011	influenza	II:SVM/NLP
Heaivilin et al.[25]	2011	dental pain	I:KS
Krieck et al.[29]	2011	disease outbreaks	I:KS
Paul et al.[38]	2011	syndromic	I
Prier et al.[37]	2011	tobacco/smoking	I:LDA
Signorni et al.[47]	2011	H1N1	II
Speriosu et al.[50]	2011	public mood	III
West et al. [42]	2012	drinking	I
Salathe et al.[46]	2012	health perceptions	III:PMI/SWN
Chew et al.[6]	2013	H1N1	I:KS
Kanhabua[28]	2013	disease outbreaks	I:KS
Myslin M [35]	2013	tobacco/smoking	II:NB/KNN/SVM
Copper al.[22]	2014	mental health	II
Huang et al.[26]	2014	tobacco/smoking	II:NB
Dai et al.[11]	2015	influenza	II:NB/NLP
Ofoghi et al.[36]	2016	public mood	II:NB, III:FN
Xiang et al.[52]	2016	health concerns	III

Category: (I) keywords-based (II) learning-based (III) lexicon-based
Methods: LDA: latent dirichlet allocation, NB: naive bayes,
SVM: support vector machine, KS: keywords/terms statistics
KNN: k nearest neighbors, NLP: natural language processing
PMI: pointwise mutual information, SWN: SentiWordNet
FN: FrameNet

II. Related Work

The recent approaches in public health are summarized into three main categories: keywords-based approaches, learning-based approaches and lexicon-based approaches (Table I).

1) Keywords-based Approaches: Most earlier studies rely on the keywords analysis such as word occurrences and word frequency. Lampos et al. [30] detected flu-related keywords to track influenza rates in the U.K. The flu-related keywords were used to learn a flu-score according to the weights of the

978-1-5386-1539-3/17/$31.00 ©2017 IEEE

keywords in each document. Culotta et al [10] collected flu-related keywords from Twitter and analyzed the correlation with national health statistics. Achrekar et al. [1] performed a similar study to collect tweets using flu-related keywords. They then counted the number of tweets at each time step per keyword to predict flu trends. Scanfeld et al. [44] investigated antibiotic abuse or antibiotic overuse. In their work, they tracked the tweets that mentioned antibiotics to find the antibiotics-related tweets. Heaivilin et al.[25] evaluated whether Twitter users broadcast information relating to dental pain and assessed the content of the information being communicated. A representative sample of tweets was collected using keywords/terms search. Prier et al.[37] used Latent Dirichlet Allocation (LDA) to analyze terms and topics from the entire dataset as well as from a subset of tweets created by querying general, tobacco use-related terms. The study of [42] was to examine the extent to which individuals tweeted about problem drinking, and to identify if such tweets corresponded with time periods when problem drinking was likely to occur. Tweets were identified that contained words reflective of problem drinking. Chew et al. [6] focused on the diversity in keyword lists for dynamics of change in circulated tweets for H1N1 virus. More similar studies are [1], [47], [54].

These methods rely on the keyword analysis and disregard context, grammar and even word order. They cannot sufficiently capture the complex linguistic characteristics of words [49].

2) Learning-based Approaches: These approaches have been intensively studied during the past decade, which require labeled data for training. Naive Bayes, k nearest neighbors (KNN), maximum entropy, and support vector machines (SVM) have been applied to a lot of health classification problems and achieved satisfactory results. Signorni et al. [47] employed a SVM classifier to determine the flu-related tweets. They then used the classified tweets to track the public sentiment with respect to H1N1 activities. Collier et al. [7] developed two supervised classifiers (SVM and Naive Bayes) to classify tweets for bio-surveillance. Unigrams, bigrams and regular expressions were used for feature selection. Myslin M [35] built machine learning classifiers using Naive Bayes, KNN, and SVM algorithms to classify tobacco-related tweets for sentiment analysis towards tobacco. Huang et al. [26] investigated e-cigarettes related tweets for public health mentions and smoking cessation mentions by using Naive Bayes machine learning methods.

