March - April 2020
ISSN: 0193-4120 Page No. 2123 - 2128

Develop Extensive Approaches of Topic Models through Assessment of Social Data For Healthcare

Dr. A. Rajesh[1], K. Kishore[2], MD Asim[3]

[1]*Associate Professor, Dept. of Computer Science and Engineering,Vels Institute of Science, Technology and Advanced Studies (VISTAS), Chennai.*
[2]*Research Scholar, Dept. of Computer Science and Engineering, Vels Institute of Science, Technology and Advanced Studies (VISTAS), Chennai.*
[3]*Assistant Professor, Dept. of Computer Science and Engineering, Dr.K.V.Subba Reddy College of Engineering for Women, Kurnool, A.P.*

Article Info
Volume 83
Page Number: 2123 - 2128
Publication Issue:
March - April 2020

Abstract
Data clustering in social networks is an emerging need for categorizing the user's data according to similarity of topics. Twitter is a great source for providing platform to social media users for sharing their views or opinions, or exchanges the ideas. Social media provides a large amount of health-related data and tends to more scope for its research in the direction of early monitoring and predict risk factors. Existing system uses two problems in the development of healthcare intelligent system by social media data, these problems are namely health transition detection and health transition prediction. Health topic models are widely used techniques in text mining for extraction of social data features. Traditional health topic models, namely Latent semantic indexing (LSI), Probabilistic latent semantic indexing (PLSI), Latent dirichlet allocation (LDA), non-negative matrix factorization (NMF) are used for extraction of latent variables or hidden topics of social data. As a part of the research work an attempt will be made to develop Ailment Topic Aspect Model (ATAM) is a new latent model that can be dedicated for capturing the topics from heath tweet data. It aims to extract health-related topic transitions by minimizing the prediction error on topic distributions between consecutive posts at different time and geographic granularities. Healthcare costs are driving the demand for big data-driven healthcare applications. Technology decision-makers in healthcare systems can't ignore the increased efficiencies, the attractive economics, and the rapid pace of innovation that can now be applied to delivering and paying for healthcare. Social system is a great source for sharing views or conversations by different people on health-related topics such as types of diseases, symptoms and medicines. Extraction of sentiments from such kind of social data is an emerging need in healthcare and recent research shown social recommended solutions for healthcare. Social data clustering is performed by LSI, PLSI, LDA, and NMF and they deliver health clustering results without knowing the knowledge of prior cluster tendency. Estimation of number of clusters for given social data is known as cluster tendency. This problem is intractable by exiting topic models with the information of tweet-term matrix of social health related data. In proposed frame work, it can be addressed by finding topic-document dense matrix through assessment of similar topic. The similarity features are computed and tweets are re-ordered according to similarity features during assessment of social data cluster tendency. Visual approaches are proposed for visualizing health clusters that useful for knowing prior number of clusters and improve the efficiency proposed topic models in social data health clustering.

Article History
Article Received*: 24 July 2019*
Revised*: 12 September 2019*
Accepted*: 15 February 2020*
Publication*: 18 March 2020*

Keywords: *Public health, ailments, social media, topic models.*

Published by: The Mattingley Publishing Co., Inc.

March - April 2020
ISSN: 0193-4120 Page No. 2123 - 2128

1. INTRODUCTION:

Public health-related topics are difficult to identify in large conversational datasets like Twitter. This study examines how to model and discover public health topics and themes in tweets. Over recent years, social network sites (SNS) like Facebook, Twitter have transformed the way individuals interact and communicate with each other across the world. These platforms are in turn creating new avenues for data acquisition and research. Such web-based applications share several common features, and while there are slight variations in actual implementations,
each service enables users to (1) create a public profile, (2) define a list of other users with whom they share a connection, and (3) view and discover connections between other users within the system. Since SNS allow users to visualize and make public their social networks, this promotes the formation of new connections among users because of the social network platform. Not only do SNS enable users to communicate with other users with whom they share explicit social connections, but with a wider audience of users with whom they would not have otherwise shared a social connection. Twitter, in particular, provides a medium whereby users can create and exchange user-generated content with a potentially larger audience than either Facebook or Twitter.
Twitter is a social network site that allows users to communicate with each other in real-time through short, concise messages (no longer than 140 characters), known as \tweets." A user's tweets are available to all of his/her \followers," i.e., all others who choose to subscribe to that user's profile. 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 [11]. 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].
Our challenges are: (i) identify health-related tweets, (ii) determine when health-related discussions on Twitter transitions from one topic to another, (iii) capture different such transitions for different geographic regions.

2. PROPOSED METHODOLOGY:

1. 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 non textual 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.

