Trends on Health in Social Media: Analysis using Twitter Topic Modeling

Mohsen Asghari
Department of Computer Engineering
and Computer Science CECS
University of Louisville
Louisville, KY, USA
m0asgh02@louisville.edu

Daniel Sierra-Sosa
Department of Computer Engineering
and Computer Science CECS
University of Louisville
Louisville, KY, USA
d.sierrasosa@louisville.edu

Adel Elmaghraby
Department of Computer Engineering
and Computer Science CECS
University of Louisville
Louisville, KY, USA
adel@louisville.edu

***Abstract*—There is a growing interest on social networks for topics related to Healthcare. In particular, on Twitter, millions of tweets related to healthcare can be found. These posts contain public opinions on health, and allow to understand how is the popular perception on topics such as medical diagnosis, medicines, facilities, and claims. In this paper we present an adaptive system designed using 5 layers. The system contains a combination of unsupervised and supervised algorithms to track the trends of health social media. As it is based on a word2vec model, it also captures the correlation of words based on the context, improving over time, enhancing the accuracy of predictions and tweet tracking. In this work we focused on United States data and use it to detect the trending topics of each state. These topics are followed including new social network contributions. The supervised algorithm implemented is a Convolutional Neural Network (CNN) in conjunction with the Word2Vect model to classify and label new tweets, assigning a feedback to the topic models. The results of this algorithm present an accuracy of 83.34%, precision of 83%, recall 84% and F-Score of 83.8% when evaluated. Our results are compared with two state of the art techniques demonstrating an advantage that can be leveraged for further improvements.**

Keywords—Health Tweets, LDA, Classification, Deep learning

I. INTRODUCTION

Free text in healthcare is classified in two groups namely Biomedical and Clinical text. Biomedical text includes books, abstracts and articles. Clinical text comprises medical personnel reports, such as the diagnosed patient pathologies, personal and medical history [1]. However, healthcare related texts are also available in social network as free/unstructured text constituting yet one more group of healthcare related text.

Social network sites such as Twitter or Facebook provide a communication platform. Recent studies show that on Twitter users tend to share advice on health related information [2, 3]. These sources contain general public health beliefs, and have the potential to expand the understanding of topics such as diagnosis, medicines and claims. There are reported almost 140 potential healthcare uses of Twitter [4], the most common uses are: disaster alerting and response, diabetes management, drug safety alerts from the Food and Drug Administration, biomedical devices data capture and reporting, shift bidding for nurses and other healthcare professionals, diagnostic brainstorming, rare diseases tracking and resource connection, smoking cessation assistance, infant care tips to new parents and post-discharge patient consultations and follow-up care [4].

As an example of how Social Networks portrays medical information, even when the safety and effectiveness on the human papillomavirus (HPV) vaccine have been proved, Social Network trends reports low efficacy in some countries including United States. However, this negative opinion and information is induced by news, celebrities or trend-setters, and they impact the public trust on this particular topic [5].

Due to the impact of social networks on healthcare there is a growing interest on model development and its analysis. Prieto et. al. (2014) present the analysis from the value of tweets related to healthcare, this study uses machine learning techniques in order to evaluate these tweets; they gather the data based on regular expression in Spain and Portugal, then they narrow the document to four selected categories "Pregnancy", "Depression", "Flu" and "Eating Disorder" they utilized two traditional machine learning methods KNN, SVM [6].

Prier et. al. (2011) propose a model based on LDA model and set the model to generate 250 topics, they select "Tobacco" as a topic for validate the model [7]. Two other studies, one in U.K. and the other in U.S. find the correlation between twitter's sentiment analysis and the quality of healthcare services [8,9].

In this work, an automated Twitter topic modeling method is presented. This system will not be seeded or initialized and will improve from positive and negative feedback. The system as developed collects tweets and by use Latent Dirichlet Allocation (LDA) as unsupervised model, labeling each tweet, identifying patterns. This method is intended to process tweets related to public beliefs on healthcare. We design a CNN combined with Word2Vect model. The Word2Vect model was trained on 7,821 medical abstracts as a first iteration of learning. The results from this training enrich the vocabulary related with healthcare, improve the method of detecting related tweets, and improve the general topic modeling for detection on new tweets.

