
Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Information Processing and Management 57 (2020) 102340

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

# A topic modeling framework for spatio-temporal information management

Mohsen Asghari*, Daniel Sierra-Sosa, Adel S. Elmaghraby

*Department of Computer Science and Engineering, University of Louisville, KY, USA*

ARTICLE INFO

*Keywords:*
Spatio-temporal real time analysis
Traceability
Topic modeling
Visualization
Artificial intelligent
Transfer learning

ABSTRACT

Real-time processing and learning of conflicting data, especially messages coming from different ideas, locations, and time, in a dynamic environment such as Twitter is a challenging task that recently gained lots of attention. This paper introduces a framework for managing, processing, analyzing, detecting, and tracking topics in streaming data. We propose a model selector procedure with a hybrid indicator to tackle the challenge of online topic detection. In this framework, we built an automatic data processing pipeline with two levels of cleaning. Regular and deep cleaning are applied using multiple sources of meta knowledge to enhance data quality. Deep learning and transfer learning techniques are used to classify health-related tweets, with high accuracy and improved F1-Score. In this system, we used visualization to have a better understanding of trending topics. To demonstrate the validity of this framework, we implemented and applied it to health-related twitter data from users originating in the USA over nine months. The results of this implementation show that this framework was able to detect and track the topics at a level comparable to manual annotation. To better explain the emerging and changing topics in various locations over time the result is graphically displayed on top of the United States map.

## 1. Introduction

In Social Network Sites (SNS) and Location Base Social Network (LBSN) systems, such as Twitter, user's behavior can be considered as sensors sending data from different locations over time. Such data is known to exhibit Spatio-temporal characteristics. The constant tweet stream at a high speed is a data-flow that represents the data velocity (Katal, Wazid, & Goudar, 2013). This real-time Spatio-temporal data source is a rich and valuable source for public health studies and researches (Asghari, Sierra-Sosa, & Elmaghraby, 2018; Hawkins et al., 2016; Martinez-Millana, Fernandez-Llatas, Bilbao, Salcedo, & Salcedo, 2017; Tursunbayeva, Franco, & Pagliari, 2017).

This data source is utilized for many purposes such as disaster detection, public opinion evaluation, outbreak prediction, and Spatio-temporal analysis. The studies show that mining twitter messages could give us information faster than other media such as TV, this could represent an advantage for early disaster detection. For example, an earthquake in the US in 2011, the London riots, Hurricane Irene, or influenza-like surveillance were assessed using SNS data (Gesualdo et al., 2013; Thom, Bosch, Koch, Wörner, & Ertl, 2012).

Analysis of public views is the other usage of this data. Companies and service-providers receive many reviews and comments related to their products and services from customers. The majority of these comments will be shared on twitter. Doing Analysis on

* Corresponding author.
*E-mail address:* m0asgh02@louisville.edu (M. Asghari).

https://doi.org/10.1016/j.ipm.2020.102340
Received 10 December 2019; Received in revised form 11 June 2020; Accepted 11 June 2020
Available online 06 July 2020
0306-4573/ 

them would have added value for the service-providers. For example, Griffis et al. (2014) used Twitter data to evaluate hospitals across the USA. Besides, the effectiveness of this data could be studied to predict outcomes, Alessa and Faezipour (2019) are utilized this kind of data to predicting preliminary Flu outbreak.

The location-based characteristic of twitter data opens new horizons for research, such as tracking the Flu outbreaks over the USA to detect the influenza hot-spots by Zadeh, Zolbanin, Sharda, and Delen (2019). The Spatio-temporal analysis adds value to the results by including new dimensions to the analysis. Finding hidden patterns and correlation from SNS and LBSN data is a demanding requirement to extract semantics.

Topics extraction from messages over time is challenging, due to the dynamic of conversations in SNS and LBSN. Our Framework enhances the topic extraction step by introducing a new topic modeling selection procedure. Koylu (2019) utilized LDA for topic extraction and evaluation; However, in this research, we demonstrate that LDA-MALLET and LSA in our case study (Healthcare Database) outperform the LDA model. in other research (Serban, Thapen, Maginnis, Hankin, & Foot, 2019) they perform classification techniques to detect the evaluation of messages. Classification methods need accurate annotated datasets with many records which is costly as it is time consuming. In our research, we combined the unsupervised and supervised evaluation methods to enhance the results. This research addresses the need for a more robust procedure that allows for the topic extraction more accurately in dynamic environments.

In this research, we propose a framework for Twitter data topic modeling. Our contributions are:

- Introducing a robust decision-making procedure for selecting a model to explain the dynamic conversation topics. We design an adaptive framework to use gained knowledge for improving the results over time.
- Using neural network transfer learning techniques to enhance the framework ability to detect unrelated messages over twitter data streams.
- Create an automatic deep cleaning method to enhance the quality of data to perform better classification in a noisy environment.

We augmented the quality of data through deep cleaning and using deep transfer learning to reduce the negative effect of non-related healthcare messages, to enhance the results. In this paper, we propose a framework that includes: fully automatic data collection, cleaning, and noise cancellation modules. In addition to an automatic topic and event detection procedure in a dynamic environment.

## 2. Related work

Twitter data used for many purposes such as detecting opinions, risk analysis, and event detection (Hwang, Wang, Cao, Padmanabhan, & Zhang, 2013; Laylavi, Rajabifard, & Kalantari, 2017; Sierra-Sosa, Asghari, Gordon, & Elmaghraby, 2019; Zhao, Chen, Lu, & Ramakrishnan, 2015). Recently Zahra, Imran, and Ostermann (2020) utilize the twitter data to classify users in two groups, eyewitness and those that were not. They designed a model to detect eyewitness tweets to extract valuable and accurate information for analysis. Time and location are the characteristics shared between eyewitness generated tweets and others. Involving these two properties in the analysis will add value to it. This applied to different practices, such as management, recommendation, and monitoring systems in different areas including natural disasters, city traffic, and healthcare.

Thom et al. (2012) used LBSN data to detect anomalies and visualize them. They applied their system and synthesized information from an earthquake in the US in 2011, the London riots, and hurricane Irene. They proposed a visual analysis method that assesses Spatio-temporal anomalies and achieves more dynamic results than they obtained when performing clustering on Twitter data streams.

We can use Twitter data not only for early disaster detection but also for covering the news and understanding the peoples' perceptions. Guo, Vargo, Pan, Ding, and Ishwar (2016) presented a methodology including two approaches, dictionary-based text analysis and Latent Dirichlet Allocation (LDA), as an unsupervised topic modeling approach to give better insight to journalists. They conducted empirical research on two Twitter datasets for the 2012 presidential election. They concluded that both techniques generated valuable results, such as detecting which users on twitter discussed both Barack Obama and foreign affairs as well as a tax policy problem related to the presidential candidate Mitt Romney.

Since velocity is an inevitable part of the LBSN datasets, a data management system is required. Hwang et al. (2013) proposed a scalable pipeline infrastructure and database schema to collect and analyze data. They presented a case study for flu risk surveillance. They collected their data using twitter stream data with a pipeline infrastructure that they processed for early detection of a flu epidemic in the United States. Rathore, Ahmad, Paul, Hong, and Seo (2017) proposed a Hadoop based framework to perform real-time analysis on social media data. They made this framework general, allowing the detection of diverse events ranging from natural disasters or healthcare information.

Detecting the subject and topic of short messages on twitter is a common point focus on most published research related to twitter analysis. Creating a framework for automatic topic detection in a data source has been researched for several applications, such as Krestel, Fankhauser, and Nejdl (2009) tagging recommendations, and Griffiths and Steyvers (2004) identifying scientific topics in documents.

