Mehran University Research Journal of Engineering & Technology
Vol. 38, No. 3, 687-704 July 2019
p-ISSN: 0254-7821, e-ISSN: 2413-7219
DOI: 10.22581/muet1982.1903.13

Predicting Collective Synchronous State of Sentiments for Users in Social Media

NIDA SADDAF KHAN*, AND MUHAMMAD SAYEED GHANI*

RECEIVED ON 26.03.2018 ACCEPTED ON 30.10.2018

ABSTRACT

The increasing use of social media offers researchers with an opportunity to apply the sentiment analysis techniques over the data collected from social media websites. These techniques promise to provide an insight into the users' perspectives on many areas. In this research, a sentiment analysis model is proposed based on HMC (Hidden Markov Chains) and K-Means algorithm to predict the collective synchronous state of sentiments for users on social media. HMC are used to find the converged state while K-Means is used to find the representative group of users. For this purpose, we have used data from a well-known social media site, Twitter, which consists of the tweets about a famous political party in Pakistan. The time series sequences of sentiments, of each user are passed on to the system to perform temporal analysis. The clustering with three and four number of clusters are found to be significant giving the representative groups. With three clusters, the representative group constitute of 82% of users and with four clusters, two representative groups are found having 45 and 36% of users. Analyzing these groups helps in finding the most popular behavior of users towards the concerned political party. Moreover, the groups perhaps tend to influence the opinion of other users in the network causing changes in their sentiments towards this party. The experimental results show that the proposed model has the power to distinguish behavior patterns of different individuals in a network.

Key Words: Synchronous State, Hidden Markov Modal, Sentiment Analysis, Social Media.

1. INTRODUCTION

With the rapid rise in the use of social media, sentiment analysis of users has become of increasing interest to researchers. Sentiment analysis has often been defined as categorizing the text as positive and negative [1]. Sentiment analysis can be performed to identify the perspectives of social media users, such as their religious and political preferences and the associated issues [2]. Similarly, predictions can be made about a variety of real-life and marketing problems such as movies' rating [3], and the perceptions of parents about getting their children vaccinated at a particular locality [4].Various machine learning algorithms have successfully been used in this field, for example: KNN (K-Nearest Neighbor) and Naïve Bayes on the reviews of movies and hotels [5], a hybrid model of Naïve Bayes and SVM (Support Vector Machine) for Twitter's data [6], a combination of classification and clustering [7], Deep Learning for movies data [8], among many others.

Authors E-Mail: (nskhan@iba.edu.pk, sghani@iba.edu.pk)
* Department of Computer Science, Institute of Business Administration, Karachi, Pakistan.

HMM (Hidden Markov Models) are widely accepted as a statistical tool among the research community for modeling a wide range of time series data [9]. An intuitive picture for complex systems can be drawn using the HMM technique.They are often used for modeling generative sequences which have applications in areas of signal processing, speech processing, and natural language processing. The linear sequence labeling problems have also been efficiently solved by probabilistic models based on HMM.

The novel micro blogging social media service, Twitter, was launched in 2006. According to an estimate, the number of monthly unique visitors to the site is 20 million [10]. Tweets are short messages (up to 140 characters) published by the users. These tweets are visible through the public message board provided by the website and can also be accessed by several third-party applications. Twitter has been supporting worldwide communication of more than one million messages per hour. Twitter users not only post their personal statuses on the website but, in addition, a wide range of topics including politics, product information and reviews are covered.The format of tweets is varied, some comprise short sentences, some have links to the websites and some have direct messages to other users. Due to the increasing scope of Tweets, researchers find it useful to develop techniques to exploit sentiment analysis of this social media platform.