II. METHODOLOGY

Text data in healthcare can be categorized in three domains Clinical, Biomedical and Social, each of them gathered by separate group of people. Biomedical text is collected by scientists and medical doctors who have experience in the laboratory. The clinical notes generated by medical personnel refer to a specific patient, biomedical text on the other hand refers to the general population of patients. Social network text provides is related with an idea, people advices or specific information, but the truthfulness of these information is not guaranteed.

In this study tweet data related to healthcare was collected for one month, and the hashtag topic correlation was modeled using the LDA technique. Also we implement a method to

978-1-5386-7568-7/18/$31.00 ©2018 IEEE

detect new documents related with the topics in order to gather future data.

A. Dataset

A dataset of 144,922 tweets in English was collected, the employed keywords were: healthcare, health, doctors, homecare, digitalhealth, and digital health. During the data gathering a number of tweets related with job offers were collected, as these data is out of the scope from the current study, tweets containing the keywords: job, Job, (hir\w+), career and hiring were excluded. In this work our interest was to collect information inside the U.S, therefore another limitation was imposed, the tweets were filtered based U.S state name, County name and Federal Information Processing Standard Publication known as FIPS code.

After filtering the data 37,910 health care related tweets match the criteria. The final dataset contains tweets from 43 U.S. states, being California with 5,923 tweets the most populated and South Dakota with 43 tweets the least populated. The data was collected from October 2018 for one month.

B. Data Preparation

In order to prepare the data for processing the first step is to remove line characters such as enter characters and Tabs, then remove all the quotes, hashtags, number and non-characters by using regular expression patterns. Also, all the URLs ("*https?://[A-Za-z0-9./]+*") and references *(@[A-Za-z0-9]+*) were removed with regular expression patterns.

The second step consist in the conversion of all the text to tokens, for this purpose genism library was employed [10]. Then the stop words such as "*a, an, the*" were removed, these words don't add any value to the text analysis, adding noise to the data.

The last step is the Lemmatizing and Stemming process, these allows for obtaining the required features. By using Stemming, the inflectional endings, prefix and suffix from words are removed. With the Llemmatization, a morphological analysis of the words is performed, this method requires to predefine a vocabulary for the target language; this process was conducted based on a dictionary provided by genism [10]. In this system the selected words were 'NOUN', 'ADV', 'ADJ' and 'VERB'.

The data pre-processing was performed using two lists: the first list named "V" is the vocabulary of text containing each word with their corresponding frequency, and the second one named "W" is the list of the tokenized words for each document. Based on W and V a Bag of Words (BOW) was created. Therefore, the documents are represented by a list of vectors with the length of "V". From the preprocessing a matrix with each document on its rows and the vocabulary on the columns is obtained.

C. Automated Topic Detection

The topic modeling consists on find patterns or relevant words inside a bag of unlabeled documents. To perform this task a LDA based on a three-level hierarchical Bayesian model was implemented [11].

The LDA was trained on three unlabeled databases. This model allows for finding were the data is denser, thus define the topic. Like the unsupervised techniques, the challenge is to define the number of clusters that will represent the topics; this implies defining a metric for the density, leading to the optimal number of topics (clusters) in the corpus. Two metrics from Natural Language Processing were employed, Perplexity and log-likelihood defined in (1) and (2) respectively.

$$perplexity(D_{test}) = exp\left\{-\frac{\sum_{d=1}^{M}\log p(w_d)}{\sum_{d=1}^{M}N_d}\right\} \qquad (1)$$

Were M is the number of documents, $p(\,)$ is the probability of a given word w_d, and N_d is the total number of words per document.

$$L = \log(P(w_d|w_1,w_2,\ldots,w_i) = log\left(\frac{freq(w_1\ldots w_i)}{freq(w_1\ldots w_{i-1})}\right) \qquad (2)$$

Were P is the conditional probability of a given word w_d.