Recently Serban et al. (2019) proposed a real-time social media processing framework named SENTINEL by focusing on disease surveillance. Although, they annotated 9353 twitters manually to apply classification techniques which would limit the scalability of the system in case of the new event appear in the system. In other recent research Koylu (2019), proposed a framework to do the prepossessing, topic extraction, and pattern exploration. Topic extraction utilizes Latent Dirichlet Allocation (LDA) techniques and

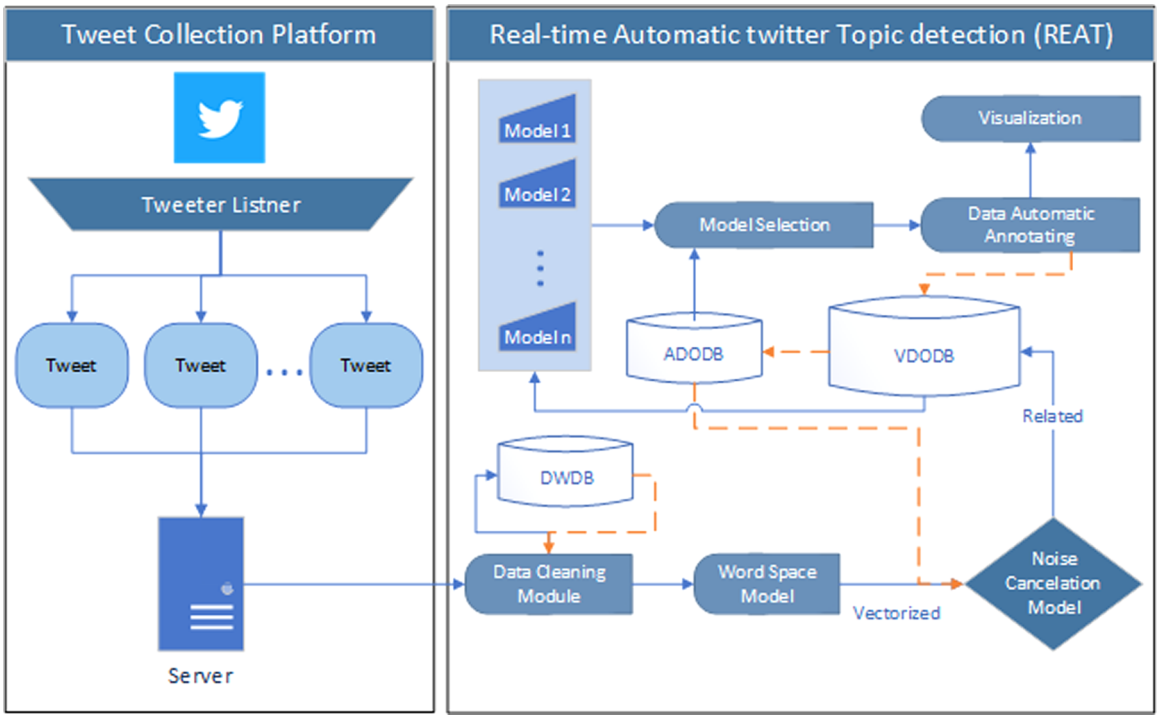


**Fig. 1.** Real-time Twitter topic modeling architecture and software modules.

cosine similarity to calculate the pairwise similarity between topics.

In contrast to these two recent kinds of research, we propose a procedure to enhance topic detection and extraction in a dynamic environment; this characteristic enables the framework and improves the model performance making it more efficient in comparison to manual annotation.

## 3. Materials and methods

The dynamic behavior from Twitter data given the velocity characteristic in different times and locations makes the analysis challenging. Designing a supervised methodology would require lots of resources and time for annotating the data. This implies we need an unsupervised learning approach for tracking the topics. Fig. 1 illustrates the proposed Twitter stream analysis framework for automatic pre-processing, unsupervised topic modeling, knowledge representation, and visualization.

This system formed by two major modules, *"Tweet Collection Platform"* and *"REal time Automatic twitter Topic detection"* (*REAT*). *Tweet Collection Platform* is an implementation of a listener using Tweepy API (Tweepy, 2019). Collected twitters will store on a server to analyze by the REAT module.

The *REAT* consists of five sub-modules, including Data Cleaning, Word Space, Related Tweets, Model Selector, and Visualization; each of these modules will explain in detail.

The *REAT* defines three databases. First is the *"Bag of Word DataBase"* (*BWDB*), to store the tokens. This will grow incrementally, helping to improve the tokenizer procedure, which implemented in the Data cleaning module. Second is the *"Vectorized DOcuments DataBase"* (*VDODB*), this database will store the data after transforming, and removing the noises, they will process by the *"Word Space Model"* and *"Noise Cancellation Model"* respectively. Third is the *"Annotated DOcument DataBase"* (*ADODB*), to boost the topic modeling selector. This database store annotated twitters. ADODB will help the framework to calculate homogeneity, completeness, and V-measure; these metrics will discuss in more detail later. To generate the ground truth, we annotated a random sample of collected data by assigning a team of graduate students and stores in ADODB.

We create a bridge between Vectorized Twitter Database and Annotated Database to make a feedback response to the model selector; to provide confidence level, that would be used as a threshold to decide which part of the data sent to the ADODB. We recommend having an expert interfere in this level for increasing the quality of ADODB by checking on annotated messages.

### 3.1. Data cleaning module

Fig. 2 illustrates the tweet cleaning and tokenization process. In this research, we focus on healthcare in the USA. We set English in API's language properties, and assign health-related hashtags for the first iteration of the cleaning process. As we targeted the USA, so we used API's Location properties, as well as Census information. The other aspect of cleaning is retweet messages, which will not

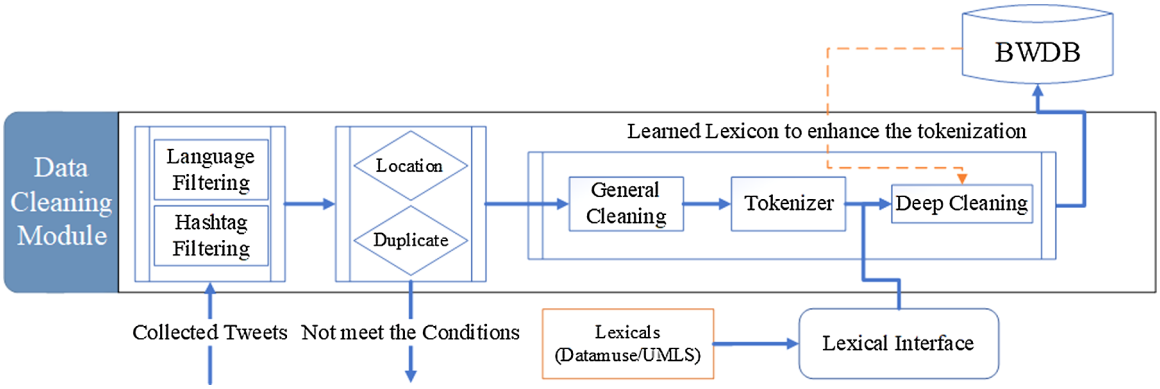


**Fig. 2.** Data Cleaning Module - Input collected Tweets and return list of words cleaned in two layers of general and deep.

add any value to analysis, so we removed all those duplicated messages.

We used Regular Expressions to perform general cleaning. We created several expressions to remove IPs, URLs, HTML tags, and repeated characters, as an example "helllllooo" which correspond to "hello". Besides, stemming, lemmatization, and tokenization process are applied.