There are many studies [11-12] that have been presented to model synchronization phenomena in a large network of interacting elements to simulate collective behavior under certain assumptions. This research is based on a model proposed by Liang and Ng in [13], where they proposed a probabilistic model (HMM) to discover the collective synchronous behavior in a network of users.The researchers in [13], took the inspiration of their research using the theory of synergetics [14] in Physics which deals with the formation of structures that are self-organized and spontaneous. In our research work, we have used this probabilistic approach to propose a model for sentiment analysis. This analysis is performed on the time series data of interactive users in a network. The HMM model is used to find the collective synchronous state of the system and the groups with the most popular behavior. The dataset that has been chosen is based on the tweets about a political party in Pakistan. There have been various studies where the social media role has been discussed in politics specially in analyzing public opinion about political parties. For example, [15], researchers have proposed a framework for analyzing public opinion, measuring sentiments and information discourse before elections. In [16-17] researchers have analyzed public sentiments about politicians standing for Brazilian presidential elections of 2014. However, to the best of our knowledge, our study is possibly the first attempt where HMM is applied for developing a model for finding converged synchronous state of sentiments over the data collected from social media.

The rest of this paper has been organized as follows. Section 2 presents a brief overview of related literature. Section 3 explains the methodology of the research conducted, while Section 4 presents the details of experiments and results. Section 5 highlights some of the findings and discusses the problems faced in this study. Finally, section 6 concludes the work and suggests the directions for future research.

2. LITERATURE REVIEW

SA/OM (Sentiment Analysis/Opinion Mining) is a well-established field in which a lot of work has been done. A brief overview of existing work in the field of SA was

presented in survey [18]. In this survey, the techniques and algorithms presented by different researchers for Feature Selection and Sentiment Analysis were presented. It also discusses new related fields which lately have become attractive for researchers. There are other research surveys [19-20] where various machine learning algorithms have been presented for sentiment analysis and their respective advantages and disadvantages have also been discussed. Another research work [21] claims to accurately detect the sentiment value of microblogs in a disaster context. They compared four different methods which comprise different combinations of following approaches: SentiWordNet, list of Emoticons, AFNN list, and Bayesian Network.

To the best of our knowledge, [13] is the only work in finding a collective synchronous behavior of users in a social network which is converged over a period, however, there are many other researches where the synchronous and collective behaviors of users on social media have been studied. In [22], a method is proposed to predict the view counts of videos on YouTube using synchronous sharing behavior. This research is based on synchronous sharing pattern in social media where the aim is to find if a video will go viral or not based on its sharing pattern in a group of users. Another research [23] has presented a technique for collective sentiment classification. In this research, features of product popularity and user leniency have been used while the model was tested on two real-world datasets of user reviews. Their experimental results have proved that this approach is better than baseline methods which use n-gram of words as features.

As mentioned in the introduction, our research is based on a model proposed by the researchers in [13], which consists of multiple HMC. The model was trained to predict the collective synchronous behavior and calculate a score for every user to measure the degree of dependence on its connected neighbors. The model considers the network information at two levels i.e., individual (micro level) and system (macro level). It was tested on synthesized datasets that mimic collective synchronous behavior. The synthesized data was created having three types of users; highly reactive, neutrally reactive and lowly reactive. The aim was to find the users who follow other users in a network, evolving the process of convergence. We used this model to analyze the sentiments of users on social media to see the phenomena of collective synchronous behavior convergence, i.e., how individuals follow the sentiments of their friends about any concerned topic over a period and converge to a specific opinion. Although, we have used the model proposed by the researchers in [13], however, our research has two significant contributions over their research work. First, we use a real-world dataset taken from Twitter, rather than using synthesized data as done by [13]. A synthesized dataset is the one which is created under controllable parameters. In contrast, a real-world dataset is taken from a real-world problem domain having many complexities which requires extra techniques and care to be applied. The real contribution of any theoretical concepts can only be seen by applying them on a real-world problem. Our second contribution is that we analyze the opinions of users to find the collective synchronous state of users in terms of sentiments and this is a more useful and complex scenario than the one discussed in [13].