The extreme values of these metrics are selected as the best number of topics for the corpus. The LDA model was applied for different number of components, this number is interpreted as the best number of topics that describe the corpus. The model was run for 2, 5, 10, 20, 50, 100, 200 topics. Given the metrics, five was the best number of topics with a Perplexity of 798.806 and value of -442786.843 for the log-Likelihood.

To validate these metrics Principal Component Analysis (PCA) was conducted. In Figure 1 the Distance map of the topic modeling obtained by the PCA is presented. With this model it can be observed that using the five selected topics, without overlap; in this figure the size of each circle represents the population of the each topic in the corpus. Table I presents the distribution of each topic.

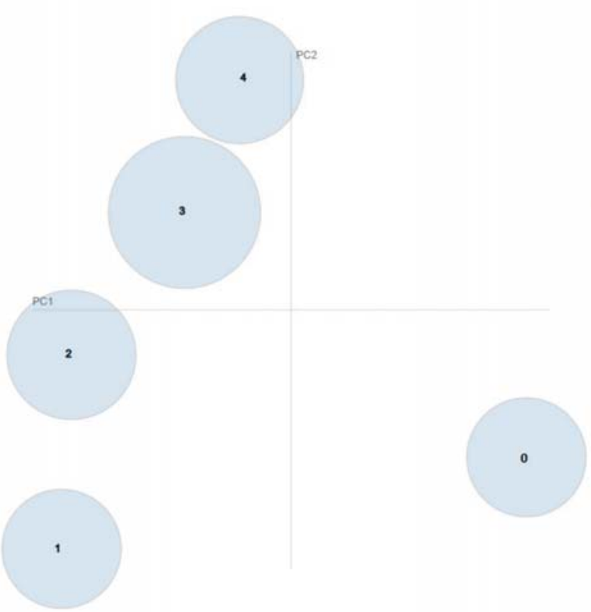

Figure 1 Inter-Topic Distance Map

We represent the most relevant terms based on "LDAvis" a web based tool [17]. In Figures 2 through 6 the 30 most relevant terms inside each topic and respective frequencies are presented, the blue bars represent the frequency of that term over all the documents and the red bars are sorted by the terms relevance inside each topic.

TABLE I. FREQUENCY AND PROPORTION OF EACH TOPIC IN 37,910 COLLECTED TWEETS

Topic	Frequency	Proportion
0	7,459	19.68
1	6,070	16.01
2	7,994	21.09
3	9,919	26.16
4	6,468	17.06
TOTAL	37,910	