To enhance data quality, we implement extra cleaning steps. Implementing these steps require metadata and expert knowledge. We used trustworthy lexicons, such as Beeferman (2019), and UMLS Bodenreider (2004) provided by NIH. Deep cleaning module, use the Datamuse lexicon for spell checking. Also, the UMLS lexicon helped us to solve the joined and separated words in the healthcare domain, which we called this problem the *"Connected Word"* problem. These are affecting the bag of words and the model's reliability, for example, "breast cancer" should transform to "breastcancer".

Finally, we store the words in the *DWDB* and use it as an internal lexical, which gives feedback to the next iteration of the tokenization process. This database would enhance and speed up the process of deep cleaning by using previously generated knowledge of words.

### 3.2. Word space model

A word-space model is a method that converts text to a vector. There are several ways to implement this model. One solution is utilizing one-hot encoded vectors, but this representation will not be satisfactory to our purpose. Since each vector needs an array size of created vocabulary (BWDB), this technique would not be practical due to the size of the vocabulary and dynamic of the environment.

Continuous Bag of Words (CBOW)(Mikolov, Chen, Corrado, & Dean, 2013) technique, can solve the vocabulary size problem by defining the vector dimensionality. Fig. 3, represents the structure of the CBOW Deep learning model used as Word2Vec in this study. $n$ represent window size, and $W$ is a word, $W(t)$ will be a word inside a document in position $t$, and create a sequence of $W(t-n)...W(t-1)W(t+1)W(t+n)$. We assumed $V$ is the size of vocabulary and $D$ is the dimensionality; therefore, we create a matrix size of $V$*$D$. This matrix is known as Embedding Layer.

To create the Embedding layer, define a loss function is an essential step. If we assume the predicted target word is $W(t^{'})$, and the actual word is $W(t)$, then based on them we can define the loss function. We implement a Dense layer (size of Vocabulary) with a Softmax activation to predict the target $W(t^{'})$. Table 1, present an example of input and target sample by window size 2, which represents two words before and after the target word. The last update of the weight in the Dense layer represents each word as a vector length of $D$.

### 3.3. Noise cancellation model

This module by filtering not related twitters to the target plays a noise cancellation role for topic modeling and trend detection. To enhance the data collection process and topic detection, we create a classification model. In this model, the CBOW used as an Embedding layer. If we assume, the length of each document is $l$ and the dimensionality of each vector assigned by the CBOW model is $D$, then each document can be represented by a matrix in size of $l$*$D$.

Using Cross-domain databases reduce the cost and time for annotating. In this paper, we used transfer learning to enhance the knowledge of unrelated tweets. Cross-domain word representation used in many applications (Abdelwahab & Elmaghraby, 2016; Asghari et al., 2018), in our case study, we trained a COWB on a 20 News Databases (Albishre, Albathan, & Li, 2015) as a source of not related healthcare records. Fig. 4 illustrates the designed architecture of the model for noise cancellation.

In this research, to create the Noise Cancellation Model, we annotated part of the experiment database, and generate a binary classification database, using two labels "related" and "unrelated". Detecting "unrelated" tweets will be challenging when these two categories share the same words. To tackle this issue, we used transfer learning by transferring the knowledge from trained COWB on a database (20 News Database). To perform this transfer learning, we assume a word vector in the COWB module is $V_w$, and the same

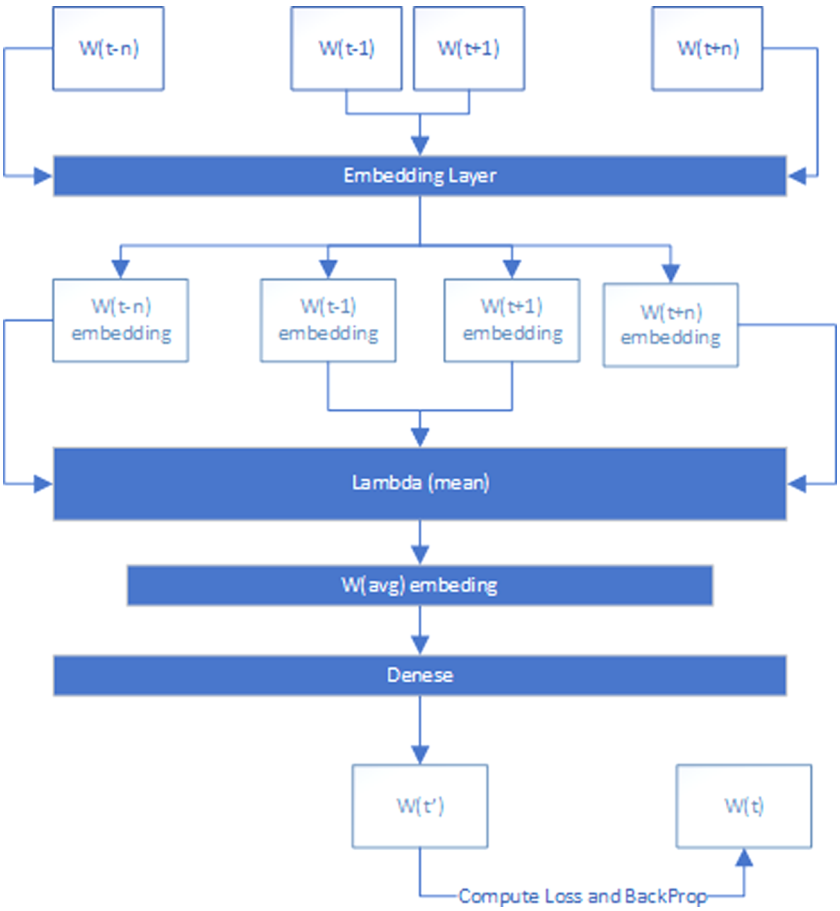


**Fig. 3.** CBOW Deep Learning Layers.

**Table 1**
Sample input and output target in the CBOW model.

| Input | Target |
|---|---|
| 'opposite', 'think', 'cancer', 'replace' | roll |
| 'think', 'roll', 'replace', 'control' | cancer |
| 'roll', 'cancer', 'control', 'month' | replace |
| 'cancer', 'replace', 'month', 'provide' | control |
| 'replace', 'control', 'provide', 'let' | month |
| 'control', 'month', 'let', 'cell' | provide |

word in transferred COWB is $V_w^{'}$. Then we add up two vectors to create a new vector for that word $V_w = V_w + V_w^{'}$. The result will be the input to a dense layer with binary cross-entropy to classify the input as related or unrelated healthcare.

### 3.4. Models selection

To detect a topic, we need to extract the main ideas discussed in a text. These ideas represent themes, subjects for discussion, or conversations. At this stage, we can process the data in different granularities such as the topic of a sentence, a paragraph, or an article. Twitter's messages are like a sentence; Therefore in this research, we analyze the data with sentence granularity.

We select Latent Semantic Analysis (LSA) as a classic algorithm, Latent Dirichlet Allocation (LDA) as the most commonly used technique, LDA-MALLET as an enhanced version of LDA, and Biterm Topic Modeling technique as a specific targeted short message topic modeling, to evaluate the proposed framework for selecting a proper model for topic modeling. In this section, we discussed each of these techniques individually, and then we explain the evaluation metrics.