Apart from finding synchronous state, there are many other researches that have been carried out to study the use of HMM in analyzing time series data. In [24], a novel method for the prediction of Time Series was proposed which uses

a combination of HMM and soft computing techniques i.e. ANN (Artificial Neural Networks) or FIS (Fuzzy Inference System). HMM was used to perform the shape-based clustering on the basis of similarity of data. Then ANN/FIS were trained which were used for the prediction of the system. The performance of the model was evaluated by four different time series and benchmarked against MGTS (Mackey Glass Time Series).

Research shows that the unstructured data from microblogging sites could be formulated into time series and analyzed accordingly, using methods to solve for the temporal data [25]. In this research, the authors have analyzed the microblog data to predict the change of sentiments of a crowd on a given topic over time and to identify key features that contributed in bringing that change. HMM are also applied on the social network's data to classify multiple online user streams with low computational complexity [26]. A Multi-HMM i.e. DMMAP (Discrete Markov Multi Arrival Process) was proposed which uses K-Means for clustering and Baum-Welch algorithm to calculate the parameters of the model. The DMMAP effectiveness was validated by comparing the mean, standard deviation, skewness and autocorrelations from un-clustered and Multi-HMM generated user actions.

In [27], some techniques have been created to analyze and classify the time series with a hybrid approach. In this article, the authors performed classification in two stages. In the first stage, they used an ordinary classifier (NN, SVM etc.) while in the second stage they used HMM to incorporate the associated temporal information in the model. Their results show clear improvements by using HMM in the second cycle of classification where the weaknesses of other classification methods are overcome by using temporal data.

Markov processes have very deep roots in analyzing financial time series as well. In [28], the researchers have developed an algorithm based on HMM to predict stock prices of S&P 500 index. They used four different techniques to determine the optimal number of states for HMM, and subsequently this optimized HMM is used to forecast monthly closing prices of the index. The model was compared against the historical average model and was claimed to outperform this model in forecasting and trading.

HMM has also been extensively used in the field of image processing specially to identify and segment the objects in images by making use of the temporal data [29-30]. Authors in [29], claimed to propose a new approach that constructs HMM model in 2D (Two-Dimensional) for recognizing facial images in 2D. Whereas, in [30] a comparison was presented by the same author for 1D and 2D data models. Another study [31] used discrete HMM for the recognition of 3D gestures obtaining an accuracy of 80% for simple gestures and 60% for complex ones.

Many researchers have used HMM for the clustering of time series data [32-33]. The authors in [32] introduced a novel approach for clustering multivariate time series which consists of a mixture of categorical and continuous values. The time series was taken from health care having health trajectories of individuals. The model was tested on various time series and found to be successful in clustering time series of categorical variables. The authors in [33] proposed a novel hierarchical EM (HEM) methodology, to cluster HMMs on the basis of similarity and to find a representative HMM for each cluster. The proposed model was claimed to have solutions in many problems i.e. automatic retrieval and annotation of music, hierarchical clustering of motion capture sequences, handling of online-handwriting data, etc.

HMM has its application in the analysis and prediction of human activities and its states by utilizing the temporal sequence patterns [34]. In a recent literature [35], PU (Primary User) channel state future prediction has been widely investigated for predicting PU channel state based on time series and HMM. A brief overview of other research work has been presented in this section. In next section, complete detail of model and approach is described.

3. METHODOLOGY

The methodology of this research is comprised of the following: system description, data acquisition, data preprocessing, and model building and prediction. System description is comprised of the technical background of this research. Data acquisition and preprocessing describe the procedure of data retrieval, cleaning and transformation while model building and prediction describe the details of the model and its predictions.

3.1 System Description

This research study is based on the technique proposed in [13]. They propose a model Co Sync which predicts the Collective Synchronous behavior of people on social media. They predict the steady future state of people and measure Reactive Factor to quantify the degree of dependence of each individual to other users on the basis of their observations. Their model stood on the following assumptions: (1) All users can be seen by other users, (2) Any pair of users can interact with each other, (3) collective movements triggered by some external changes or events. Below are some notations to represent the model:

(1) M is the total number of individuals/users.