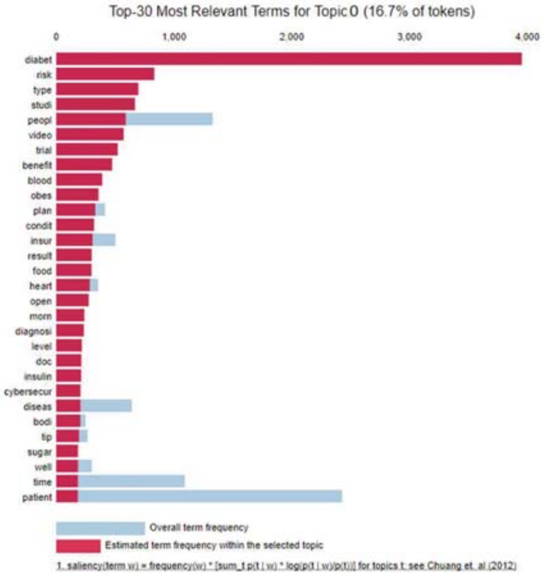

Figure 2 Top 30 Most Relevant Terms for Topic 0

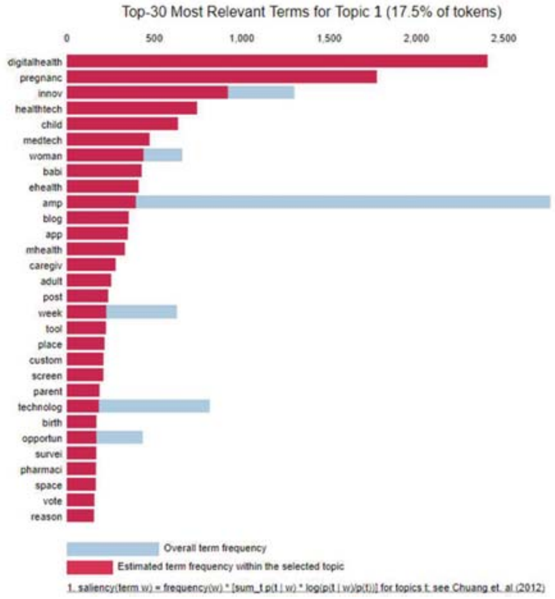

Figure 3 Top 30 Most Relevant Terms for Topic 1

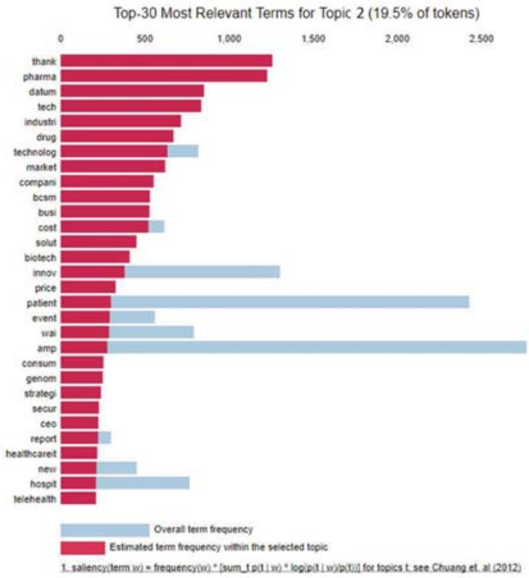

Figure 4 Top 30 Most Relevant Terms for Topic 2

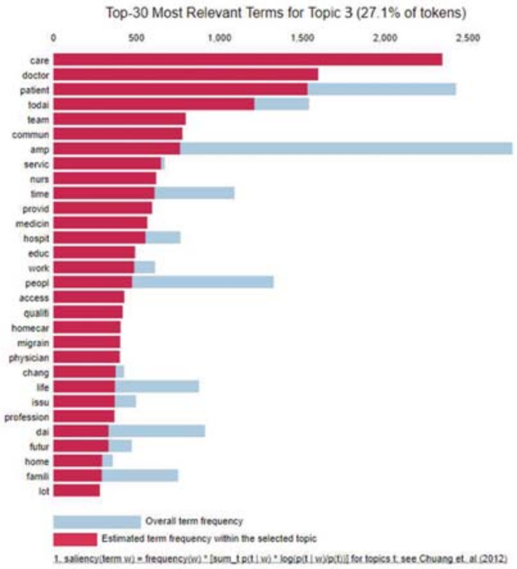

Figure 5 Top 30 Most Relevant Terms for Topic 3

D. Prediction Metrics

LDA used as unsupervised model to categorize the tweets. Five is the optimal number of topics based on the perplexity and log-likelihood metrics. The selected topics described in Table II are diabetics, digital health, drug market, health care services, and cancer and research. LDA assign 5 scores (number of topics) to a tweet, each of them represents the similarity score with predefined topics, then we select the highest score and labeled it as the corresponding Topic. To validate our results we used LDA labeling as ground truth and compare it with predicted labels. Therefore, utilizing the LDA model transfer the data from an unsupervised to a supervised technique.