Our previous work [11] presented a hybrid classification method that combines NLP preprocessing, rule-based classifiers and machine learning classifier. Our experimental results achieved a better performance than any single approach. The method improves the classification process because it takes advantage of the multiple approaches. The Naive Bayes model is used for the machine learning classifier, which assumes that all features are independent. It is computationally efficiency, but it avoids the contextual meanings of words.

In summary, the learning-based approaches suffer from the limitation of labeling training datasets, which requires experts to read the tweets and ascertain the category to which they belong. For a large-scale twitter data, it is difficult to manually label the large-scale training tweets. Therefore, it will be not easy to apply on other diseases.

3) Lexicon-Based Approaches: The other direction is knowledge-based, which is also called a dictionary method or knowledge-based method. It is considered to be a part of the unsupervised learning method. Speriosu et al. [50] used label propagation to incorporate labels from knowledge about word types encoded in a lexicon, which does not need annotated labels for training. Salathe et al. [46] proposed an unsupervised classification method for evaluation in health perceptions. The method combines PMI with a lexicon called SentiWordNet. Ofoghi et al. [36] implemented a simple unsupervised baseline emotion classier using a lexicon-based vector model form the FrameNet. However, the performance of these approaches depends on the lexicons [8].

4) Word Embedding Based Approach: Unlike other approaches in public health, we present a word embedding based clustering method. Word embedding is the one of the strongest trends in Natural Language Processing at this moment. It learns the continuous vector representation of words from context words and the vectors can represent the semantic information of words. A tweet can be represented as a few vectors and divided into clusters of similar words. According to similarity measures of all the clusters, the tweet then can be classified as related or unrelated to a topic (e.g., influenza). Our approach is unsupervised and does not require annotated data.

III. Research Design and Methodology

A. Architecture of Word Embedding Based Clustering Classification

Figure 1 shows the architecture of the proposed method, which involves the following 3 steps:

- **Step 1: NLP preprocessing** - Social media are informal, less structured, contain misspellings and nontextual information. NLP preprocessing is recommended to clean data for further analysis [11].
- **Step 2: Clustering process** - This step divides a tweet into clusters of words. Not all words in a tweet are helpful for classification. Some words actually distract identifying the topic and these words introduce bias. It is insensitive and fuzzy to use all words of a tweet for classification. However, it is too sensitive to use every single word.
- **Step 3: Similarity measure** - It identifies whether one of the clusters is related to flu according to cosine similarity measure. In this study, we consider if one of clusters is related to flu, the tweet then is related to flu.

B. NLP preprocessing

Tweets contain various noisy contents such as hash tags, slangs, abbreviations, links, etc. (Table II) and need to be tokenized or normalized, which is called text preprocessing [11]. It involves the following:

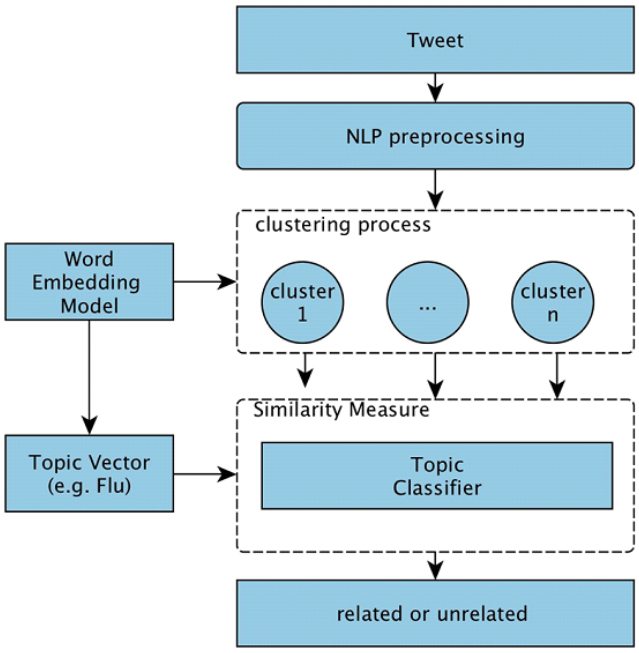

Fig. 1. Architecture of Word Embedding Clustering Classification

- Throw away special characters, punctuations, digits, HTML tags, quote, additional spaces, URLs and replies to users (@usernames) - They often appear in tweets, but do not contain any information for identifying topic.
- Capitalization, case folding - convert all words to lower case
- Correct spelling mistakes
- Nested words - filtering words by length.
- Stopwords removal - stopwords (such as prepositions, articles, a, is, the, with etc) have a high frequency of occurrence in the tweets. They do not carry much meaning and are not typically related to topic classification. Classifiers on average are more accurate without stopwords [24].

TABLE II
TWEETS CONTAIN NOISY CONTENTS

Feeling so miserable :(Having a flu fever Did not go to school DD: I will stay home, do some gentle stretching and nourish myself with herbal teas or veggie juices.
Turkey sandwich @__@, anyone? https://t.co/DZ.u
#_# Let's go @panthers! So excited to watch the Super Bowl!!! #SB50

C. Word Embedding Model

The word embedding model we used in this research is Word2Vec [32] [33]. It is an open source deep learning toolkit from Google based on word analogies that probes the finer structure of the word vector space. Once a word embedding model is created, each word in a tweet can be represented as a continuous space vector. For example, one can input a word (e.g., influenza) to the model, it returns a vector like this:

[-0.09220404 0.17788577 -0.15402232 -0.0221551 0.12370043
-0.11695234 -0.10891347 -0.02606416 0.12587149 0.11295457
0.05856625 0.09467842 0.08716864 0.0077392 -0.11854415
-0.13117599 0.11624993 0.10040938 -0.03850672 -0.17260635
-0.08380257 0.08499301 -0.01977218 -0.07082637 0.16041237
-0.10684048 -0.10911188 0.03601867 0.05466199 0.03672563
0.63015562 0.07612631 0.10935818 -0.08508652 0.16213417
0.15687755 0.01537053 0.32331052 -0.08125523 -0.10982797
0.12150883 -0.01701115 0.00365523 0.00532982 -0.0067344
-0.02842223 0.02850888 0.16477957 0.10853034 0.18133394]

The vectors catch semantic relationships between words [4] [9] [21] [33]. We collected a few words of 4 different topics (food, sports, weather and animals), then plotted their vectors in Figure 2. Since these vectors are high dimensional, we use a dimensionality reduction algorithm t-SNE [31] to visualize them in 2D [2].

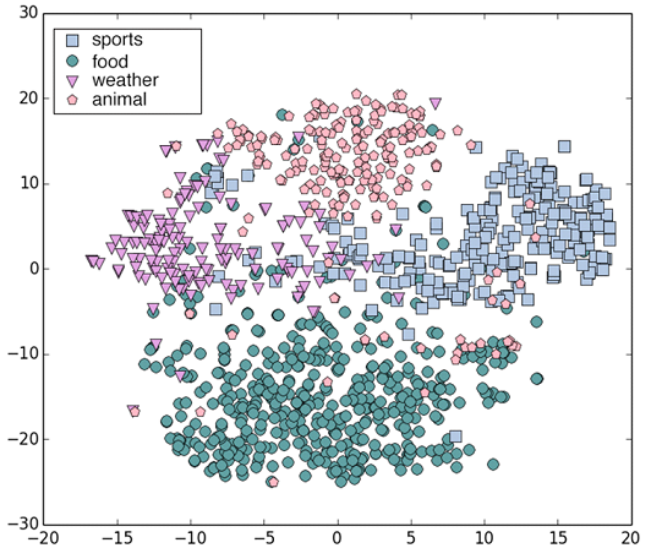

Fig. 2. words of different topics in a vector space

We can see that the words are clustering together with topics in a vector space. Our clustering process is inspired by this. A tweet can be divided into clusters of words, then identify whether each cluster is related to flu.