(2) Time t is taken as discrete, till T of stochastic process. e.g. $t = 0,1,2,3 \ldots T$.

(3) $X_p^{(t)}$ is a random variable which is discrete to represent an individual p at time t.

(4) R_t is a random variable to represent the statistical aggregate state of system for all users.

(5) S is the finite state space with N total states. i.e. $s_n \in S, n = 0,1,2, \ldots N$.

(6) π is stationary probability distribution over N states.

(7) A is transition matrix.

(8) W is emission matrix.

The behavior of users was modeled by many MMC with the assumption that the state of any user at time t depends upon its own state, and on the state of other users in that network at time t-1.

$$\Pr\left(X_p^{(t)} \middle| X_p^{(t-1)}, X_p^{(1-2)}, X_p^{(1-3)}, \ldots, X_p^{(1)}, X_p^{(0)}\right) = \Pr\left(X_p^{(t)} \middle| X_p^{(t-1)}\right) \quad (1)$$

$$\Pr\left(X_p^{(t)} \middle| X_{p-1}^{(t-1)}, X_{p-2}^{(1-2)}, X_{p-3}^{(1-3)}, \ldots, X_1^{(t-1)}, X_0^{(t-1)}\right) = \Pr\left(X_p^{(t)} \middle| R_{t-1}\right) \quad (2)$$

These assumptions made the model to have multiple coupling MMC as shown in Fig. 1, where $X_p^{(t)}$ is a random variable for an individual p, at time t.

Since the network may have many users which makes the process of inferring very time-consuming for all random variables (users). For simplification, a variable R_t is

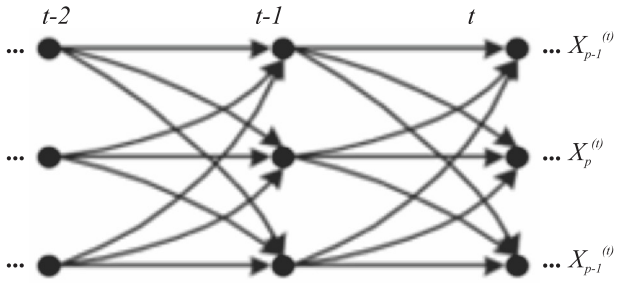

FIG. 1. COUPLING MARKOV CHAIN [13]

introduced to represent the aggregate state of the system and it is referred as the macro variable having the statistical result of all states over all users. The previous state of the system i.e. at previous time step, given an individual X_p and system variable R, determines the probability of a user X_p. This representation gives a simplified model by decomposing coupling MMC into a set of connected HMM models as shown in Fig. 2.

The terms System Synchronous State and Reactive Factor are defined as follows:

System Synchronous State: A system that contains P individuals and $Pr(R_t = s_n)$ where $s_n \in S$ is defined as a probability distribution for N system states. If sufficient time is given for evolving, the system synchronous state is the largest steady probability state in $Pr(R_t = s_n)$ as shown in Equation (3).

$$S_{syn} = \{s_n \,|\max(Pr(R_t = s_n))\}, 1 << n << N, t \infty \quad (3)$$

Reactive Factor: The Reactive Factor (RF) for each individual is defined in Equation (4).

