ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/321453109

Information Diffusion on Twitter: Pattern Recognition and Prediction of Volume, Sentiment, and Influence

Conference Paper · December 2017

DOI: 10.1145/3148055.3148078

CITATIONS
2

READS
211

3 authors:

Amartya Hatua
University of Southern Mississippi
10 PUBLICATIONS **10** CITATIONS
SEE PROFILE

Trung Nguyen
University of Southern Mississippi
12 PUBLICATIONS **13** CITATIONS
SEE PROFILE

Andrew H. Sung
University of Southern Mississippi
17 PUBLICATIONS **24** CITATIONS
SEE PROFILE

Some of the authors of this publication are also working on these related projects:

Deepfake Video Detection and Prevention View project

Building a Learning Machine Classifier with Inadequate Data for Crime Prediction View project

All content following this page was uploaded by Trung Nguyen on 31 October 2018.

The user has requested enhancement of the downloaded file.

Information Diffusion on Twitter: Pattern Recognition and Prediction of Volume, Sentiment, and Influence

Amartya Hatua
School of Computing
The University of Southern Mississippi
Hattiesburg, MS 39406, USA
amartya.hatua@usm.edu

Trung T. Nguyen
School of Computing
The University of Southern Mississippi
Hattiesburg, MS 39406, USA
trung.nguyen@usm.edu

Andrew H. Sung
School of Computing
The University of Southern Mississippi
Hattiesburg, MS 39406, USA
andrew.sung@usm.edu

ABSTRACT

Characterizing, predicting, and quantifying the impact of postings, tweets, messages, etc. on social media platforms is a topic of growing interest due to the increasing reliance on using social media as a means for various purposes by individuals and organizations alike. In this paper, we describe an information diffusion model on the social network of Twitter. The model treats information diffusion on social media as a multivariate time series problem and deals mainly with three different dimensions of Twitter data and the different patterns of information diffusion. These dimensions are the volume of tweets, the sentiment of tweets and influence of tweets. To discover different patterns of information diffusion on Twitter, time series clustering is used where Dynamic Time Warping distance is adopted as the distance measure. To predict different parameters of each of the three dimensions, the linear time series model of Autoregressive Integrated Moving Average (ARIMA) and the non-linear time series model of Long Short-Term Memory (LSTM) Recurrent Neural Networks are used and their performance is compared. Results indicate that LSTM models achieve far better performance and hold great potential to be utilized for real-world applications.

CCS CONCEPTS

• **Human-centered computing** → **Social network analysis**; *Social networks*; • **Information systems** → *Social networks*;

KEYWORDS

Twitter; information diffusion; time series clustering; DTW; time series forecasting; ARIMA; RNN; LSTM

1 INTRODUCTION

Since the inception of online social media, its usage has evolved in many facets, and nowadays social networks has become one of the most important means of communication-it is used as a major tool for viral marketing, political messaging, opinion formation and many other things. The abundance of information in the social network and its huge impacts have made the social networks a fast growing industry in recent years. People who share a common interest get connected over the social network; they share and spread information over a period of time, or, in other words, the information diffuses over the social network. The pattern and behavior of information diffusion in the social network, therefore, can be analyzed to help predict the success, failure, popularity and opinion about different events.

Information diffusion model in the social network has been studied using the graph-based method [10] or considered as an analogous model for the viral spread of diseases [3]. Earlier researches on this topic showed that the popularity of a topic or meme depends on the number of friends or acquaintances of users who have responded to that topic or meme [4]. However, information diffusion shows much more intelligent and complex behaviors which are beyond just the phenomenon of exposure of meme and it depends on many other factors. Some of the most significant factors are the sentiment of the meme, topic of discussion, the network structure of the user, the physical location of user and presence of some influential users on a topic. In [8] Ferrara shows the complex dynamics between the sentiment of tweets and information diffusion; the papers initial parts discuss the effect of sentiment on the diffusion speed and on content popularity while latter parts of the paper addresses the relation between different type sentiments and their temporal evolution; the paper, however, focuses only on the sentiment of the text of the tweets but does not consider another important aspect of the text, i.e. the topic of it. In [21] Pinto et al. introduce a framework to model information diffusion in social networks based on linear multivariate Hawkes processes and the latent Dirichlet allocation topic model, which mainly focused on the relation between information diffusion and the topics of discussion in social networks. Other than the contents of the social network, the network structure also plays a vital role in information diffusion on the social network. In [27] Yang et al. propose an information diffusion model in implicit networks using Linear Influence Model, where the authors discuss the roles of the different participants of the social network in the dynamics of diffusion. As can be easily understood a very popular person can influence many of their acquaintances over a topic.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
BDCAT'17, December 5–8, 2017, Austin, TX, USA
© 2017 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-5549-0/17/12.
https://doi.org/10.1145/3148055.3148078

Other than factors like this, network structure is another important factor for information diffusion. In [23] Reagans et al. discuss how network structure plays an important role in information diffusion; the authors conclude that social cohesion and network range are more important than the strength of the tie between two people for effective knowledge transfer, social capital, and information diffusion. Recently, Kafeza et al [15] conducted a research of predicting information diffusion patterns in Twitter. They collected tweets data for a couple of hours that related to a single hashtag. Based on retweet action, they analyzed different Tree-Shaped Tweet Cascades patterns and then reduced to four most popular tree structures. Later, they built an information diffusion pattern prediction model based on linguistic features of the tweet, user profile features and their corresponding tree-shaped tweet pattern labels. However, in their work, time components have not been considered and the size of dataset is also small and not diverse enough. In our current research, we extend their work by using a larger dataset, analyzing the data in time series model, and labelling them by using Dynamic Time Wrapping (DTW) clustering. In most of the study on information diffusion, the researchers focused on the pattern of information diffusion and the factors that affect it. Although there are many research papers on different factors of information diffusion, there is very little research that has been carried out to predict the future behavior of these factors. To predict and forecast different factors related to information diffusion, we propose a prediction model in this paper.