D. Clustering Process

This process is unsupervised, it divides a tweet into clusters of words (Figure 3). The algorithm is adapted from Chinese Restaurant Process (CRP) [3] of Dirichlet Process. The algorithm reads word by word from a tweet. The first word is added to the first cluster. The succeeding word has 2 options: add to existing cluster/clusters or add to a new cluster according to the similarity measure (equation 1) and an updated probability. Cosine similarity is a measure of similarity between two non-zero vectors of an inner product space that measures the cosine

of the angle between them. Two vectors are highly similar if their cosine similarity value is approaching 1.

$$\text{sim}(A, B) = \frac{\sum_{i=1}^{n}(A_i B_i)}{\sqrt{\sum_{i=1}^{n} A_i^2} \sum_{i=1}^{n} B_i^2} \quad (1)$$

where A and B are the vectors of length n

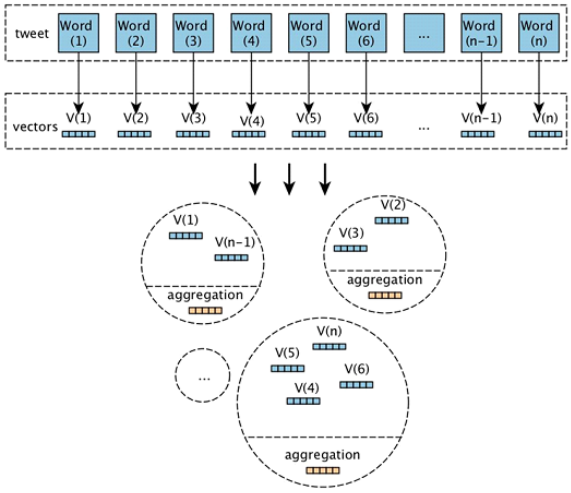

Fig. 3. Divide a tweet into clusters of words

Algorithm: Clustering Process

t: *an array of vectors that represents a tweet*
n: *number of clusters*
p: *probability*

1) $n = 1$
2) $p = 1/(1 + n)$
3) *append the first vector* $t[0]$ *to the first cluster* v_1
4) **loop** *the remaining vectors in* t
   a) *generate a random variable* r *between (0, 1)*
   b) **if** $r < p$
      i) *add a new cluster,* $n = n + 1$
      ii) *update* p
      iii) *append the current vector* $t[i]$ *to a new cluster*
   c) **else**
      i) *compute similarity* s_j *between* $t[i]$ *and each existing cluster* v_j
      ii) *append* $t[i]$ *to the cluster* v_j *where* $s_j = \max(s_1, s_2, ..., s_j)$
5) **return** *all clusters* $v_1, v_2, ..., v_j$

The number of clusters varies in different tweets. By practice, most cases end up with 3-5 clusters for a tweet, which satisfies our purpose of extracting topics. For example, the first tweet in Table II can be divided into the clusters of words in Table III.

TABLE III
CLUSTERING PROCESS

#	Cluster of words
C_1	`Feeling so having did not go I will stay do some gentle stretching myself`
C_2	`flu fever school`
C_3	`miserable`
C_4	`nourish herbal teas veggie`

E. Similarity Measure to Identify Related or Unrelated

We use similarity measure to identify whether the clusters of words are related to flu. We denote C (Cluster Vector) as the aggregation vector of a cluster. Averaging word embeddings of all words in a text is widely used to aggregate text embeddings [16] [19] [53]. Following this recommendation, one can calculate C by averaging all vectors in the cluster. We then choose a topic word, this topic word should obviously indicate a topic (e.g., influenza). We denote T as the vector of the topic word. One can calculate a similarity score from topic vector T and the cluster vector C (equation 1). For example, we use the vector of flu as topic vector T_{flu}. We then compute the similarity score s between T_{flu} and each cluster C (Table IV).