#### 3.4.1. Latent semantic analysis

The probability of finding a set of words with the same meaning is more likely in a set of documents with the same topic. Latent

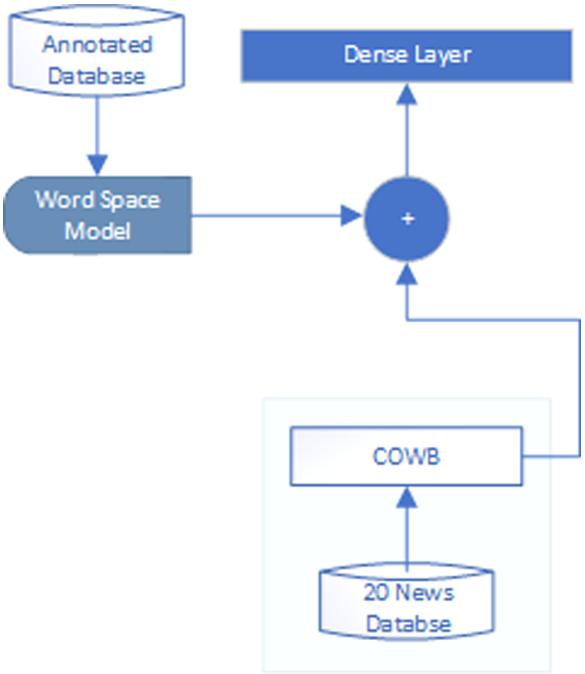


**Fig. 4.** Architecture of model trained for noise cancelation classification.

Semantic Analysis is based on this assumption to find a group of words that can create clusters of documents. Deerwester, Dumais, Furnas, Landauer, and Harshman (1990) proposed a model to find these groups of words between documents. Creating an LSA model requires three steps.

The first step is generating a matrix, that summarizing documents and words. In this matrix, the rows represent unique words and the columns represent the documents and will grow with the number of documents. Dumais (2004) explained in his research, that cumulating frequencies in a sublinear fashion would improve the result. The second step is using Singular-Value Decomposition (SVD) to reduce the dimensionality. Finally, the generated matrix by SVD used to calculate the similarity between entities. Since both the document and terms are in the same space as vectors then we can define a document by some of the terms known as a topic.

### *3.4.2. Latent dirichlet allocation - LDA MALLET*

Latent Dirichlet Allocation (LDA) is a three-level hierarchical Bayesian model (Blei, Ng, & Jordan, 2003). The purpose of LDA is to reduce the dimensionality of data and create a representation of documents by topics where each topic defined by a group of words (Blei et al., 2003).

LDA following generative process for nth word $w_n$ in a document as follow: $z_n \sim Multinomial(\gamma)$ where $z_n$ and $\gamma$ represent the sample a topic and document topic probability matrix respectively, and $w_n \sim Multinomial(\varphi)$ where $w_n$ and $\varphi$ represent Sample a word, topic word probability matrix respectively. The purpose of following this process is to create a sequence of words that represent a topic. Recently a team of researchers at UMASS AMHERST created a toolkit named MALLET (McCallum, 2019). The toolkit is public and contains multiple useful topic modeling techniques, such as the sample-based implementation of LDA.

### *3.4.3. Biterm topic modeling*

This model assesses the problem of the sparsity of data in short texts. Yan, Guo, Lan, and Cheng (2013) proposed the Biterm model based on word connectivity pattern, instead of a document. The Biterm model creates terms, such as “breast cancer” and “digital healthcare” by searching the connectivity between the words over the corpus. In this model, each topic is $z$, word Dirichlet distribution in each topic is $\phi_z$, and topic distribution is $\theta$. Each Biterm contain two words ($w_i$, $w_j$) then they calculate the joint probability of these words as the likelihood of the whole corpus as follow:

$$P(B) = \prod_{(i,j)} \sum_{z} \theta \phi_{i|j} \phi_{j|z} \tag{1}$$

## *3.5. Evaluation metrics*

In this section, we explain two types of evaluation metrics. First, the coverage of selected words as a topic over documents that are specifically defined to evaluate the topic modeling techniques, such as Coherence measurement. Second, involve the expert opinion in topic evaluation. Rosenberg and Hirschberg (2007) used V-measured to evaluate K-means’ performance over document clustering.

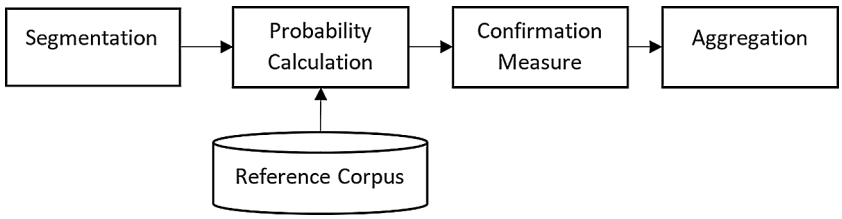


**Fig. 5.** Coherence algorithm steps from segmentation through Aggregation.

They represent that the V-measure can be performed transparently over distinct data-set. Approaching this technique needs manual labeling. In this paper, we used homogeneity, completeness, and v-measure to evaluate topic modeling techniques.

### 3.5.1. Coherence

Measuring Coherence among members of a set, or subset of words is based on the probabilistic analysis. Röder, Both, and Hinneburg (2015) defines a score for describing the quality of document groups generated by a model. A set of words define a topic, and the proposed score is based on a comparison among these sets. We assume the pair for comparison defined by $S = (w', w^*)$ which $w'$,$w^*$ represents two sets of words where $w' \cap w^* = \phi$. Eq. (2) define log-likelihood and Eq. (3) define log-ratio, these are basic elements of coherence calculation.

$$ll = \log \frac{p(w'|w^*) + \epsilon}{p(w'| \neg w^*) + \epsilon} \tag{2}$$

$$lr = \log \frac{p(w'|w^*) + \epsilon}{p(w')^* p(w^*)} \tag{3}$$

All the steps of calculation defined in Fig. 5, the outcome ranges from 0 to 1. However, in our research, we did not rely just on unsupervised metrics, we also used supervised metrics for evaluating generated topics to select the robust model.

### 3.5.2. Homogeneity, completeness and V-measure

Homogeneity explains the number of true-labeled elements in a single class. The goal is that each cluster contains members of the single class. Completeness supplements homogeneity by defining how many of the members of a single class are assigned to the same cluster. These two metrics are defined in a range from 0 to 1, being the latter the optimal performance indicator. The harmonic-mean of these two measurements gives V-measure. Rosenberg and Hirschberg (2007) These metrics determine how close a given cluster is to its ideal definition by examining the conditional entropy of class distribution given the proposed clustering. Annotation of these functions are represented in Table 2.

$$H(C|K) = -\sum_{k=1}^{|K|}\sum_{c=1}^{|C|} \frac{a_{ck}}{N} log \frac{a_{ck}}{\sum_{c=1}^{|C|} a_{ck}} \tag{4}$$

$$H(C) = -\sum_{c=1}^{|C|} \frac{a_{ck}}{N} log \frac{\sum_{k=1}^{|K|} a_{ck}}{n} \tag{5}$$

Based on these formulas, we can calculate homogeneity by following:

$$p_{GL} = \begin{cases} if H(C|K) = 0 & then \quad h = 1 \\ else & h = 1 - \frac{H(C \mid K)}{H(C)} \end{cases} \tag{6}$$

Completeness is symmetrical to homogeneity. So, we can calculate this metric as:

$$p_{GL} = \begin{cases} if H(K|C) = 0 & then \quad c = 1 \\ else & h = 1 - \frac{H(K \mid C)}{H(K)} \end{cases} \tag{7}$$

**Table 2**
Annotations definition in homogeneity and completeness metrics.

| Description | Annotation |
|---|---|
| C as a Set of Classes | $C_{i_{1..n}}$ |
| K as a Set of Cluster | $K_{j_{1..m}}$ |
| Represent a member of class “c” which is element of cluster “k” | $a_{ck}$ |
| Number of Data point | N |