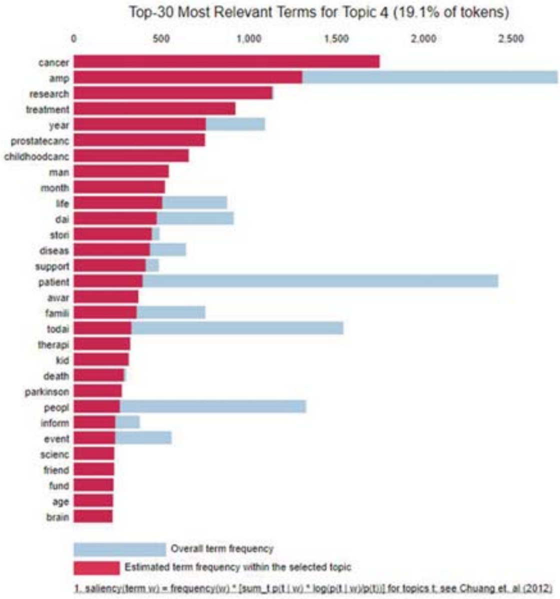

Figure 6 Top 30 Most Relevant Terms for Topic 4

TABLE II. GENERATED TOPIC BASED ON LDA MODEL

Topic 0	Topic 1	Topic 2	Topic 3	Topic 4
diabet	digitalhealth	thank	care	cancer
risk	pregnanc	pharma	doctor	amp
type	innov	datum	patient	research
studi	healthtech	tech	todai	treatment
peopl	child	industri	team	year
video	medtech	drug	commun	prostatecanc
trial	woman	technolog	amp	childhoodcanc
benefit	babi	market	servic	man
blood	ehealth	compani	nurs	month
obes	amp	bcsm	time	life

The classification results are evaluated by four metrics: accuracy (3), recall (4), precision (5) and F1-score (6), on these equations *TP* is true positives, *TN* is true negatives, *FP* is false positives and *FN* is false negatives. Precision represents the exactness of the classifier and measure how many predicted labels are related, recall represents how many true records are predicted and the F-scores quantify the harmonic average from the precision and recall.

$$accuracy = \frac{TP + TN}{(TP + FP + TN + FN)} \qquad (3)$$

$$precision = \frac{TP}{(TP + FP)} \qquad (4)$$

$$recall \;= \frac{TP}{(TP + FN)} \qquad (5)$$

$$F1 - Score \;= 2 * \frac{Precision * Recall}{(Precision + Recall)} \qquad (6)$$

E. Architecture and classification results

The proposed model is built by layers. On the first layer the data is collected using a python tool called Teewpy [12]. The second layer contains the cleaning and preprocessing methods described, converting the tweets to vectors that can be processed. The third layer is a Word2Vec method that create a matrix based on the vector received by the last layer and used that matrix for initializing a neural network to predict the labeled the tweet.

A CNN classifier is the fourth layer, were unseen tweets coming from the Word2Vec are labeled. Usually, a sequence modeling is related to Recurrent Neural Network (RNN), however the results indicate a different perspective. The Convolutional Neural Network (CNN) provides notable results in NLP [13]. Yih et al. 2011 applied CNN on semantic parsing [14], Shen 2014 utilized it for query retrieval [15], KalchBanner 2014 used it for sentence modeling [16], and Yoon Kim 2014 connected Word2Vec model with CNN [13]. As part of this research project we were exploring classification models that we can use as part of an adaptive system. Feature selection is one of the challenges and the Convolutional Neural Networks (CNN) are appropriate as they do not require a priori feature selection. A limitation of CNNs is that they require a fixed input size, as the tweets are limited to 280 characters we can use padding for shorter tweets preserving the fixed input sizes. Our architecture comprises three convolutional layers with a 128, 64 and 32 kernel sizes respectively. The system has a drop out of 0.5. It was iterated by 200 epochs using batches of 100 records with a learning rate of 0.00001. The weights were obtained using Adam-Optimizer.

On this layer some tweets may not fit into the selected topics, these are flagged as unlabeled data and stored in an independent dataset. If the tweets fit into the selected topics they are classified and labeled accordingly.

In the fifth layer, unlabeled data is feed to a LDA model and new topics may be created. The CNN model will be trained again, updating both the topics and the Word2Vec model.

In Figure 7 the system's architecture is presented. This architecture is intended to provide an active learning of topics and reinforce the CNN model classification.