TABLE IV
SIMILARITY MEASURE FOR EACH CLUSTER

#	Similarity Score
s_1	0.134835
s_2	0.590763
s_3	0.106400
s_4	0.122449

We use $\rho_i \in \{0, 1\}$ to indicate the ith cluster whether related to flu. We denote τ as similarity threshold. If $s_i >= \tau$, $\rho_i = 1$,else $\rho_i = 0$. Table V shows an example when $\tau = 0.5$

TABLE V
IDENTIFY EACH CLUSTER

#	Related?
ρ_1	0
ρ_2	1
ρ_3	0
ρ_4	0

Since we consider if one of clusters is related to flu, the whole tweet is related to flu. $\rho_1 \cup \rho_2 \cup \rho_3 \cup \rho_4 = 1$. Therefore, this tweet is related to flu.

IV. EVALUATION AND RESULTS

A. Datasets

1) Test Set: We collected $2,270$ tweets through Twitter APIs and manually labeled them for testing our classifier. $1,070$ tweets are labeled as related to flu, the other $1,200$ tweets are labeled as unrelated to flu.

2) Pre-trained Vector Set: The quality of the word vectors increases significantly with amount of the training data. Google's pre-trained vector set [23] is used for our research purpose. It constructs a vocabulary from the training text data (Google News dataset) and then learns vector representation of words. The pre-trained word2vec model contains 3 million words.

B. Evaluation

The performance of the proposed method can be evaluated by four criteria calculated as the following equations:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{2}$$

In addition to accuracy, precision and recall are the most common measurements to evaluate classifiers.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{3}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{4}$$

The F1 measure is defined as the weighted harmonic mean of precision and recall:

$$\text{F1} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \tag{5}$$

where TP is the number of correctly classified as related tweets, TN is the number of correctly classified as unrelated tweets, FP is the number of false classified as related tweets, FN is the number of false classified as unrelated tweets, as defined in the Table VI.

TABLE VI
CONFUSION MATRIX

	Predicted Related	**Predicted Unrelated**
Actual Related	TP	FN
Actual Unrelated	FP	TN

The proportion of correctly classified observations is the estimated classification rate. The higher this proportion, the better the classifier. We evaluated the proposed method on 3 different similarity thresholds τ.

TABLE VII
EVALUATION OF WORD EMBEDDING CLUSTERING METHOD

τ	**Precision**	**Recall**	**F1**	**Accuracy**
0.8	99.6%	41.5%	65.2%	**65.2%**
0.7	99.6%	47.7%	64.5%	**75.3%**
0.6	96.2%	75.6%	84.6%	**87.1%**
0.5	77.1%	95.7%	84.7%	**84.6%**
0.4	55.1%	99.5%	70.9%	**55.1%**
0.3	48.9%	99.9%	65.7%	**50.8%**

The higher thresholds τ (0.7 and 0.8) have better precisions, but increase FN (the number of false classified as unrelated tweets), therefore, recalls get down. On the contrary, The lower thresholds τ (0.3 and 0.4) have better recalls, but increase FP (the number of false classified as related tweets).

A superior algorithm should tradeoff between precision and recall. F1 measure is defined as the weighted harmonic mean of precision and recall. It shows excellent performance (F1 and accuracy) when $\tau = 0.5$ and 0.6.

C. Comparison with Supervised Naive Bayes Method

We also applied the same dataset on the classical Naive Bayes classification method for baseline mechanism comparison. We implemented the Naive Bayes classifier with Python and scikit-learn machine learning library [45]. The dataset was randomly divided into a training set (75%), and a testing set (25%). The Table VIII shows the results of performance. Our proposed method is better than the standard Naive Bayes method when $\tau = 0.5$ and 0.6. The classical supervised Naive Bayes classification method is better than our proposed method when $\tau < 0.5$ and $\tau > 0.6$

TABLE VIII
PERFORMANCE OF NAIVE BAYES METHOD

Classifier	**Precision**	**Recall**	**F1**	**Accuracy**
Naive Bayes	73.4%	76.4%	74.9%	**75.6%**

V. CONCLUSIONS

In this paper, we summarize the existing classification approaches in public health. The approaches such as keywords or related words analysis, word occurrences analysis do not capture semantic similarity beyond a trivial level. Moreover, the approaches of supervised learning methods are difficult as it requires the labor intensive process of manually labeling a large corpus of training tweets. Furthermore, handcrafted patterns and external sources of structured semantic knowledge cannot be assumed to be available in all circumstances and for all domains in lexicon-based approaches.