We aim in this work to build a prediction model which can predict the volume of tweets and reachability [6] for a hashtag, and which can also predict the sentiment and number of people who will be influenced by that hashtag over a period of time. The proposed model predicts three different facets of information diffusion, they are i) volume ii) sentiment and iii) influence of different popular memes of a social network. We have chosen Twitter as our target social media platform because Twitter provides publicly available data. Specifically the objectives of this research are as follows: 1) Understanding the pattern of information diffusion related to a hashtag and its relationship with the volume of tweets and number of people who are using that hashtag. After that, predicting the number of tweets and people who will use the same hashtag over a period of time. 2) Finding the relation between the sentiment of a tweet and its effect on information diffusion, and predicting the sentiment of tweets related to a hashtag. 3) Finding the direct and indirectly affected users by a hashtag, and predicting the number of total affected users by a particular hashtag over a period of time.

2 METHODOLOGY

This section describes our methodology. Fig. 1 describes our general approach to analyze, recognize different information diffusion patterns on Twitter and then build prediction models for every hashtag on Twitter social network platform.

2.1 Modeling the Information Diffusion Process

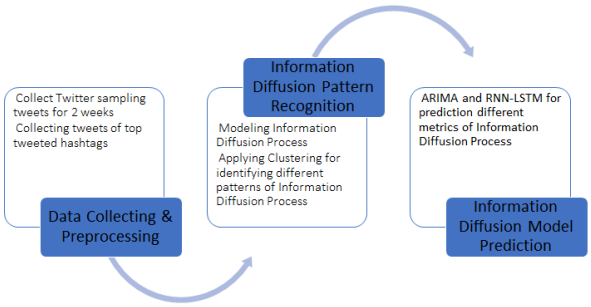

Figure 1: The architecture overview

Traditional methods to model information diffusion on online social media require an explicit knowledge of the network. Nevertheless, there are many parameters which are difficult to track such as recommendations, links, tags, topics, phrases or memes. Therefore, in this research, we model the information diffusion process on social network as a multivariate time series problem. This approach can help address the above issue as no explicit knowledge of the network is required. The information diffusion process of a topic (which is a hashtag on Twitter) can be modeled in three-time series dimensions with a total of 10 features, as described below:

2.1.1 Volume: This dimension contains two features:
#tweet: the total number of tweets of a corresponding hashtag.
#retweet: the total number of retweets of a corresponding hashtag.

2.1.2 Influence: This dimension contains two features:
#direct_influence_user: the total number of users and mentioned users that associated with all tweets contain such hashtag.
#indirect_influence_user: the total number of followers of all users and mentioned users that associated with all tweets contain such hashtag.

2.1.3 Sentiment: This dimension contains six features:
#positive_percentage: the percentage of positive sentiment among the tweets.
#neutral_percentage: the percentage of neutral sentiment among the tweets.
#negative_percentage: the percentage of negative sentiment tweets.
#positive_average_score: the average score of positive sentiment tweets.
#neutral_average_score: the average score of neutral sentiment tweetsAn Approach for Pattern Recognition and Prediction of Information Diffusion Model on Twitter.
#negative_average_score: the average score of negative sentiment tweets.

2.2 Recognizing Information Diffusion Patterns

We believe that a good understanding of the different patterns in information diffusion processes will help to improve the prediction models. So, after modeling the information diffusion processes as a multivariate time series problem using the 10 features, time series clustering techniques will be employed to analyze and recognize such patterns. In this study, clustering techniques such as TADPole clustering, Hierarchical clustering, partitional clustering with Dynamic Time Warping (DTW) distance are used to determine the number of different information diffusion patterns and their shapes.

2.3 Predicting Information Diffusion Processes

Based on the time series clustering step, we obtain different groups of clusters corresponding to the 10 features in our model. To determine the information diffusion patterns of new Twitter hashtags, k-NN can be used. With enough initial time steps data of a new Twitter hashtag, we can use k-NN with DTW distance to determine which clusters this hashtag belongs to, which in turn will help to recognize the information diffusion patterns of the new hashtag.

The next step is building time series forecasting models for each cluster. Every cluster will have its different shape and characteristics. To demonstrate the novelty of our approach, we compare the performance of such models for each cluster with the prediction models without using the clustering results. Experimental results show that prediction model using LSTM with patterns information delivers better performance than traditional time series forecasting methods (ARIMA) and without knowing the patterns beforehand.

3 DATA COLLECTING AND PREPROCESSING

3.1 Twitter Data Collecting

One of the objectives of this research is analyzing the pattern of information diffusion on Twitter. Hashtags have a very important role in the information diffusion process on Twitter. Twitter users usually tag posts with hashtags. The adoption of hashtags has created a global information transmission effect on Twitter because hashtags help users keep track of information topics and therefore can form dynamic communities or groups. Therefore, we collected 2-weeks of Twitter sampling streaming data using Tweepy Python library from July 1 to July 14, 2017. From this collected dataset , we collected all hashtags that were used in tweets in this time with their corresponding total number of tweets. Then we sorted these hashtags into descending order based on the volume of tweets. We identified that about 1 million different hashtags were used in this time, and discarded hashtags that did not have a significant volume of tweets (we applied a threshold of 200). As a result, we obtained a list of top 1,686 hashtags with the highest volume of tweets. According to Twitter,the streaming API will only return 1% of real-time tweets data at a time, so we think that we may not have enough tweets of those 1,686 hashtags to analyze. Therefore, we began to collect all tweets that related to those 1,686 hashtags using Twitter Search API in the next 3 weeks, from July 15 to August 4, 2017. Finally, we collected about 27.5 million tweets that contain those 1,686 hashtags that we want to analyze.

3.2 Data Preprocessing

From the 27.5 million tweets, we composed our multivariate Twitter information diffusion time series dataset based on the model described in section 2 above. Our time series is measured in hourly basis. Accordingly, for each hashtag, we have 10 time series data rows:

#tweet: the total number of tweets that contain such hashtag in each hour

#retweet: the total number of retweets that contain such hashtag in each hour

#direct_influence_user: the total number of users and mentioned users that associated with all tweets contain such hashtag in each hour

#indirect_influence_user: the total number of followers of all users and mentioned users that associated with all tweets contain such hashtag in each hour

#positive_percentage: the percentage of positive sentiment tweets in each hour

#neutral_percentage: the percentage of neutral sentiment tweets in each hour

#negative_percentage: the percentage of negative sentiment tweets in each hour

#positive_average_score: the average score of positive sentiment tweets in each hour

#neutral_average_score: the average score of neutral sentiment tweets in each hour

#negative_average_score: the average score of negative sentiment tweets in each hour.