To compare the proposed system SVM and CNN techniques were implemented independently to predict and label new tweets. Nonetheless, the prediction capacities from both methods were limited due to the unbalanced datasets. The results using the prediction metrics are presented on Table II.

TABLE III. MODELS COMPARISON

Algorithm	**Accuracy**	**Precision**	**Recall**	**F-Score**
SVM	39.5%	67.6%	39.5%	34.9%
CNN	57%	58.8%	55.1%	56%
CNN-static	83.34%	83%	84%	83.8%

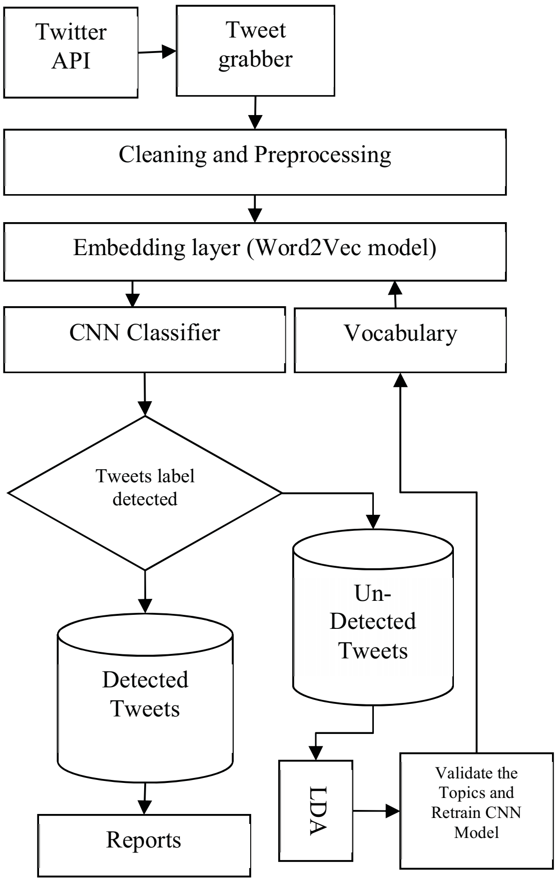

Figure 7 Proposed Architecture to track the Twitter Topics

III. Results

The collected tweets were processed using the LDA model, and with the perplexity and log-likelihood metrics the topics were defined, Table III presents the 10 keywords of each of the topics. An example of the collected tweets classification is presented in Table IV, where the similarity percentage to each topic is depicted.

After topic modeling, we represent the distribution of each topic over the different States by using a heat map. Red areas represent those States where more people is interested on that topic. This report provides the possibility to depict the time evolution of a topic over the different States.

Each tweet belongs to a location, represented by the corresponding abbreviation of State name. To calculate the percentage of interest from each state in a particular topic we divided the population of tweets in that state by the population of the specific topic in that state.

In Figure. 8 it can be observed that California is the state more interested on topic 0 "Diabetics" with 85.22% and Virginia with 30.98% is the less interested. Figure. 9 shows that New York is the state more interested in topic 1 "Digital Health" with 77.3% and Wisconsin with 29.13% the less interested.

In the case of topic 2, the most frequent word cannot be used as the topic label, but the contained words on this topic allows to label it as "Market Drug and Pharmacy". On Figure. 10 the distribution for this topic is depicted, here California is the more interested state with a 100% and Oklahoma with 21.89% the less interested.

TABLE IV. Sample Tweet and Topic Similarity based on the LDA model

Sample Tweets	***0***	***1***	***2***	***3***	***4***
Pregnant Women Should Eat More Fish	10%	**59%**	10%	10%	10%
Healthcare examines How to Protect Against New Mobility and #IoT Security Threats in #Healthcare by Tech #cybersecurity # healthdata #healthIT #HIT	**55%**	5%	30%	5%	5%
Reframing and addressing horizontal violence as a workplace quality improvement concern #digitalhealth #innovation #health tech #medtech #li #fb	3%	22%	3%	**50%**	20%
A patient with BMI of 32 with #metabolic disease #diabetes #hypertension high #cholesterol etc would need #BariatricSurgery amp #MetabolicSurgery more than a patient with #BMI 37 without these conditions Unfortunately an outdated BMI threshold of 35 is still our criteria'# healthtech	**40%**	10%	2%	15%	33%
how the White House healthcare blueprint attempts to inject value into the US #healthcare system #pbm #drugprices #pharma'	20%	22%	**51%**	3%	3%