We present a word embedding based clustering method for health-related classification using social media. Unlike other common approaches in public health, our method is based on word embeddings. The vectors of word embeddings are able to represent the semantic information of words. A tweet can be divided into clusters of similar words according to semantic similarity. According to the similarity measure of the clusters of words, the tweet can be classified.

We evaluated the performance in terms of precision and recall. The higher threshold τ get better precision. The lower threshold τ gets better recall. The algorithm gets good tradeoff between precision and recall when $\tau = 0.5$ and 0.6. We also evaluated the performance in F1 and accuracy. The performance is excellent when $\tau = 0.5$ and 0.6. We compared the proposed method to classical supervised Naive Bayes classification method. Our proposed method is better than the standard Naive Bayes method when $\tau = 0.5$ and 0.6.

Our method is unsupervised and it does not require labors to label data for training. It can be readily extended to other classification problems or other diseases. Once the temporal disease-related tweets are collected through the proposed classification method, we can use distance-based outliers method [13] for detecting outbreaks.

REFERENCES

[1] H. Achrekar, A. Gandhe, R. Lazarus, S. Yu and B. Liu, “Predicting Flu Trends using Twitter Data,” in Computer Communications Workshops (INFOCOM WKSHPS), 2011 IEEE Conferenceon, pp. 702–707, 2011

[2] R. Agrawal, A. Kadadi, X. Dai and F. Andres, "Challenges and opportunities with big data visualization", In Proceedings of the 7th International Conference on Management of computational and collective intElligence in Digital EcoSystems (pp. 169-173). ACM. 2015

[3] Aldous, David J. "Exchangeability and related topics." Ecole dete de Probabilites de Saint-Flour XIII—1983. Springer Berlin Heidelberg, 1985. 1-198.

[4] Y. Bengio, H. Schwenk, "Neural probabilistic language models", In Innovations in Machine Learning, 2006

[5] CDC, “Overview of Influenza Surveillance in the United States,” [online] Nov 2016, http://www.cdc.gov/flu/weekly/overview.htm (Accessed: 21 Nov 2016)

[6] C. Chew and G. Eysenbach, “Pandemics in the age of twitter: Content analysis of tweets during the 2009 h1n1 outbreak,” PLoS ONE, 2013

[7] N. Collier, N. Son and N. Nguyen, “OMG U got flu? Analysis of shared health messages for bio-surveillance”, Journal of Biomedical Semantics, 2011

[8] E. Cambria, B. Schuller, Y. Xia and C. Havasi, "New avenues in opinion mining and sentiment analysis", IEEE Intelligent Systems, 28(2):15–21, 2013.ter to Examine Smoking Behavior and Perceptions of Emerging Tobacco Products," Journal of Medical Internet Research, 2013

[9] R. Collobert, J. Weston, L. Bottou, M. Karlen, K, Kavukcuoglu, "Natural language processing (almost) from scratch", Journal of Machine Learning Research, 2011

[10] A. Culotta, “Detecting influenza outbreaks by analyzing Twitter messages” in Proc. 2010 Conf. on Knowledge Discovery and Data Mining, 2010

[11] X. Dai and M. Bikdash, “Hybrid Classification for Tweets Related to Infection with Influenza,” Proceedings of the IEEE SoutheastCon 2015, April 9 - 12, 2015, Fort Lauderdale, Florida, 2015

[12] X. Dai and M. Bikdash, "Trend Analysis of Fragmented Time Series for mHealth Apps: Hypothesis Testing Based Adaptive Spline Filtering Method with Importance Weighting", IEEE Access, 2017