And as we mentioned before V-measure is calculated with the harmonic-mean between homogeneity and completeness, so the calculation is:

$$V - Measure = \frac{(1 + \beta)*h*c}{(\beta*h) + c} \tag{8}$$

We assume that $\beta$ is equal to 1 then the calculation can be done in this way:

$$V - Measure = \frac{2*h*c}{h + c} \tag{9}$$

We proposed $\alpha$ as a dynamic decision making metric to select the proper model among several models for topic detection. This measurement is the multiplication of V-measure and coherence, which is between [0,1]. In clustering, sometimes the probability of overlapping a set of words is high; Although they need to be captured as different topics. This situation happens in subtopics, for example, women as a big topic and sub-topic such as diabetic in women or pregnancy. The v-measure as a supervised evaluation can contribute to the metric to sharpener the decision boundaries. Therefore, Coherence beside V-measure creates a combination to handle the dynamic of twitter in topics detection. In Eq. (10) *cr* refers to coherence, and *h* refers to homogeneity, also *c* represents the completeness.

$$\alpha = cr*\frac{2*h*c}{h + c} \tag{10}$$

### 3.6. Visualization

After calculations and decision making to create topics, we needed to find a way to represent and visualize the trend over the tweets to make the framework be able to represent trends track over location and time. The trend detection process is represented in a visual format to provide a better perception of the topic trend over time and across locations. In our experiment, we include two factors to visualize. Topic changes in each state, and the nationwide change over time. We used a color-coding structure, which will represent these factors for a given topic. RGB channels are utilized for color-coding and data structured as presented in Fig. 6. Where each row represents the location change over time. In addition, the RGB color mapping is represented by three channels per row.

Processing the data and running the topic modeling detect trends. These topics will change based on location and time-frequency. We assume $N$ as the total number of generated topics and $n$ as the individual topic. Location is $l$, and time is $t$, then The topic frequency in a specific location and time is $(n)_l^t$. To calculate the influence of a state we used Eq. (11). In our color-coding the Red channel will be dedicated to the State Influence $SI\,(n)_l^t$. Detecting the red gradient on the map infer that a state has more influence on that topic.

$$SI\,(n)_l^t = \frac{f\,(n)_t^l}{\sum_{topic=1}^{N} f\,(topics)_t^l} \tag{11}$$

The next level of representation will be the influence of a trend over a country. To calculate this, we need to assess how many tweets posted about that topic at a specific time. If we assume the total number of locations at $L$. $f\,(l)_n^t$ represents the frequency of tweets in a specific location ($l$) followed by a topic ($n$) and time ($t$). Eq. (12) presents the country influence calculation.

$$CI\,(n)_t^n = \frac{f\,(n)_t^n}{\sum_{location=1}^{N} f\,(location)_t^n} \tag{12}$$

The green channel represents the States that are trendsetters for the whole country. Similarly, the yellow color means the topic not only effects the State but also effects the country however they are not a trendsetter.

## 4. Experimental results and analysis

The stream collection process ran for 9 months, from October 2018 until May 2019, resulting in 765,715 collected tweets. Table 3 reports the tweets distribution for two of the higher and lower states over the observation period. Florida was the state with the

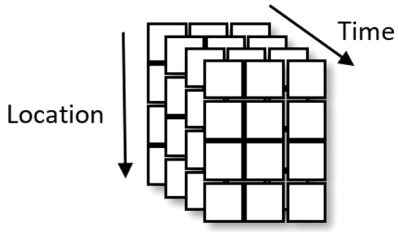


**Fig. 6.** Tensor of data representation.

**Table 3**
Most representative states in terms of the total number of tweets over the observation period.

| State | 2018/8 | 2018/9 | 2018/10 | 2018/11 | 2018/12 | 2019/1 | 2019/2 | 2019/3 | 2019/4 | Total |
|---|---|---|---|---|---|---|---|---|---|---|
| **Florida** | 1240 | 8864 | 997 | 7946 | 4838 | 7697 | 1298 | 24,741 | 15,228 | 92,849 |
| **California** | 1752 | 9990 | 1332 | 7570 | 4496 | 7568 | 8338 | 4106 | 3664 | 8816 |
| **South Dakota** | 4 | 52 | 4 | 106 | 57 | 126 | 256 | 301 | 122 | 1028 |
| **Wyoming** | 19 | 82 | 8 | 44 | 25 | 73 | 127 | 135 | 116 | 629 |

largest number of tweets and Wyoming had the lowest number of tweets.

After collecting the data on a server, the framework cleaning Module will clean and filter the twitter message automatically. Then we Annotated a portion of the database (1,275 messages). A team of graduate students annotates the twitters, and we use them as ground truth for evaluation metrics. Labeled data was required to calculate the Homogeneity, Completeness, and V-measure. The statistics of the eight most common topics in the Annotated database represented in Table 4.

### *4.1. Noise cancellation*

We applied the CBOW deep learning model with a window size of 2, and a dimension of 100, leading to a matrix size of 169, 394*100, representing each word in a vocabulary of 169,394 words by vector length of 100. Then we apply the Noise cancellation model to filter not related messages. To evaluate the noise cancellation model, we used the annotated database (ADODB) to classify “related” and ”not related” tweets, then we compare regular training on News20 (Albishre et al., 2015) dataset with proposed transfer learning. This dataset contains a wide range of topics and among all the topics there is a group related to healthcare named “sci.med”, which contains 990 documents. This topic removed before we train CBOW on News20 Dataset.

Using the transfer knowledge approach to select not related healthcare database, shows better result compare to the regular one. Classification results show that almost 10% improvement on F1-Score achieved by transfer knowledge from other resources. The complete results of this classification based on F1-score, Precision and Recall reported in Table 5. We report the result of running the models over a 20% balanced test dataset.

### *4.2. Topic modeling selection*

The next step is selecting the proper model to assign topics to the messages. To experiment with the model selector module we repeat the experiment by a different number of topics $t$ (in the range of 2 to 30). We report Coherence in Fig. 7 for each model based on a different t. In our database, 30 topics using LDA-MALLET is the best. The coherence select 30 as the optimal number of topics. However, the annotated dataset contains 28 individual topics. This implies that evaluating and making decisions based on coherence might not be accurate enough and we need more evaluation and analysis over selecting the practical number of topics.

Our framework is using Homogeneity, Completeness and V-measure to support the decision. These metrics require a ground truth to evaluate the selected models. We used our Annotated database to evaluate the results. Based on these metrics we observer two groups of models. The first is LDA and BITERM, which they are show approximately same behavior. The second is LDA-MALLET and LSA. We show the result of Homogeneity, Completeness and V-Measure metrics in Figs. 8, 9 and 10.

This experiment not only will help us to choose number of topics, but also will represent which model performed more efficient. The Alpha Value summarize 4 metrics in one to let the model to have a metric to rank the models.

In Fig. 11 it can be concluded that the LDA-MALLET shows better results compare to other models also we can see that 28 is the best number of topics. This metric help us to understand the behavior of different models. Having this robust measurement will improve the quality and reliability of reports generated by this framework.

Although, part of this experiment needs experts interfere, which is the part for annotating database. Experts could contribute by evaluating labeled messages partially. This can decrease the cost of manually annotated and help the system to do the self-correction in model selector by using expert feedback beside get advantages from unsupervised properties.