To understand the information diffusion patterns of tweets sentiments, two measurements, the average sentiment score and the sentiment percentage, were used. Sentiment scores (positive, negative and neutral scores) for each tweet are determined by using the Python NLTK library with Wordnet corpora [22]. Later on, the average positive, negative and neutral sentiment scores of all tweets for every time steps (hourly basis) are calculated. Similarly, the percentage of positive, negative and neutral tweets for every hourly time steps are also calculated.

The preprocessed data is available in [7].

4 INFORMATION DIFFUSION PATTERNS

4.1 Motivation

Data collection and data preprocessing are the initial major tasks of this research work. After these two steps, we need to analyze the data and find the different patterns in

our information diffusion time series dataset. There are several patterns that exist in the data and the patterns are also unknown, or in other words there is no label or class name corresponding with each hashtag. To divide hashtags into groups of similar patterns, many time series clustering techniques are employed. As our information diffusion contains 10 features, clustering for each of those features is done separately.

4.2 Time Series Distance Measure

In the present scenario, our dataset contains multiple sequences that were taken at successive, equally spaced points in time which is similar to other time series data like stock market data or weather data. To measure the similarity between two temporal sequences which may vary in speed, the Dynamic Time Warping (DTW) distance is one of the most popular measures. Hence DTW is used in conjunction with time series clustering techniques in our experiments.

4.3 Dynamic Time Wrapping (DTW)

The distance is the most useful parameter in time series analysis which helps to measure the dissimilarity between two time series data. DTW employs a dynamic programming algorithm to find the distance between two time series; though an effective distance measure, it requires a lot of computation. In Fig. 2 alignment between two sample time-series are shown. To find DTW initial and final points of the series must match. The deatils of DTW calculation can be find in [9].

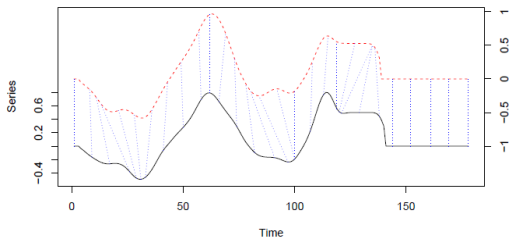

Figure 2: Sample alignment performed by the DTW algorithm between two series

The dashed blue lines exemplify how some points are mapped to each other, which shows how they can be warped in time. Note that the vertical position of each series was artificially altered for visualization. The detailed algorithm to calulate DTW distance can be found in [9].

DTW distance is usually used jointly with clustering algorithms on time series data. Some of the wellknown clustering techniques are Hierarchical clustering, Partitional clustering, and TADPole clustering. Brief descriptions of those clustering algorithms are described in the following subsections.

4.4 Hierarchical Clustering

This clustering algorithm tries to create a hierarchy of groups in which, as the level in the hierarchy increases, clusters are created by merging the clusters from the next lower level, such that an ordered sequence of groupings is obtained [9]. The created hierarchy can be visualized as a binary tree where the height of each node is proportional to the value of the intergroup dissimilarity between its two daughter nodes.

4.5 Partitional Clustering

Partitional clustering follows a stochastic procedure and it starts with a fixed number of random points from its dataset. The number of data points is decided by the required number of clusters. Some of the most popular algorithms of this type are k-means [20] and k-medoids [13]. In the first step of this algorithm a fixed number of data point is randomly selected (say k points) and assigned as centroids. In subsequent steps, one by one, all the remaining data points are clustered based on similarity to the centroids, and after each iteration new centroids are calculated.

4.6 TADPole Clustering

It is a relatively new method for time series data clustering with DTW distance. In this algorithm, the centroid of the clusters is always the element of dataset, so it can be also considered as PAM clustering. Depending on cutoff value of distance the clustering algorithm is deterministic in nature. To find close neighbors in DTW space, the algorithm initially uses the upper and lower bounds of the DTW distance. To do a faster calculation of clustering, the algorithm tries to prune as many DTW calculations as possible.

4.7 Cluster Evaluation

Clustering is an unsupervised procedure, so performance evaluation of clustering maybe somewhat subjective at least. Much research has been done to develop a cluster evaluation metrics by cluster validity indices (CVIs), and there are many indices proposed by different researchers. In this paper, we employ with some of the very popular [1] indices among them.

Every index has its own range of values. For some indices, if the value is high then the quality of cluster is better; on the other hand some indices show exactly the opposite characteristics. Some indices do not concern how the clustering is happening internally, or how the partition works. For example, Silhouette index is an internal CVI and Variation of Information [17] is an external CVI.

Time-Series Clustering Algorithms mainly represent a group of different types of clustering algorithms such as Hierarchical clustering, Partitional clustering, TADPole clustering, Fuzzy clustering. In our experiments, TADPole clustering gave the best results among all the clustering algorithms.

5 PREDICTION MODELS

As described in section 2 above, we model information diffusion process on Twitter as multivariate time series problem. Time series analysis is one of the difficult problems in Data Science and is still an active research interest area. There are many Time Series data examples around us. Predicting the

stock price, predicting the energy price, sales forecasting or predicting energy consuming load, etc. The stochastic nature of these events makes time series forecasting a very difficult problem.

Traditional Time Series analysis follow parametric methodology by decomposing the data into many components such as trend, seasonal and noise components [18]. Techniques such as Auto regression, Moving average, and ARIMA (p, d, q), etc. are used to analyze time series. However, because of the ability of capturing complex structure of time series models, stateful RNNs such as LSTM is found to be very effective in Time Series analysis recently.