[13] X. Dai and M. Bikdash, "Distance-based Outliers Method for Detecting Disease Outbreaks using Social Media", Proceedings of the IEEE SoutheastCon 2016, Norfolk, VA, 2016

[14] S. Doan, B. Vo, and N.Collier, “An analysis of Twitter messages in the 2011 Tohoku Earthquake”, eHealth 2011 Conference, Spain, 2011

[15] A. Eiji, S. Maskawa, and M. Morita. "Twitter catches the flu: detecting influenza epidemics using Twitter." In Proceedings of the conference on empirical methods in natural language processing, pp. 1568-1576. Association for Computational Linguistics, 2011

[16] M. Faruqui, D. Jesse, K. Sujay, "Retrofitting word vectors to semantic lexicons." arXiv preprint arXiv:1411.4166 2014.

[17] E. Gabrilovich and S. Markovitch, “Overcoming the Brittleness Bottleneck using Wikipedia: Enhancing Text

[18] E. Gabrilovich and S. Markovitch, "Wikipedia-based Semantic Interpretation for Natural Language Processing," Journal of Artificial Intelligence Research, 2009

[19] S. Gershman and T. Joshua. "Phrase similarity in humans and machines." In Proceedings of the 37th Annual Conference of the Cognitive Science Society. 2015.

[20] J. Ginsberg, M. Mohebbi, R. Patel, L. Brammer, M. Smolinski, and L. Brilliant, "Detecting influenza epidemics using search engine query data", Nature, 457(7232):1012–1014, 2008.

[21] Y. Goldberg, O. Levy, "word2vec Explained: deriving Mikolov et al.’s negative-sampling word-embedding method", arXiv preprint arXiv:1402.3722, 2014 Categorization with Encyclopedic Knowledge”, Proceedings of the Twenty-First National Conference on Artificial Intelligence, 2006

[22] G. Coppersmith, C. Harman and M. Dredze, "Measuring Post Traumatic Stress Disorder in Twitter," AAAI Publications, Eighth International AAAI Conference on Weblogs and Social Media, 2014

[23] Google word2vec. https://code.google.com/archive/p/word2vec/ (Accessed: 25 Oct. 2016)

[24] S. Hassan, Y. He and H. Alani. "Semantic sentiment analysis of twitter." International Semantic Web Conference. Springer Berlin Heidelberg, 2012.

[25] N. Heaivilin, Gerbert B, Page JE and Gibbs JL., “Public health surveillance of dental pain via Twitter”, J Dent Res. 2011 Sep;90(9):1047-51. doi: 10.1177/0022034511415273. Epub, 2011

[26] J. Huang, R. Kornfield, G. Szczypka and S. Emery, "A cross-sectional examination of marketing of electronic cigarettes on Twitter," Tobacco Control, 2014

[27] H. Kanayama and T. Nasukawa, "Fully automatic lexicon expansion for domain-oriented sentiment analysis," Proceeding EMNLP ’06 Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, Pages 355-363, 2006.

[28] N. Kanhabua and W. Nejd, “Understanding the diversity of tweets in the time of outbreaks,” WWW ’13, pp. 1335-1342, 2013

[29] M. Krieck, J. Dreesman , L. Otrusina and K. Denecke, “A new age of public health: Identifying disease outbreaks by analyzing tweets”, Proceedings of Health WebScience Workshop, ACM Web Science Conference, 2011

[30] V. Lampos and N. Cristianini. “Tracking the flu pandemic by monitoring the Social Web,” In 2nd IAPR Workshop on Cognitive Information Processing (CIP 2010), pp. 411-416, 2010

[31] L. Maaten and G. Hinton,"Visualizing High-Dimensional Data Using t-SNE", Journal of Machine Learning Research, 9(Nov):2579-2605, 2008

[32] T. Mikolov, K. Chen, G. Corrado and J. Dean , "Efficient Estimation of Word Representations in Vector Space", ICLRWorkshop, 2013

[33] T. Mikolov, S. Sutskever, C. Kai, S. Greg and D. Dean. "Distributed representations of words and phrases and their compositionality." In Advances in neural information processing systems, pp. 3111-3119. 2013.