2. NLP preprocessing

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

Published by: The Mattingley Publishing Co., Inc.

March - April 2020
ISSN: 0193-4120 Page No. 2123 - 2128

- 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 stop words [14].

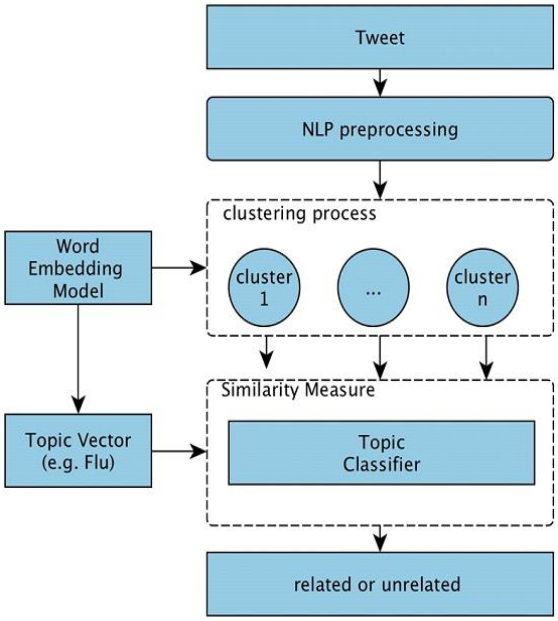

Fig. 1. Architecture of Word Embedding Clustering Classification

3. Clustering Process

This process is unsupervised, it divides a tweet into clusters of words (Figure 2). 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 nonzero 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.

$$Sim(A,B) = \frac{\sum_{i=1}^{n}(AiBi)}{\sqrt{\sum_{i=1}^{n} A_i^2 \ \sum_{i=1}^{n} B_i^2}} \qquad (1)$$

Where A and B are the vectors of length *n*

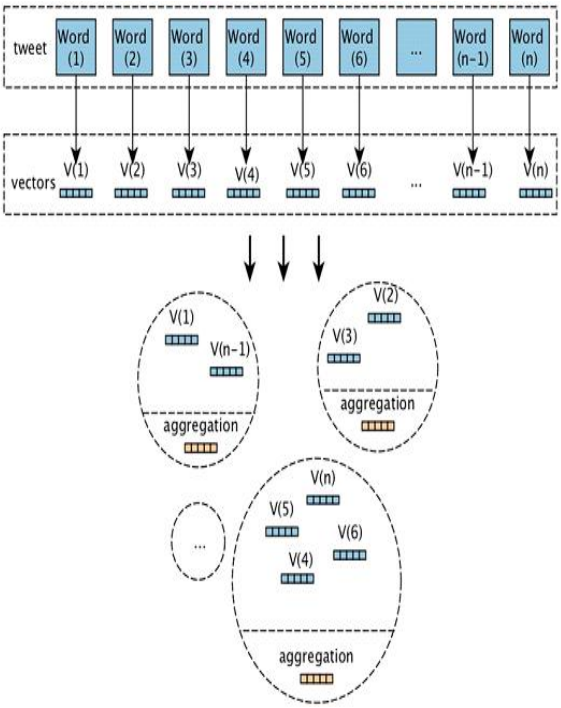

Fig. 2. Divide a tweet into clusters of words

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 v1
4. loop the remaining vectors in t
i) generate a random variable r between (0,1)
ii) if r<p
a) add a new cluster, n=n+1

Published by: The Mattingley Publishing Co., Inc.

March - April 2020
ISSN: 0193-4120 Page No. 2123 - 2128

b) update p

c)append the current vector t[i] to a new cluster

iii) else

a) compute similarity s_j between t[i] and each existing cluster v_j

b) append t[i] to the cluster v_j where sj = 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 I.

CLUSTERING PROCESS

S.No	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

3. RESULTS AND DISCUSSIONS:

1. Datasets

1) ***Test Set:*** We collected 2270 tweets through Twitter APIs and manually labeled them for testing our classifier. 1070 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 [13] 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.

2. Evaluation

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

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (2)$$

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

$$Precision = \frac{TP}{TP + FP} \quad (3)$$

$$Recall = \frac{TP}{TP + FN} \quad (4)$$

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

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} \quad (5)$$

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 II.

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 .

TABLE III

EVALUATION OF WORD EMBEDDED 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%

Published by: The Mattingley Publishing Co., Inc.

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 (F1and accuracy) when τ = 0:5 and 0:6.