**Table 4**
Eight Top common Generated Topics in 1275 tweets that annotated by graduate students.

| Topic | Count | Percent |
|---|---|---|
| notrelated | 251 | 20% |
| innovation, health tech, health it, tech, iot, medtech, mhealth, cybersecurity, bigdata, machine learning | 126 | 10% |
| cancer, bcsm, week, day, share, love, risk, childhood cancer, read | 76 | 6% |
| industry, medicine, future, check, digital, pharma, learn, blog, big, global | 71 | 6% |
| Digital health, data, technology, health tech, innovation, blockchain, tech, industry, digital, iot | 65 | 5% |
| pregnancy, pharma, migraine, women, pregnant, love, baby, biotech, pain, fda | 63 | 5% |
| diabetes, news, research, education, free, women, pregnancy, information, study, food | 61 | 5% |
| patients, hospital, access, services, providers, quality, hospitals, home, telehealth, doctor | 56 | 4% |

**Table 5**
Regression classification results with and without transfer learning.

| Class Label | Precision | Recall | F1-score |
|---|---|---|---|
| **Result the Classification with Transfer Learning** | | | |
| Not Related | 0.82 | 0.86 | 0.84 |
| Related | 0.82 | 0.86 | 0.84 |
| **Result the Classification without Transfer Learning** | | | |
| Not Related | 0.68 | 0.76 | 0.72 |
| Related | 0.81 | 0.74 | 0.77 |

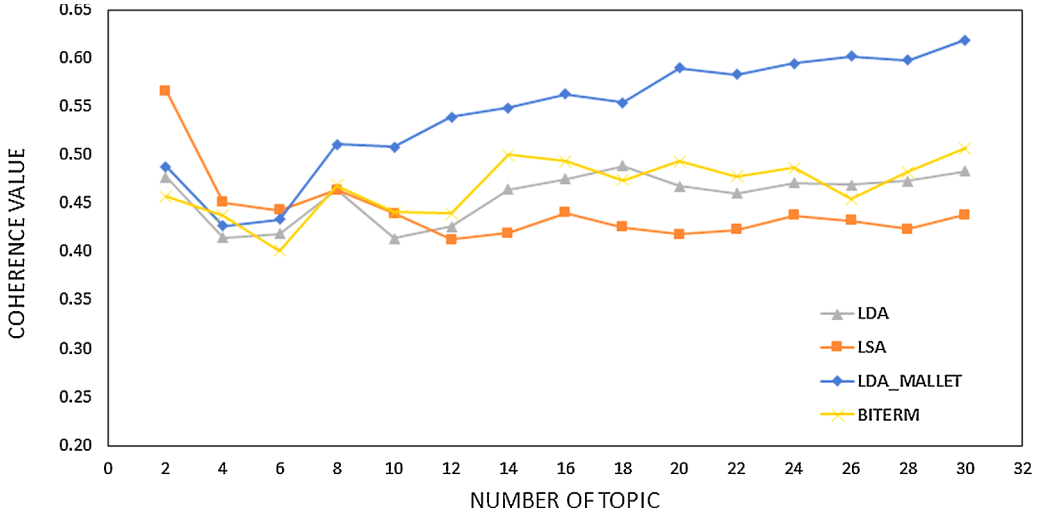


**Fig. 7.** Coherence measurement for LDA, LSA, LDA-MALLET, BTM.

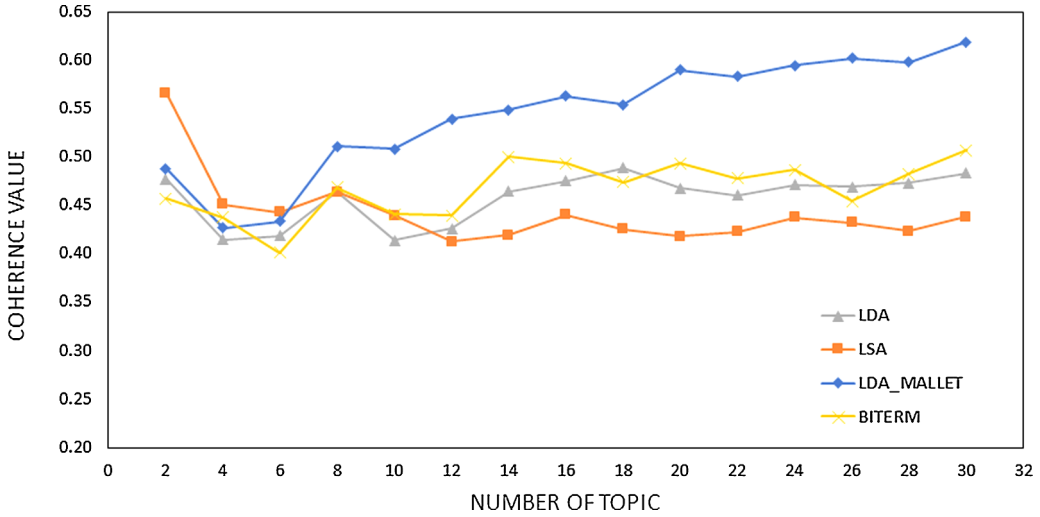


**Fig. 8.** Homogeneity over annotated database based on the model created over Vectorized Twitter Database.

## 5. Discussion and case study

We selected a trending topic related to diabetics that appeared while conducting this research. Media outlets such as US News discussed about the insulin crisis and there were general interest and research studies related to this concern as the one presented in Fralick and Kesselheim (2019). Therefore, we found the ”Cost of Insulin” as a demonstrative case study. To visualize the results we used the color to highlight the intensity of the trending topic over space and time distributed on the U.S map. To serve as example, Fig. 12 shows the topic interest change as time-lapse with monthly granularity.

We observe low saturation color (both red and green channel) over Alaska, this implies that this topic was not of concern for this