5.1 ARIMA

ARIMA stands for Autoregressive Integrated Moving Average. It is a class of model that can capture the temporal structure with different cyclicity in a time series data [5]. ARIMA is most generally used for time series data which can be made to be stationary by differencing (if necessary). A random variable that is a time series is stationary if all its statistical properties are constant over time. A random variable of this form can be viewed as a combination of a signal and noise. The series wiggles around the mean with constant amplitude. ARIMA is a generalization of simpler Autoregressive Moving Average (ARMA) model that adds the notion of integration. This acronym is descriptive, capturing the aspects of the model itself [2]:

AR: Autoregression. The component of the model to forecast the interest variable using linear combinations of past values of that variable.

I: Integrated. The use of differencing steps (i.e. subtracting values at current timestep from values at the previous time steps) to make the time series stationary.

MA: Moving Average. Another component of the model that uses past residual errors in a regression-like model to forecast.

ARIMA (p, d, q) model has three parameters p, d, q:

- p: number of autoregressive terms or the lag order
- d: the number of non-seasonal differences
- q: the size of the moving average window

In terms of Y, the general forecasting equation is:

$$\Big(1 - \sum_{i=1}^{p} \phi_i L^i\Big)(1 - L)^d Y_t = \Big(1 + \sum_{i=1}^{q} \theta_i L^i\Big)\epsilon_t \qquad (1)$$

In Equation 3, L is the lag operator which operates on a value of a time series to produce the previous value, ϕ_i are the parameters of the autoregressive, and θ_i are the moving average parameters, following the convention introduced by Box and Jenkins. To identify the appropriate ARIMA model for Y, firstly the order of differencing (d) needing to stationarize the series must be determined. Later, the gross features of seasonality characteristics in time series Y are removed in conjunction with a variance-stabilizing transformation (logging or deflating). After above steps, the differenced series can merely fit a random walk or random trend model. However, this stationarized series may still have autocorrelated errors. Therefore, some number of AR terms $(p \geq 1)$ and/or some number MA terms $(q \geq 1)$ are also required in the final predicting model.

5.2 Recurrent Neural Network (RNN)

Recurrent Neural Networks (RNN) [14] are a class of neural network which are well suited for sequential data. This makes them a compelling model for time series, forecasting tasks etc. RNN can be built in many ways. One of the simplest ways to understand RNNs is to think of them as a feed forward neural network that has been unfolded in time. Fig. 3 below describe the process of unfolding visually in a RNN. At each time step, the network emits an intermediate output o_t and maintain an internal state s_t , x_t's form the sequential input being fed to the network.The following equations describe the update equations:

$$\begin{aligned} a_t &= b + W s_{t-1} + U.x_t \\ s_t &= tanh(a_t) \\ o_t &= c + V.s_t \\ y_t &= softmax(o_t) \end{aligned} \qquad (2)$$

The matrices U, V and W form the parameters of the model which are learnt by standard propagation. In practice, RNNs have limited usefulness because they suffer from the problem of vanishing and exploding gradients.

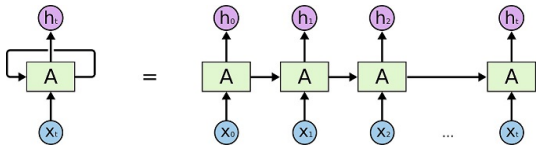

Figure 3: Recurrent Neural Network with loop

The vanishing gradient problem occurs when the gradient values become zero and the exploding gradient problem occurs when the gradient values blow up to infinity. Without going into much details, the reason why this happens is as follows. Because of the chain rule of differentiation, during the propagation step, gradients at each time step are multiplied together. If this value <1, the successive multiplications will drive this value to 0. If this value is >1, successive multiplications drive this value to infinity.

5.3 LSTM for Time Series Prediction

Long Short-Term Memory (LSTM) [14] is a type of recurrent neural network which protects gradients from harmful changes during training and can capture dependencies when there are time lags of unknown size. LSTM can remove or add information to the cell state by regulated gates. The key to this ability is that there is no activation function within the recurrent components. Thus, the stored value, is not iteratively squashed over time and the gradient term doesnt tend to vanish or explode when backpropagation through time is applied to it. Fig. 4 shows the internal gates and connections

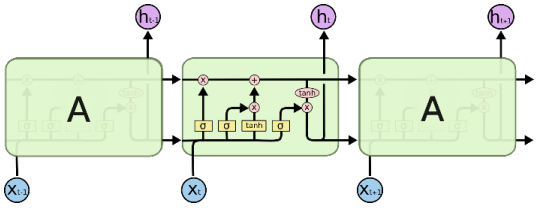

Figure 4: Recurrent Neural Network with loop

of a standard LSTM cell. The following equations describe the update equations of LSTM model:

$$\begin{aligned} i_t &= g(W_{x_i}x_t + W_{h_i}h_{t-1} + b_i) \\ f_t &= g(W_{x_f}x_t + W_{h_f}h_{t-1} + b_f) \\ o_t &= g(W_{x_o}x_t + W_{h_o}h_{t-1} + b_o) \\ c\_in_t &= tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \\ c_t &= f_t c_{t-1} + i_t c\_in_t \\ h_t &= o_t tanh(c_t) \end{aligned} \tag{3}$$

The variables i_t, f_t, o_t are the input, forget and output gates respectively. The gates values can be reset either after feeding each batch or after feeding the entire sequence.

6 RESULTS AND DISCUSSION

In this section, we evaluate the performance of time series clustering and prediction on our collected dataset from Twitter. We first describe the datasets, experimental setup, and then evaluate the performance and present the results.

6.1 Data Descriptions

After preprocessing our dataset as described in section 3, our dataset contains 10 time series subset that are corresponding to 3 dimensions in our information diffusion model: Volume dimension (#tweet and, #retweet), Network Influence dimension (#direct influence user, #indirect influence user), and Sentiment dimension (#positive sentiment percentage, #neutral sentiment percentage, #negative sentiment percentage, #average positive sentiment score, #average neutral sentiment score, and #average negative sentiment score). Each subset of those time series data contains 1,687 samples with 467 measured time steps in hourly basis.

6.2 Time Series Clustering

In clustering operations, the prior decision about the number of clusters carries a lot of importance in obtaining satisfactory results. In this case we performed ten different experiments to do clustering, and the number of clusters is changed every time. So, these experiments are performed for cluster number 4 to 10. After the clustering, standard cluster validity indices (CVIs) are used to determine the best cluster number between 4 to 10. In this case we have used internal CVIs for cluster evaluation. Internal CVIs and their optimization conditions are mentioned below:

Sil: Silhouette index [16] to be maximized to get better cluster.