The topic 3 "Care" is shown in Figure. 11. California and New York are the more interested on this topic with 80% and Virginia with 41% is the less interested. The last topic represented by the labels "cancer", "amputation" and "research" depicted in Figure. 12, shows that there is a generalized interest on these topics being California, Texas, New York, Wisconsin and Ohio the most interested states.

IV. Discussion and Conclusions

In this paper the implementation of a tweet analysis system is presented. The system comprises from the collection process to the classification of new gathered documents. To classify the tweets a Convolutional Neural Network (CNN) in conjunction with a Word2Vect model was implemented. This system provides feedback to the topics, allowing for the generation of new topics. The classification algorithm has an accuracy of 83.34%, precision of 83%, recall 84% and F-Score of 83.8%. The trends on topics analyzed with this system can be depicted and reported in a localized manner, in this paper tweets from the U.S were used as Case Study, the heat maps corresponding to those trends were presented and described.

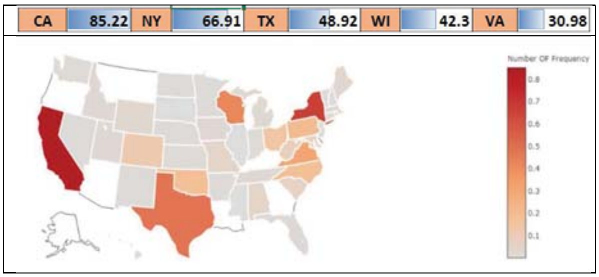

Figure 8 Distribution of Topic 0 over the United States

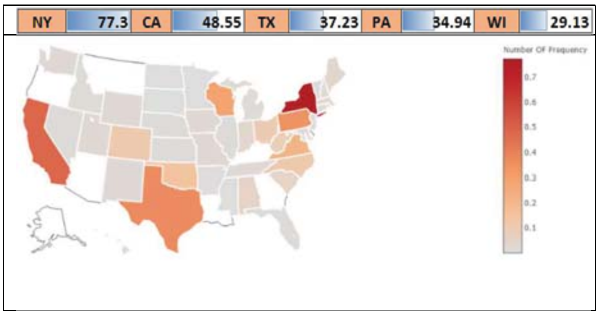

Figure 9 Distribution of Topic 1 over the United States

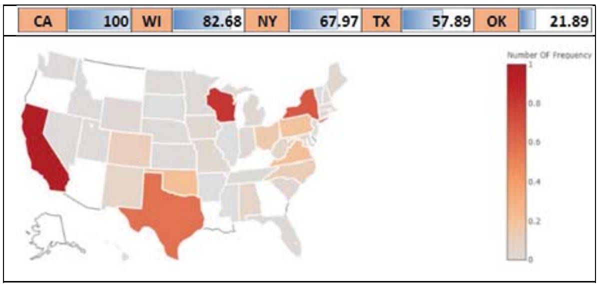

Figure 10 Distribution of Topic 2 over the United States

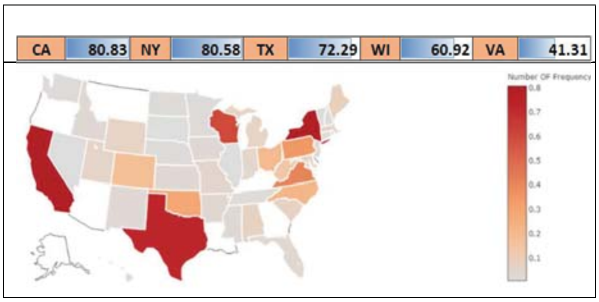

Figure 11 Distribution of Topic 3 over the United States

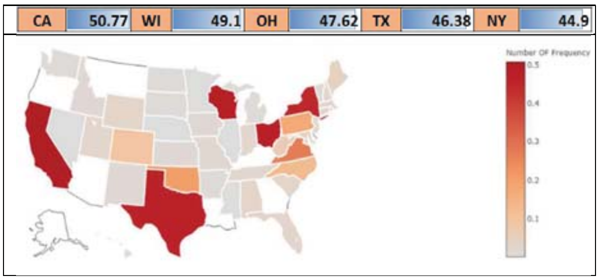

Figure 12 Distribution of Topic 4 over the United States

REFERENCES

[1] Reddy, C. K., & Aggarwal, C. C. (2015). Healthcare data analytics. Chapman and Hall/CRC.

[2] Scanfeld, D., Scanfeld, V., & Larson, E. L. (2010). Dissemination of health information through social networks: Twitter and antibiotics. American journal of infection control, 38(3), 182-188.

[3] Prier, K. W., Smith, M. S., Giraud-Carrier, C., & Hanson, C. L. (2011, March). Identifying health-related topics on twitter. In International conference on social computing, behavioral-cultural modeling, and prediction (pp. 18-25). Springer, Berlin, Heidelberg.

[4] Chapman, B. E., Lee, S., Kang, H. P., & Chapman, W. W. (2011). Document-level classification of CT pulmonary angiography reports based on an extension of the ConText algorithm. Journal of biomedical informatics, 44(5), 728-737.