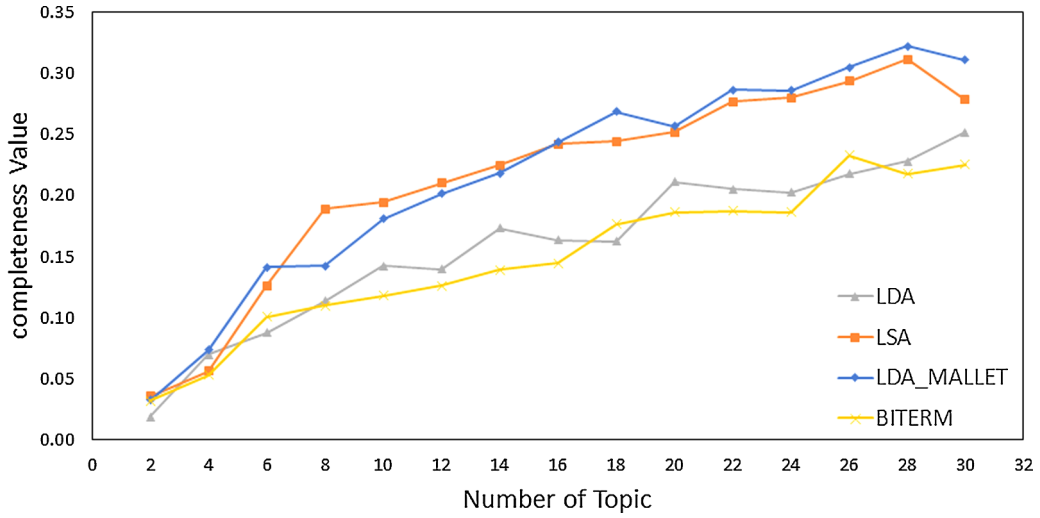


**Fig. 9.** Completeness over annotated database based on the model created over Vectorized Twitter Database.

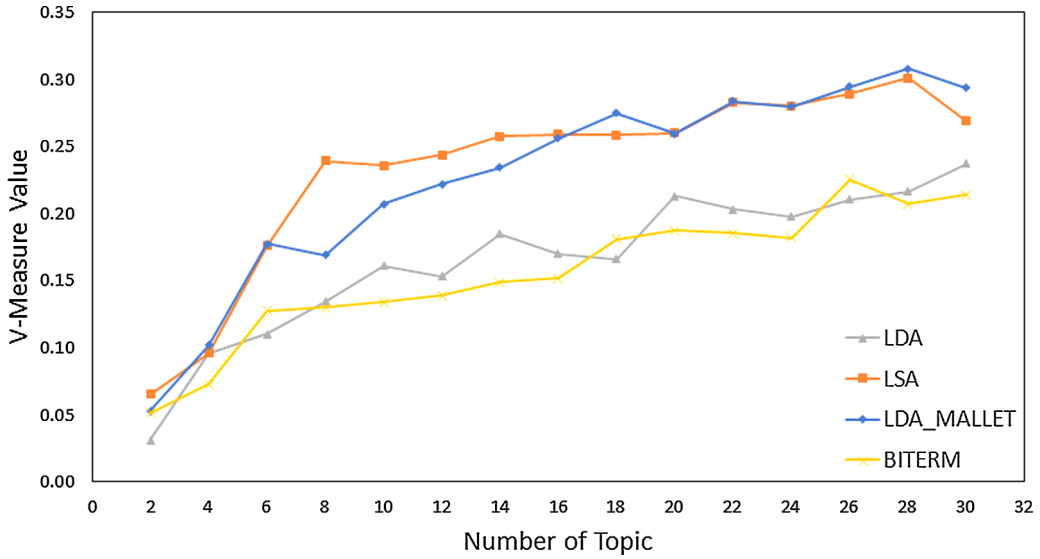


**Fig. 10.** V-measure over annotated database based on the model created over Vectorized Twitter Database.

state at the beginning. As we move through time, we can observe that color migrates to red, meaning the topic interest grew. Florida and California from the beginning had green color, meaning they were trendsetters among all the states. We can also observe, Washington and Texas are colored yellow, meaning their influence is at the national level, but they were not trendsetters. If we put these frames in an animation, we can observe a movement of the topic from October 2018 to April 2019 for seven months. This movement started from Florida and California and spread up through the country until it reached Alaska. New York, Washington DC, and Chicago were already interested in this topic.

## 6. Conclusion

Topic detection has many applications, it is useful in grouping scientific papers, understanding customers' concerns about a product, and for detecting and tracking trends in social media. These can be helpful for many institutions, such as governments or news agencies. In this paper, we proposed a general framework to perform a fully automatic text collection and cleaning, alongside a semi-automatic topic detection technique by using hybrid evaluation metrics that assess the result regardless of selected models. We address two main issues, first automatic noise detection by the usage of deep transfer learning techniques, second, we address the high cost of manual labeling and annotation. We decrease the cost of labeling by creating an annotated database using entropy-based clustering measurements.

In this experiment, we focus on health-related tweets and represent the results to demonstrate the utility of our proposed

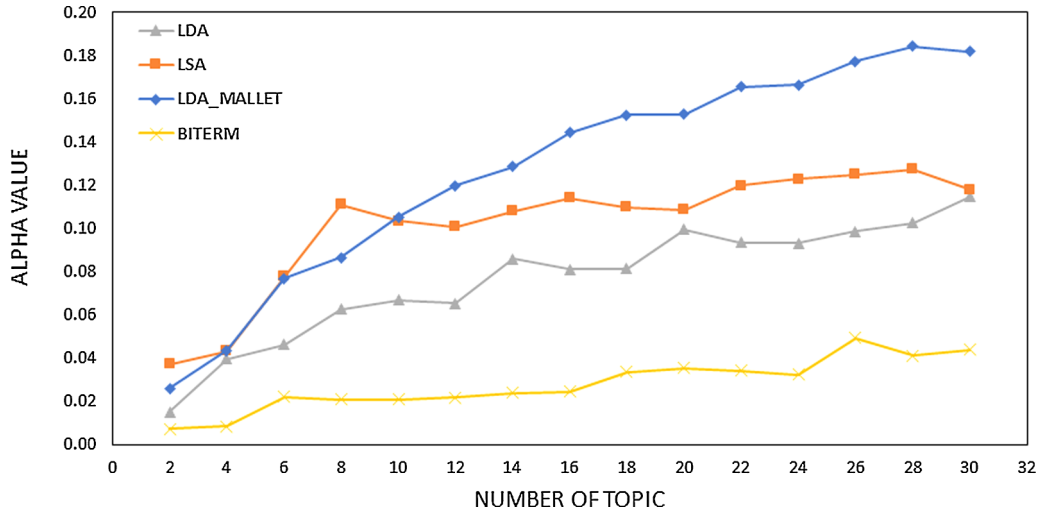


**Fig. 11.** Calculated Alpha over annotated database based on the model created over Vectorized Twitter Database.

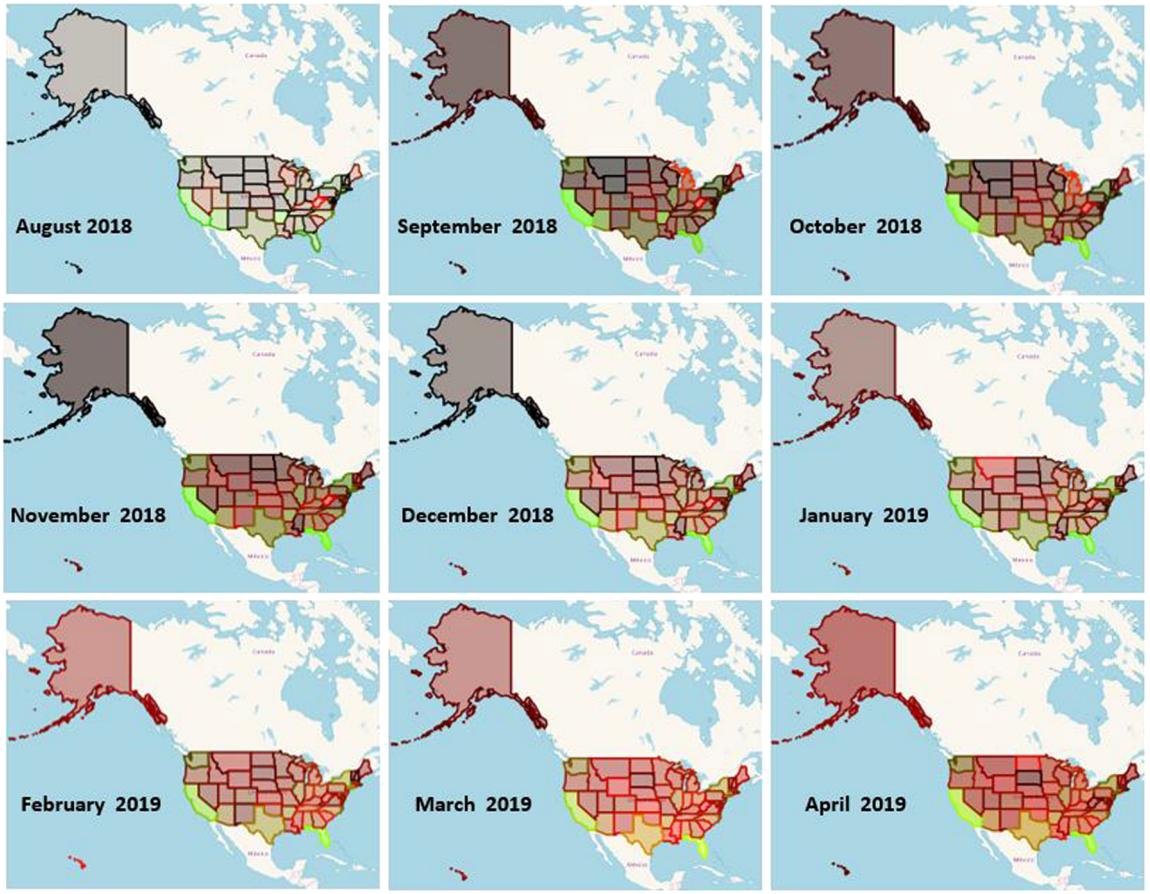


**Fig. 12.** Tracking "Cost of Insulin" Topic from August 2018 until April 2019.

framework. We measure our results by using V-Measure, Homogeneity, Completeness, and Coherence to select the number of topics and models. Besides, we visualize a sample of trend tracing in the United States, which shows the value of the analysis using this framework. This hybrid approach of using feedback from experts with our results improves performance in dynamic environments. This continuing learning approach used to address the domain shift problem in time and location as in the case of COVID-19 pandemic analysis, as follow up on this research.