D: Dunn index [16] to be maximized to get better cluster.

COP: COP index [16] to be minimized to get better cluster.

DB: Davies-Bouldin index [16] to be minimized to get better cluster.

DBstar: Modified Davies-Bouldin index [24] to be minimized to get better cluster.

CH: Calinski-Harabasz index [16] to be maximized to get better cluster.

SF: Score Function [24] to be maximized to get better cluster.

In the present context clustering is performed on the preprocessed Twitter data. The data representing mainly three different dimensions: tweet and retweet count for every hashtag; positive, negative, neutral sentiment score and percentage of tweets on those hashtags; influence count of each of each hashtag. If we consider creating four to ten clusters and comparing their CVIs as one job, a total of ten jobs are performed for each of the parameters. Two jobs for tweet and retweet volume for every hashtag, six jobs for positive, negative, neutral sentiment score and percentage of tweets on those hashtags, two jobs for direct and indirect influence count of each hashtag. While performing the experiments often it has been observed that, not all best CVIs correspond to a number of cluster. In such situations the number of cluster having maximum best values of indices are selected as the optimum number of clusters to be done.

6.2.1 Clustering for tweet and retweet volume for every hashtag. To find out the optimal number of clusters for tweet and retweet volume features, different CVIs of different number of clusters from four to ten are compared. Based on the comparison, the best number of clusters for tweet and retweet volume is six. Hence both tweet and retweet volume data are clustered into six clusters. In Table 1, different CVIs for cluster number six are displayed for tweet and retweet volume features. In this table, the column name represents the feature name and the corresponding number of clusters in parentheses. Fig. 5 and Fig. 6 are the visualization of all six different clusters for tweet and retweet volume. Where x-axis is representing the time in hours and y-axis is representing the count of tweets in each hour.

Table 1: CVIs corresponds to cluster number six for tweet and retweet volume

CVIs	Tweet(k=6)	Retweet(k=6)
Sil	0.006817	-0.0953
SF	0.001150	0.00175
CH	193.2471	55.7518
DB	1.888122	1.50510
DBStar	2.377118	1.88673
D	0	0.00201
COP	8.689435	0.89911

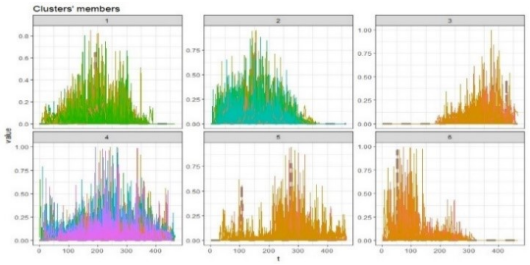

Figure 5: Different patterns of tweets volume

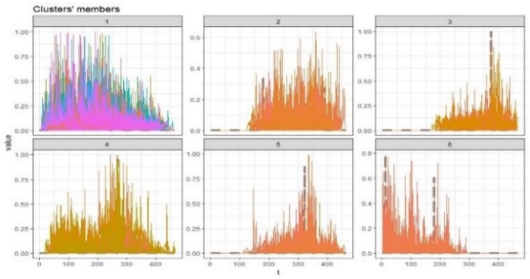

Figure 6: Different patterns of retweets volume

6.2.2 Clustering for different sentiment scores and percentages for every hashtag. Similar to tweet and retweet volume features, cluster analysis has also performed for different parameters of sentiments of tweets. Every tweet is given three different sentiments positive, negative and neutral. Each of the sentiments has two different measures: first is the average sentiment score and second is sentiment percentage which ranges from 0 to 1. The best number of clusters for all the sentiment features is six. While Table 2 shows different CVIs values for sentiment percentage features, Table 3 displays different CVIs values for average sentiment scores features.

Table 2: CVIs corresponds to percentage of positive, negative and neutral sentiment of tweets

CVIs	Positive(k=6)	Negative(k=6)	Neutral(k=6)
Sil	5.896364	0.6213710	$2.4595e^{-02}$
SF	7343.473	0.0316833	$7.4177e^{-09}$
CH	255.7330	389.8150492	$1.0173e^{+03}$
DB	1.386791	1.0448693	1.395693
DBStar	1.627891	1.3182946	1.865864
D	0.000000	0.0000000	0.0000000
COP	3.512364	0.3232937	0.3683004

Fig. 7, Fig. 8, Fig. 9 are the visualization of all six different clusters for percentage of positive, negative and neutral sentiment of tweets. Where x-axis is representing the time in hours and y-axis is representing the percentage of sentiments of tweets in each hour.

Fig. 10, Fig. 11, Fig. 12 are the visualization of all six different clusters for positive, negative and neutral sentiment

Table 3: CVIs corresponds to positive, negative and neutral sentiment score of tweets

CVIs	Positive(k=6)	Negative(k=6)	Neutral(k=6)
Sil	0.049204	0.86700011	0.1666151
SF	0.020461	0.06737498	$5.8312e^{-11}$
CH	540.1516	250.42973	752.1350
DB	1758284	1.83395888	5.116541
DBStar	1.987059	2.16595008	5.552838
D	0.000000	0.0381956	0.0000000
COP	0.060752	0.93276401	$3.68930e^{-01}$

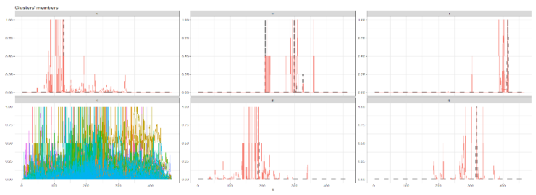

Figure 7: Different patterns of percentage of positive sentiment

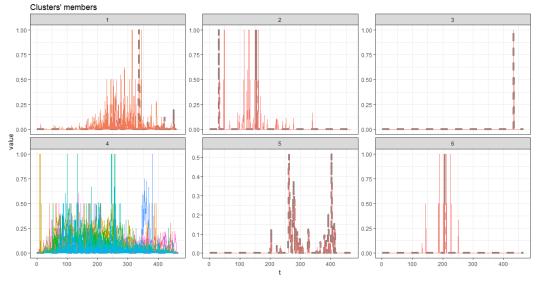

Figure 8: Different patterns of percentage of negative sentiment

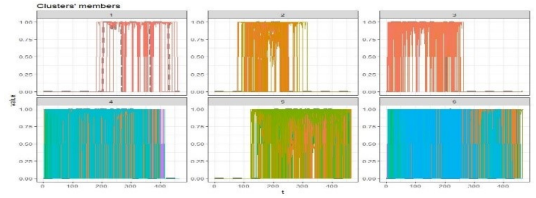

Figure 9: Different patterns of percentage of neutral sentiment

scores of tweets. Where x-axis is representing the time in hours and y-axis is representing the sentiment score of tweets in each hour.