$$RF(X_p) = Pr(X_p = s_n \,|\, R = s_n)\,, s_n \in Se \quad (4)$$

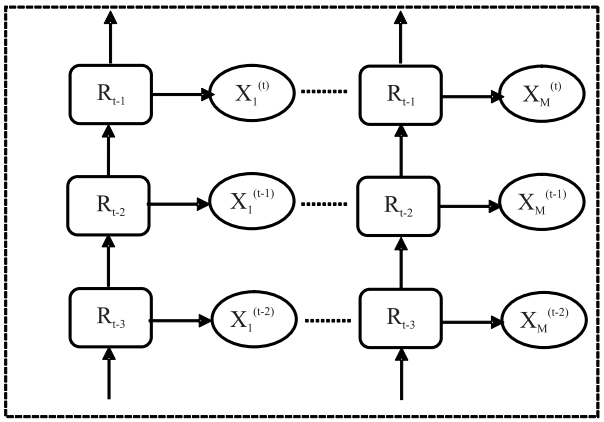

FIG. 2. COUPLING HMM

The collective synchronous behavior emerges when sufficient time for evolving is given and the stationary probability distribution of user has the coherence with the probability distribution of a complete system i.e. $Pr(X_p^{(t)})=Pr(R_{t-1})$. The equation of evolution of MMC is given below for which the state probability distribution $Pr(R_t)$ and the RF$Pr(X_p^{(t)}|R_{t-1})$, are needed.

$$Pr\left(X_p^{(t)}\right) = \Sigma_{R_{t-1}} \; Pr\left(X_p^{(t)}\middle|R_{t-1}\right)Pr\left(R_{t-1}\right) \quad (5)$$

An analogy was made by the comparison of the phenomena of collective synchronous behavior and the properties of HMM. It is based on the assumption that behavior of a user is a MMC process and there exists a probability distribution to control the intrinsic preference behavior which determines the state of a person, on N candidate states. In contrast, it is also observed that a person's behavior may be affected by its neighbors and may cause the change in its corresponding probability distribution of preferences. The preference probability distribution will tend to be stabilized and if left to evolve for a long time,will give rise to the collective synchronous behavior. As R is the macro variable representing the statistical aggregate state of all users, it will also experience Markov process converging to a stationary state. It can also be taken as a hidden variable in 1-order MMC because the system evolving states may not be observed from outside. Analogously, a series of states given by each user can be treated as the observation sequence of HMM. It is claimed that the emission probability $Pr\,(X_p^{(t)} \mid R_{t-1})$ by HMM and the conditional probability $Pr(X_p^{(t)} \mid R_{t-1})$ by MMCgiven in Equation(5), are exactly same.

The parameters of HMM i.e., transition probability matrix and emission probability matrix are learned by using

Baum-Welch Algorithm [36], which is an iterative algorithm to estimate these parameters based on observation data sequences. The proposed macro variable R was assumed to have a random walk process having transition probability and a steady state probability distribution stating the chances of being in each state after a long duration of time. This stationary probability vector π can be found out by eigenvector of transition probability matrix A shown in Equation (6).

$$\pi = \pi.A \tag{6}$$

The condition of irreducibility in Equation (7) is satisfied since it was allowed that each state is accessible by other states without restrictions.

$$\forall s_i, s_j \in S, \exists' t >> 0, \Pr(X^{(t)} = s_j \mid X^{(0)} = s_i) > 0 \tag{7}$$

The proposed model also satisfies the condition of aperiodicity as people have complete liberty to make their decisions. They can stay with their current state or come back later in time to their previous state. These assumptions guarantee unique stationary probability distribution.

A transition matrix A and a stationary probability vector π for each user are given by the coupling HMM. The purpose of this model is to find a collective synchronous state, of a subset of P people which shows similar behavior. To reflect these majorities clustering method K-Means is applied on the stationary probability vectors π of all the users to divide them into the clusters on the basis of similarities among them. The largest cluster is selected as the representative group C* of this network exhibiting the behavior of majority users. The clustering is done on the vectors of stationary probability distributions where each vector is mapped to an N-dimensional data point according to state space S. Because the converging process will evolve and converge to one of the states N hence the number of clusters K are kept equal to the number of states N. All the transition matrices of users who belong to the representative cluster (cluster having maximum members) are merged to give asingle representative state transition matrix A* of this group. The merging is achieved by taking the statistical mean of corresponding entries in all the transition matrices of the representative group. A representative stationary probability distribution π* is also calculated to give the steady state behavior of this representative group.