[34] C. Manning, P. Raghavan and H. Schütze, An Introduction to Information Retrieval, Online edition, Cambridge University Press, 2009

[35] Myslin M, Zhu SH, Chapman W and Conway M., "Using Twit

[36] B. Ofoghi, M. Mann, K. Verspoor. Towards early discovery of salient health threats: A social media emotion classification technique. InPacific Symposium on Biocomputing. Pacific Symposium on Biocomputing 2016 (Vol. 21, p. 504).

[37] K. Prier, M. Smith, C. Giraud-Carrier and C. Hanson, "Identifying health-related topics on twitter: an exploration of tobacco-related tweets as a test topic," Proceeding SBP’11 Proceedings of the 4th international conference on Social computing, behavioral-cultural modeling and prediction, Pages 18-25, 2011.x

[38] M. Paul and M. Dredze, “You Are What You Tweet: Analyzing Twitter for Public Health”, AAAI Publications, Fifth International AAAI Conference on Weblogs and Social Media, 2011

[39] A. Passos and J. Wainer, “Wordnet-based metrics do not seem to help document clustering”, Proceedings of the of the II Workshop on Web and Text Intelligente, Brazil, 2009

[40] J. Pennington, R. Socher and C. Manning, "GloVe: Global Vectors for-Word Representation", Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014

[41] A. Vo and C. Ock, "Assessing Vaccination Sentiments with Online Social Media: Implications for Infectious Disease Dynamics and Control," PLoS Comput Biol. 2011

[42] J. West, P. Hall, C. Hanson, K. Prier, C. Giraud-Carrier, E. Neeley and M. Barnes, “Temporal variability of problem drinking on Twitter”, Open Journal of Preventive Medicine, 2012.

[43] World Health Organization, “Influenza fact sheet,” [online] March 2014, http://www.who.int/mediacentre/factsheets/fs211/en/ (Accessed: 1 Jan 2016)

[44] D. Scanfeld, Scanfeld V and Larson EL., “Dissemination of health information through social networks: twitter and antibiotics”, Am J Infect Control. 2010

[45] Scikit-learn, Machine Learning in Python, http://scikit-learn.org/stable/ (Accessed: 10 Nov. 2016)

[46] M. Salathe and S. Khandelwal, "Sentiment classification: A Combination of PMI, sentiWordNet and fuzzy function," Proceeding ICCCI’12 Proceedings of the 4th international conference on Computational Collective Intelligence: technologies and applications - Volume Part II, Pages 373-382, 2012

[47] A. Signorini, A. Segre1 and P. 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, 2011

[48] R. Soebiyanto, F. Adimi and R. Kiang, "Modeling and predicting seasonal influenza transmission in warm regions using climatological parameters", PLoS One, 2010

[49] M. Sofean and K. Denecke, “Medical Case-Driven Classification of Microblogs : Characteristics and Annotation,” IHI’12, Miami, Florida, USA, 2012
[50] M. Speriosu, S. Nikita, S. Upadhyay and J. Baldridge. "Twitter polarity classification with label propagation over lexical links and the follower graph." In Proceedings of the First workshop on Unsupervised Learning in NLP, pp. 53-63. Association for Computational Linguistics, 2011.
[51] Twitter, Inc., “Twitter usage,” [online] 2015, https://about.twitter.com/company (Accessed: 1 Jan 2016)
[52] J, Xiang, S. Chun and J. Geller. "Knowledge-Based Tweet Classification for Disease Sentiment Monitoring." In Sentiment Analysis and Ontology Engineering, pp. 425-454. Springer International Publishing, 2016.
[53] L. Yu, H. Karl, P. Blunsom and S. Pulman. "Deep learning for answer sentence selection." arXiv preprint arXiv:1412.1632 2014.
[54] Q. Yuan; E. Nsoesie; B. Lv; G. Peng; R. Chunara and J. Brownstein, “Monitoring influenza epidemics in china with search query from baidu,” PLoS ONE, 2013