6.2.3 Clustering for network influence dimension. To find out optimal number of clusters for network influence features, different CVIs of different number of clusters from four to ten are compared. Based on the comparison, the best number of clusters for network influence features is four. In Table 4,

CVIs values for direct and indirect network influence features are displayed.

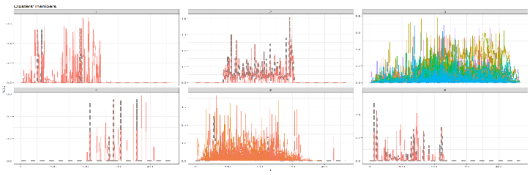

Figure 10: Different patterns of positive sentiment score

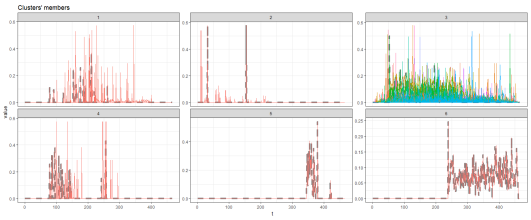

Figure 11: Different patterns of negative sentiment score

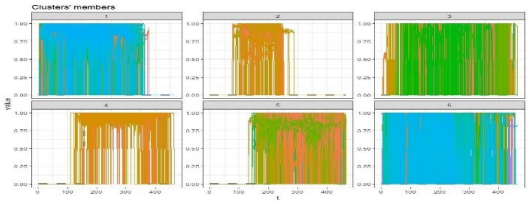

Figure 12: Different patterns of neutral sentiment score

Table 4: CVIs for number of clusters 4 to 10 for direct and indirect influenced users

CVIs	Direct(k=4)	Indirect(k=4)
Sil	$1.351231e^{-01}$	$7.713632e^{-02}$
SF	$2.663825e^{-02}$	$2.097075e^{-01}$
CH	$2.861303e^{+02}$	$2.149411e^{+02}$
DB	1.477699	1.118317
DBStar	1.536565	1.387608
D	$1.550126e^{-03}$	$1.109945e^{-16}$
COP	$9.076306e^{-01}$	$4.512777e^{-01}$

The visualization of different patterns clearly shows some common patterns are present in many features. Some of the very common and easily explainable patterns are discussed here. In Fig. 5 cluster number one is a common pattern. Where the graph grows slowly over time, reaching the peak and then gradually declines; it represents the hashtag with the similar type of popularity growth. A similar type of pattern can be observed in cluster 1 in Fig. 14. The second popular pattern is observed in cluster 4 in Fig. 5, cluster 1 in Fig. 6, cluster 4 in Fig. 7. In all these cases the graphs are showing always a high value. Regarding the volume of tweet, it can be considered as the group of hashtags which are always popular. Some of the patterns are showing very high value in their initial phase and slowly the value decreases with time. In Fig. 5 cluster 6, in Fig. 6 cluster 6 are showing this type of pattern. In Fig. 5 cluster 3 and in Fig. 6 cluster 3 are exhibiting just the opposite of the previous pattern. In these cases, the value is low in the initial time and increases with time. Other than these patterns, some of the hashtags are observed to show spike behavior-the graph suddenly gives a very high value for a very short period of time.

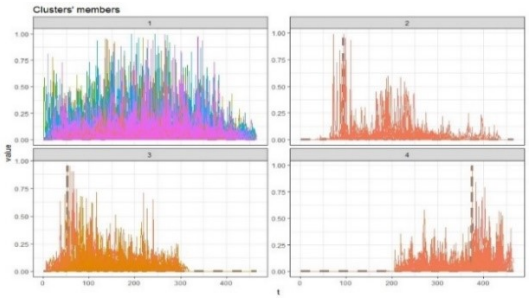

Figure 13: Different patterns of direct influence

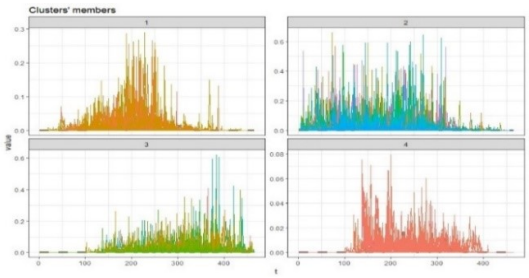

Figure 14: Different patterns of indirect influence

6.3 Classification of Information Diffusion Patterns of New Hashtags

The proposed system also supports the method to recognize the information diffusion patterns of a new popular hashtag and to predict its information diffusion features over time. In previous sections, time series clustering process helps us to identify clusters of patterns for each feature in our information diffusion model. Therefore, these cluster labels can help us to build a classification model to recognize the information diffusion patterns of a new hashtag. In this research, k-NN is used to build such classification model. The procedure is described in below steps:

Step 1: Find the DTW distance between the time series data of the new hashtag and all the time series data points in each of the cluster.
Step 2: Find k closest points from each of the clusters.
Step 3: Find the mean of those selected k points for every cluster.
Step 4: Find the distance between the new data point and the k mean data points (determined in step 3).
Step 5: Find the closest mean point and assign the new data point to cluster that has that closest mean point.

The procedure of building prediction models where ARIMA and LSTM are used is described in the next section.