After finding the representative HMC, all individual HMC are recalculated so that their corresponding emission matrices B can be readjusted. The transition matrix A' and emission matrix B' are re-estimated by Baum-Welch algorithm setting A' = A* for all HMC. The clustering is performed again on updated stationary probability vectors π' and a renewed candidate group C*' is obtained. Then again the updating in representative transition matrix A*' and representative stationary probability matrix π*' are calculated using C*'. This process repeated iteratively to improve A* and B, till π* gets stabled under a certain threshold υ of error where the error is calculated by Euclidean distance. The mechanism of evolution inside a system is represented by A* and the degree of reactive influence with respect to each system state is represented by B. The synchronous state S_{syn} of the system is predicted, which is the maximum probability in π*. The detailed algorithm is given in Table 1. Reactive factor RF of a user, is used to measure the degree of dependence of everyone to other users in network. Reactive factor could be positive RF_p and negative RF_n, where positive reactive factor measures the chances that a state acquired by system would also appear on an individual as shown in

Equation (8). On the other side, negative reactive factor calculates the chances a state appears on system and the individual will acquire a different observation state.

$$RF_p = \sum_{n=1}^{N} \frac{b_{nn}}{N} \quad (8)$$

Another type of score is calculated using the π^* by taking the average of positive reactive factor as shown in Equation (9).

$$WRF_p = \sum_{n=1}^{N} \pi_n^* * b_{nn} \quad (9)$$

The experiments were performed on a synthesized dataset created by the authors to simulate the collective synchronous model. A dataset of 100 users was created having one of three dependence properties i.e., highly reactive, neutrally reactive and lowly reactive. The highly reactive group contains the people having WRF greater than 0.7, neutrally and lowly reactive group contains individuals having WRF nearly equaling to 0.4 and 0.1 respectively.

Our research work is based on the theories presented in aforementioned research [13]. We have used the coupling MMC to model the state of sentiments of each user. The state of sentiments of a user at time t is not only dependent on his own state, but also on other users' states at time t-1. The approach of macro variable R_t is used to represent the aggregate state of system at time t so that the system can be modeled by coupling HMM. The state of a user at time t now depends only on the state of the system at time t-1. It is known that the sentiments of a user may also affect the sentiments of his friends in a social network. The collective synchronous behavior is caused by the preference probability of a group of people which will tend to be similar and stabilized, if the system is left to evolve for a long time. Rather than applying the model on a synthesized data, we have used a real-world dataset to measure its true usefulness. The dataset used for this purpose is taken from a well-known social media site, Twitter.

The proposed approach consists of the following three phases: (1) Data Acquisition, (2) Data Preprocessing, (3) Model Building (Learning Parameters) and synchronous state prediction.

3.2 Data Acquisition

To apply this analysis data is acquired by using a tool NodeXL, which is a tool designed to download and performs the network analysis. A dataset of Tweets is downloaded from Twitter containing several fields about a specific hash tag. Hash tag is used to download a specific set of networks containing the Tweets about a topic. For this research, Tweets are downloaded about a Pakistani political party to analyze public sentiments towards this party.

3.3 Data Preprocessing

In the phase of Data preprocessing following steps are taken to prepare the data for model building:

Feature Selection: The downloaded data contains several fields e.g. Sender, Receiver, Relationship, Relationship Date, Tweets, Tweets Date, URLs in Tweets, Domains in Tweets, Hashtags in Tweets, Twitter Page, Imported etc. The unnecessary fields were deleted to obtain the filtered data fields which are required for model building. After cleaning, we were left with the following fields: Sender, Receiver, Relationship, Tweets, and Tweets Date.

Missing Values: Records having missing values were deleted to clean the data.