## CRediT authorship contribution statement

**Mohsen Asghari:** Conceptualization, Data curation, Formal analysis, Writing - original draft, Writing - review & editing. **Daniel Sierra-Sosa:** Conceptualization, Data curation, Formal analysis, Writing - original draft, Writing - review & editing. **Adel S. Elmaghraby:** Conceptualization, Data curation, Formal analysis, Writing - original draft, Writing - review & editing.

## Supplementary material

Supplementary material associated with this article can be found, in the online version, at 10.1016/j.ipm.2020.102340

## References


Abdelwahab, O., & Elmaghraby, A. (2016). Uofl at semeval-2016 task 4: Multi domain word2vec for twitter sentiment classification. *The 10th international workshop on semantic evaluation (SemEval-2016),* 164–170.

Albishre, K., Albathan, M., & Li, Y. (2015). Effective 20 newsgroups dataset cleaning. *The 2015 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), 3*, 98–101.

Alessa, A., & Faezipour, M. (2019). Preliminary flu outbreak prediction using twitter posts classification and linear regression with historical centers for disease control and prevention reports: Prediction framework study. *JMIR Public Health and Surveillance, 5*(2), e12383.

Asghari, M., Sierra-Sosa, D., & Elmaghraby, A. (2018). Trends on health in social media: Analysis using twitter topic modeling. *2018 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT),* 558–563.

Beeferman, D. (2019). The datamuse api.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research, 3*(1), 993–1022.

Bodenreider, O. (2004). The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research, 32*(suppl_1), D267–D270.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American society for information science, 41*(6), 391–407.

Dumais, S. T. (2004). Latent semantic analysis. *Annual review of information science and technology, 38*(1), 188–230.

Fralick, M., & Kesselheim, A. S. (2019). The us insulin crisis-rationing a lifesaving medication discovered in the 1920s. *The New England journal of medicine, 381*(19), 1793.

Gesualdo, F., Stilo, G., Gonfiantini, M. V., Pandolfi, E., Velardi, P., Tozzi, A. E., et al. (2013). Influenza-like illness surveillance on twitter through automated learning of naïve language. *PLoS One, 8*(12), e82489.

Griffis, H. M., Kilaru, A. S., Werner, R. M., Asch, D. A., Hershey, J. C., Hill, S., ... Merchant, R. M. (2014). Use of social media across us hospitals: descriptive analysis of adoption and utilization. *Journal of medical Internet research, 16*(11), e264.

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *The National academy of Sciences, 101*(suppl 1), 5228–5235.

Guo, L., Vargo, C. J., Pan, Z., Ding, W., & Ishwar, P. (2016). Big social data analytics in journalism and mass communication: Comparing dictionary-based text analysis and unsupervised topic modeling. *Journalism & Mass Communication Quarterly, 93*(2), 332–359.

Hawkins, J. B., Brownstein, J. S., Tuli, G., Runels, T., Broecker, K., Nsoesie, E. O., ... Bourgeois, F. T., et al. (2016). Measuring patient-perceived quality of care in us hospitals using twitter. *BMJ Qual Saf, 25*(6), 404–413.

Hwang, M.-H., Wang, S., Cao, G., Padmanabhan, A., & Zhang, Z. (2013). Spatiotemporal transformation of social media geostreams: a case study of twitter for flu risk analysis. *The 4th ACM SIGSPATIAL International Workshop on GeoStreaming,* 12–21.

Katal, A., Wazid, M., & Goudar, R. (2013). Big data: issues, challenges, tools and good practices. *The 2013 Sixth international conference on contemporary computing (IC3),* 404–409.

Koylu, C. (2019). Modeling and visualizing semantic and spatio-temporal evolution of topics in interpersonal communication on twitter. *International Journal of Geographical Information Science, 33*(4), 805–832.

Krestel, R., Fankhauser, P., & Nejdl, W. (2009). Latent dirichlet allocation for tag recommendation. *The third ACM conference on Recommender systems,* 61–68.

Laylavi, F., Rajabifard, A., & Kalantari, M. (2017). Event relatedness assessment of twitter messages for emergency response. *Information processing & management, 53*(1), 266–280.

Martinez-Millana, A., Fernandez-Llatas, C., Bilbao, I. B., Salcedo, M. T., & Salcedo, V. T. (2017). Evaluating the social media performance of hospitals in spain: A longitudinal and comparative study. *Journal of medical internet research, 19*(5), e181.

McCallum (2019). Mallet: A machine learning for language toolkit.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Rathore, M. M., Ahmad, A., Paul, A., Hong, W.-H., & Seo, H. (2017). Advanced computing model for geosocial media using big data analytics. *Multimedia Tools and Applications, 76*(23), 24767–24787.

Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the space of topic coherence measures. *The eighth ACM international conference on Web search and data mining,* 399–408.

Rosenberg, A., & Hirschberg, J. (2007). V-measure: A conditional entropy-based external cluster evaluation measure. *The 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL),* 410–420.

Sierra-Sosa, D., Asghari, M., Gordon, J., & Elmaghraby, A. S. (2019). Demographic influence on opioid misuse. *2019 Fifth International Conference on Advances in Biomedical Engineering (ICABME),* 1–4.

Thom, D., Bosch, H., Koch, S., Wörner, M., & Ertl, T. (2012). Spatiotemporal anomaly detection through visual analysis of geolocated twitter messages. *The 2012 IEEE Pacific Visualization Symposium,* 41–48.

Tursunbayeva, A., Franco, M., & Pagliari, C. (2017). Use of social media for e-government in the public health sector: A systematic review of published studies. *Government Information Quarterly, 34*(2), 270–282.

Tweepy (2019). The tweepy api.

Yan, X., Guo, J., Lan, Y., & Cheng, X. (2013). A biterm topic model for short texts. *The 22nd international conference on World Wide Web,* 1445–1456.

Zadeh, A. H., Zolbanin, H. M., Sharda, R., & Delen, D. (2019). Social media for nowcasting flu activity: Spatio-temporal big data analysis. *Information Systems Frontiers,* 1–18.

Zahra, K., Imran, M., & Ostermann, F. O. (2020). Automatic identification of eyewitness messages on twitter during disasters. *Information processing & management, 57*(1), 102107.

Zhao, L., Chen, F., Lu, C.-T., & Ramakrishnan, N. (2015). Spatiotemporal event forecasting in social media. *the 2015 SIAM international conference on data mining,* 963–971.

Serban, O., Thapen, N., Maginnis, B., Hankin, C., & Foot, V. (2019). Real-time processing of social media with sentinel: a syndromic surveillance system incorporating deep learning for health classification. *Information Processing & Management, 56*(3), 1166–1184.