3. Comparison with Supervised Naive Bayes Method

We also applied the same dataset on the classical Naïve Bayes classification method for baseline mechanism comparison. We implemented the Naive Bayes classifier with Python and scikit-learn machine learning library [14]. 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 τ = 0:5 and 0:6: The classical supervised Naïve Bayes classification method is better than our proposed method when τ < 0:5 and _ > 0:6

TABLE IV
PERFORMANCE OF NAVIE BAYES METHOD

Classifier	Precision	Recall	F1	Accuracy
Navie Bayes	73.4%	76.4%	74.9%	75.6%

4. CONCLUSION:

We have demonstrated the application of a topic model in discovering relevant clinical concepts and structuring a patient's medical record. The imposed statistical structure was then used for case-based information retrieval of similar patients. The analysis of the system in terms of precision and recall is challenging due to the exhaustive requirements of generating a gold standard. However, the generation

and release of such a set is a point of future work. We are augmenting this system with additional query mechanisms including demographics and lab values.

Additionally, we are pursuing query templates consisting of different weightings of domains to answer pre-defined clinical questions. Phrasal discovery and analysis for improved topic learning is also underway. Our approach can be quickly applied to any type of clinical document corpus as it requires no customization. However, for document corpora with relatively limited variation in words and grammar, a customized, knowledge-driven approach may also be appropriate. Such a system would likely take much longer to create, but could ultimately better capture clinical notions of similarity.

We plan to explore new applications for topic models in clinical reporting and have begun implementing techniques for 1) topic-driven problem list generation; and 2) systems that analyze the expression of a topic over time for modeling the progression of a disease process.

References:

[1] SumitSidana, SihemAmer-Yahia, Marianne Clausel, M Rebai, Son T Mai, Massih-Reza Amini, “Health Monitoring on Social Media over Time”, IEEE Transaction on Knowledge and Data Engineering, Vol. 30, No.8, August 2018

[2] L. Jiang, M. Yu, M. Zhou, X. Liu, and T. Zhao, “Target-dependent twitter sentiment classification,” in Proc. Annu. Meeting Assoc. Comput. Linguistics: Human Language Technol., 2011, pp. 151–160.

[3] L. Hong and B. D. Davison, “Empirical study of topic modeling in twitter,” in Proc. 3rd Workshop Social Netw. Mining Anal., 2010, pp. 80–88.

[4] S. R. Chowdhury, M. Imran, M. R. Asghar, S. Amer-Yahia, and C. Castillo, “Tweet4act: Using incident-specific profiles for classifying

March - April 2020
ISSN: 0193-4120 Page No. 2123 - 2128

crisis-related messages," in 10th Proc. Int. Conf. Inform. Syst. Crisis Response Manag., 2013.

[5] C. Chemudugunta, P. Smyth, and M. Steyvers, "Modeling general and specific aspects of documents with a probabilistic topic model," in Proc. Int. Conf. Neural Inf. Process. Syst., 2006, pp. 241–248.

[6] Y. Wang, E. Agichtein, and M. Benzi, "TM-LDA: Efficient online modeling of latent topic transitions in social media," in Proc. ACM SIGKDDInt. Conf. Knowl. Discovery Data Mining, 2012, pp. 123–131.

[7] D. M. Blei and J. D. Lafferty, "Dynamic topic models," in Proc. Int. Conf. Mach. Learn., 2006, pp. 113–120.

[8] C. X. Lin, Q. Mei, J. Han, Y. Jiang, and M. Danilevsky, "The joint inference of topic diffusion and evolution in social communities," in Proc. 11th Int. Conf. Data Mining, 2011, pp. 378–387.

[9] O. J. Dyar, E. Castro-S_anchez, and A. H. Holmes, "What makes people talk about antibiotics on social media? a retrospective analysis of twitter use," J. Antimicrobial Chemotherapy, vol. 69, pp. 2568–2572, 2014.

[10] L. Manikonda and M. D. Choudhury, "Modeling and understanding visual attributes of mental health disclosures in social media," in Proc.CHI Conf. Human Factors Comput. Syst., 2017, pp. 170–181.

[11] A. Culotta, "Estimating county health statistics with twitter," in Proc. SIGCHI Conf. Human Factors Comput. Syst., 2014, pp. 1335–1344.

[12] L. Hong, B. Dom, S. Gurumurthy, and K. Tsioutsiouliklis, "A time-dependent topic model for multiple text streams," in Proc. 17th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2011, pp. 832–840.

[13] M. J. Paul and M. Dredze, "You are what you tweet: Analyzing twitter for public health," in Proc. Int. Conf. Weblogs Social Media, 2011.

[14] A. Saha and V. Sindhwani, "Learning evolving and emerging topics in social media: A dynamic NMF approach with temporal regularization," in Proc. 5th ACM Int. Conf. Web Search Data Mining, 2012, pp. 693–702.