[5] Surian, D., Nguyen, D. Q., Kennedy, G., Johnson, M., Coiera, E., & Dunn, A. G. (2016). Characterizing Twitter discussions about HPV vaccines using topic modeling and community detection. Journal of medical Internet research, 18(8).

[6] Prieto, V. M., Matos, S., Alvarez, M., Cacheda, F., & Oliveira, J. L. (2014). Twitter: a good place to detect health conditions. PloS one, 9(1), e86191.

[7] Prier, K. W., Smith, M. S., Giraud-Carrier, C., & Hanson, C. L. (2011, March). Identifying health-related topics on twitter. In International conference on social computing, behavioral-cultural modeling, and prediction (pp. 18-25). Springer, Berlin, Heidelberg.

[8] Greaves, F., Laverty, A. A., Cano, D. R., Moilanen, K., Pulman, S., Darzi, A., & Millett, C. Tweets about hospital quality: a mixed methods study. BMJ Qual Saf. 2014 Oct; 23 (10): 838–46. doi: 10.1136/bmjqs-2014-002875.

[9] Hawkins, J. B., Brownstein, J. S., Tuli, G., Runels, T., Broecker, K., Nsoesie, E. O., ... & Greaves, F. (2015). Measuring patient-perceived quality of care in US hospitals using Twitter. BMJ Qual Saf, bmjqs-2015.

[10] Rehurek, R., & Sojka, P. (2010). Software framework for topic modelling with large corpora. In In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks.

[11] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of machine Learning research, 3(Jan), 993-1022.

[12] Roesslein, J. (2015). Tweepy. Python programming language module.

[13] Kim, Y. (2014). Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882.

[14] Yih, W. T., Toutanova, K., Platt, J. C., & Meek, C. (2011, June). Learning discriminative projections for text similarity measures. In Proceedings of the Fifteenth Conference on Computational Natural Language Learning (pp. 247-256). Association for Computational Linguistics.

[15] Shen, Y., He, X., Gao, J., Deng, L., & Mesnil, G. (2014, April). Learning semantic representations using convolutional neural networks for web search. In Proceedings of the 23rd International Conference on World Wide Web (pp. 373-374). ACM.

[16] Kalchbrenner, N., Grefenstette, E., & Blunsom, P. (2014). A convolutional neural network for modelling sentences. arXiv preprint arXiv:1404.2188.

[17] Sievert, C., & Shirley, K. (2014). LDAvis: A method for visualizing and interpreting topics. In Proceedings of the workshop on interactive language learning, visualization, and interfaces (pp. 63-70).