6.4 Predicting Information Diffusion Process by ARIMA and LSTM

As described in the section 2, we employed two well-known techniques, which are ARIMA and LSTM, to forecast the value of information diffusion time series model. To compare the results between ARIMA and LSTM, we employ the data splitting scheme of 70-30 to divide the dataset. 70% of the total 467 time steps willbe used to train the models and then those models will be used to predict the rest 30%. Root Mean Square Error (RMSE) will be used to evaluate the performance of ARIMA and LSTM models.

6.4.1 ARIMA. For each subset of information diffusion time series, we use grid search to find the corresponding ARIMA model (p, q, d) for each hashtag. As described above, 70% of the total 467 time steps will be used to estimate the ARIMA models. Then the estimated models will be used to forecast the rest of 30% time steps. The total RMSE of prediction for each subset of our time series dataset will be the total sum of prediction RMSE of all hashtags.

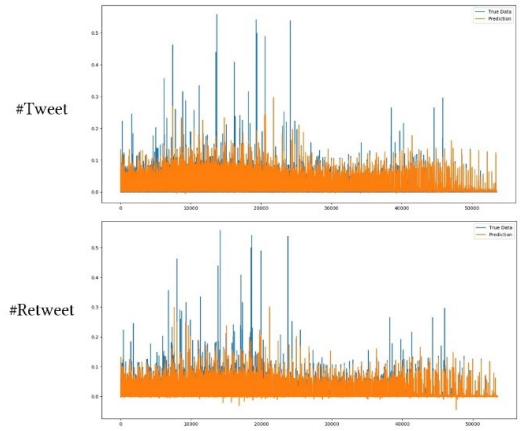

Figure 15: The comparison of true value and predictions using LSTM models on #tweet and #retweet volume datase

6.4.2 LSTM. For each subset of our time series dataset, we train LSTM models with two layers: first layer has 24 cells and second layer has 128 cells. We use the window size of 24 which means our LSTM models use 24 previous values to predict the value at current time step. Moreover, those models were trained with 100 epochs. The charts from Fig. 15 to Fig. 18 display the comparison of actual value and the prediction values of LSTM models that corresponding with 10 variables in our information diffusion models. Nevertheless, Table 5 displays the performance comparison of different ARIMA and LSTM models that were used to predict our multivariate Twitter information diffusion time series dataset.

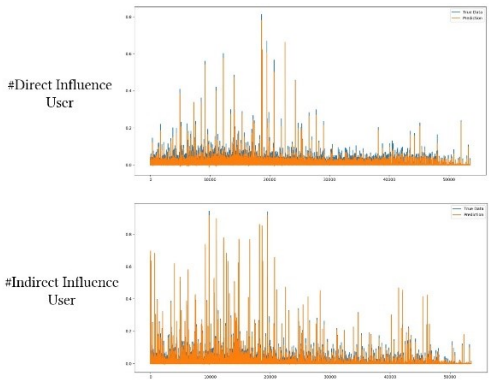

Figure 16: The comparison of true value and predictions using LSTM models on #direct and #indirect influence user dataset

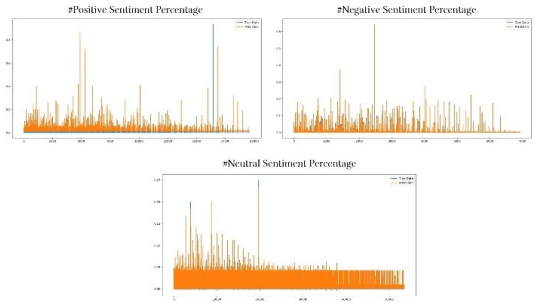

Figure 17: The comparison of true value and predictions using LSTM models on the dataset of percentage of positive, neutral, negative sentiment tweets

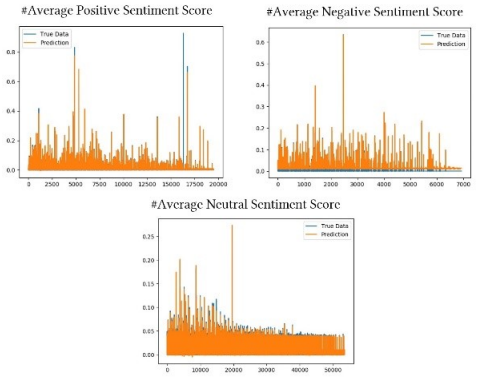

Figure 18: The comparison of true value and predictions using LSTM models on the dataset of average score of positive, neutral, negative sentiment tweets

Table 5: Comparison of testing RMSE when using ARIMA and LSTM for different Information Diffusion parameters

Information Diffusion Model Parameters	ARIMA	LSTM 24×128
#tweet	74.75	0.0089
#retweet	74.75	0.0086
#direct_influence_user	70.96	0.0037
#indirect_influence_user	47.88	0.0038
#positive_percentage	16.40	0.0155
#neutral_percentage	246.33	0.0088
#negative_percentage	1.88	0.0024
#positive_avg_score	16.13	0.01
#neutral_avg_score	230.16	0.0096
#negative_avg_score	1.91	0.0133

The performance comparison in Table 5 shows that LSTM prediction models outperform traditional ARIMA models by far. Moreover, building LSTM models for each cluster of time series can help to achieve better performance than other models.

7 CONCLUSIONS

In recent years, online social media have been increasingly utilized by individuals as well as organizations for a great variety of purposes including communication, entertainment, marketing, crowdsourcing, political messaging, promotion, propaganda, fraud, etc. Characterizing, predicting, and quantifying the key aspects of information diffusion processes on social media, accordingly, has become a research topic of growing interest.

The main contribution of our paper is a general approach to recognize the patterns of, and a model to quantitatively predict, information diffusion on Twitter. We first modeled the information diffusion processes on Twitter as a multivariate time series problem in three dimensions (volume, network influence and sentiment) with a total of 10 features. There are two features in volume dimension which are #tweet and #retweet; two features in network influence dimension which are #direct influence users and #indirect influence users; and six features to quantify the percentage and average score of positive-sentiment, neutral-sentiment, and negative-sentiment tweets. We then collected and processed 27.5 million tweets to develop our information diffusion time series dataset with the 10 features. Different temporal patterns of these features were discovered using time series clustering techniques such as TADPole clustering, hierarchical clustering, and partitional clustering. DTW was used as the distance measure in these clustering techniques.

With the patterns identified, we built an information diffusion prediction model for new topics or memes (hashtags on Twitter). Our prediction model comprises two phrases: the first phrase is determining the pattern of hashtags by using k-NN with DTW distance on our clustering result; the second phrase is building the time series forecasting models using a traditional AIRMA approach and the non-linear LSTM approach. We have built different forecasting models with and without using the pattern information. The performance comparison shows that building LSTM models for each cluster resulted in much significantly better performance than other models. Therefore, we believe that our method holds great promise to be effective in real-world applications of analyzing and predicting the information diffusion processes of new topics or memes in Twitter.

To enhance and refine our proposed model, a better measure of influence (possibly something like influencetracker.com [25]) can be used which draws more inferences than just the count of directly and indirectly influenced people. Secondly, sentiment analysis methods specifically developed for shot texts or tweets [19, 26] will need to be incorporated to provide accurate measures of the sentiments of tweets. Thirdly, a thorough analysis of network structure and its effect on information diffusion is another important direction for future research. Finally, this model should be applied to other social media platforms [11, 12] to evaluate its performance for validation and/or further development.

REFERENCES

[1] Olatz Arbelaitz, Ibai Gurrutxaga, Javier Muguerza, Jesús M Pérez, and Inigo Perona. 2013. An extensive comparative study of cluster validity indices. *Pattern Recognition* 46, 1 (2013), 243–256.
[2] Dimitrios Asteriou and Stephen G Hall. 2015. *Applied econometrics*. Palgrave Macmillan.
[3] Lars Backstrom, Dan Huttenlocher, Jon Kleinberg, and Xiangyang Lan. 2006. Group formation in large social networks: membership, growth, and evolution. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 44–54.
[4] Lars Backstrom, Dan Huttenlocher, Jon Kleinberg, and Xiangyang Lan. 2006. Group formation in large social networks: membership, growth, and evolution. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 44–54.
[5] George EP Box, Gwilym M Jenkins, Gregory C Reinsel, and Greta M Ljung. 2015. *Time series analysis: forecasting and control*. John Wiley & Sons.
[6] Richard Colbaugh and Kristin Glass. 2012. Early warning analysis for social diffusion events. *Security Informatics* 1, 1 (2012), 18.

[7] Information Diffusion. 2017. Information Diffusion. (2017). Retrieved September 30, 2017 from https://github.com/amartyahatua/informationdiffusion

[8] Emilio Ferrara and Zeyao Yang. 2015. Quantifying the effect of sentiment on information diffusion in social media. *PeerJ Computer Science* 1 (2015), e26.

[9] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. 2001. *The elements of statistical learning*. Vol. 1. Springer series in statistics New York.

[10] Mark Granovetter. 1978. Threshold models of collective behavior. *American journal of sociology* 83, 6 (1978), 1420–1443.

[11] Giannis Haralabopoulos and Ioannis Anagnostopoulos. 2014. On the information diffusion between web-based social networks. In *International Conference on Web Information Systems Engineering*. Springer, 14–26.

[12] Giannis Haralabopoulos, Ioannis Anagnostopoulos, and Sherali Zeadally. 2015. Lifespan and propagation of information in Online Social Networks: A case study based on Reddit. *Journal of network and computer applications* 56 (2015), 88–100.

[13] John A Hartigan and Manchek A Wong. 1979. Algorithm AS 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 28, 1 (1979), 100–108.

[14] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[15] Eleanna Kafeza, Andreas Kanavos, Christos Makris, and Pantelis Vikatos. 2014. Predicting information diffusion patterns in twitter. In *IFIP International Conference on Artificial Intelligence Applications and Innovations*. Springer, 79–89.

[16] Minho Kim and RS Ramakrishna. 2005. New indices for cluster validity assessment. *Pattern Recognition Letters* 26, 15 (2005), 2353–2363.

[17] Marina Meila. 2003. Comparing clusterings by the variation of information. In *Colt*, Vol. 3. Springer, 173–187.

[18] Terence C Mills. 1991. *Time series techniques for economists*. Cambridge University Press.

[19] Vu Dung Nguyen, Blesson Varghese, and Adam Barker. 2013. The royal birth of 2013: Analysing and visualising public sentiment in the uk using twitter. In *Big Data, 2013 IEEE International Conference on*. IEEE, 46–54.

[20] François Petitjean, Alain Ketterlin, and Pierre Gançarski. 2011. A global averaging method for dynamic time warping, with applications to clustering. *Pattern Recognition* 44, 3 (2011), 678–693.

[21] Julio Cesar Louzada Pinto and Tijani Chahed. 2014. Modeling multi-topic information diffusion in social networks using latent Dirichlet allocation and Hawkes processes. In *Signal-Image Technology and Internet-Based Systems (SITIS), 2014 Tenth International Conference on*. IEEE, 339–346.

[22] NLTK Project. 2017. NLTK 3.2.5 documentation. (2017). Retrieved September 30, 2017 from http://www.nltk.org/api/nltk.sentiment.html

[23] Ray Reagans and Bill McEvily. 2003. Network structure and knowledge transfer: The effects of cohesion and range. *Administrative science quarterly* 48, 2 (2003), 240–267.

[24] Sandro Saitta, Benny Raphael, and Ian FC Smith. 2007. A bounded index for cluster validity. In *International Workshop on Machine Learning and Data Mining in Pattern Recognition*. Springer, 174–187.

[25] Influence Tracker. 2017. Influence Tracker. (2017). Retrieved September 30, 2017 from http://influencetracker.com

[26] Nan Wang, Blesson Varghese, and Peter D Donnelly. 2016. A machine learning analysis of Twitter sentiment to the Sandy Hook shootings. In *e-Science (e-Science), 2016 IEEE 12th International Conference on*. IEEE, 303–312.

[27] Jaewon Yang and Jure Leskovec. 2010. Modeling information diffusion in implicit networks. In *Data Mining (ICDM), 2010 IEEE 10th International Conference on*. IEEE, 599–608.

View publication stats