Data Preparation: Parsing and cleaning is performed on the Tweets to delete the terms having '#'and '@'. On Twitter terms following # are specific topics which are tagged and terms following '@' are the user names. After cleaning, Tweets are annotated and given the polarity of Positive, Negative and Neutral based on the sentiments in the text of Tweets. To achieve the higher accuracy, the Tweets are annotated manually so that the rate of error could be minimized. The Tweets are assigned the uniform time stamps i.e. one-time stamp is comprised twelve hours giving two time stamps for one day. If a user has multiple Tweets in a single time stamp, then Mode is used to find the representative polarity of that time stamp of a particular user. And if a user does not have any Tweets in a particular time stamp then that user is assigned the polarity of its previous time stamp.

3.4 Model Building and Prediction

Collective synchronous behavior is a commonly observed behavior in social networks. People like to buy the things which their friends have bought and found them useful or joyful. Same is the case with opinion in that people often tend to have the same opinion about an event or a product that their friends have in their social network. This approach gives rise to the convergence to a synchronous state of sentiments of the system. This theory has been used to model the behavior of users in a social network and has been applied in this research to model their state of sentiments. A Stochastic HMM could be learned with the following assumption into consideration: all users are connected to each other in a social network and any two users may interact with each other [13]. The behavior pattern of Twitter users is modeled by a set of HMC. To model this system by Markov process, the second most important assumption is that the state of a user at a specific time t depends upon his state at time t-1. Hence, each user formulates a HMC of its sentiments over time. To model the collective behavior of the whole system we also need to consider the collective influence of other's opinion in a network. Hence, the state of a user at time t not only depends upon its own state but also on the state of other users at the previous time. The resultant Markov process is now changed a bit where the hidden state represents collective state of sentiments of whole network and observation signals represents the sentiments of an individual. The coupling HMM is shown in Fig. 3. The model contains three hidden states: positive (P), neutral (Nu) and negative (Ng) to represent the sentiments of the whole network. The observations state to represent output signal of an individual at a particular discrete time have the same three signals i.e. (P, Nu, and Ng). A generalized state diagram is shown in Fig. 3.

Since all the users are connected and visible to each other the behavior of some users may affect the behavior of other users causing the change in their corresponding probability distribution over N states. If the system is left

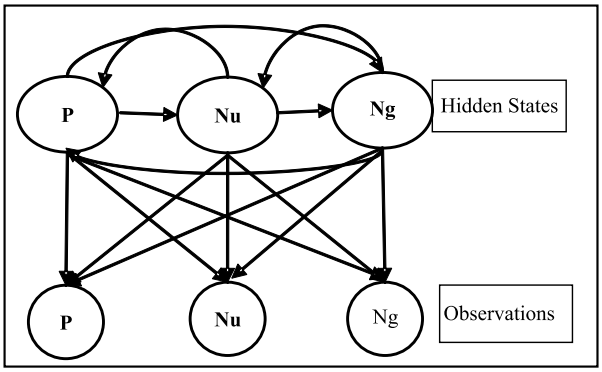

FIG. 3. THE GENERALIZED STATE DIAGRAM OF THE MODEL

for a long time, the probability distribution exhibiting the preferences of group of users will be stabilized giving the converging state of the system.

The individual observation sequences, along with the initial guessed transition table, emission table and initial system stationary probability matrices are then passed to a function called Synch Sentiment. This function calculates transition probability, emission probability and stationary probability matrices of each user. Then K-Means clustering is performed on all the stationary probability matrices and cluster having the largest records is identified. This cluster represents the behavior of majority of the users in a network. The average transition probability matrix is calculated by taking the average of all the users' transitions tables that belong to the maximum sized cluster. Based on this average transition table, the stationary probability is obtained the way given in Equation(10):

$$\pi^{*\prime} = \pi^{*} \times A^{*} \tag{10}$$

where π^{*} is initial guess for stationary probability and A^{*} is the average transition table of the largest cluster. This state having the maximum probability is the Synchronous state of the system in which it is most likely to be. By analyzing the clusters, we can also comment about the most popular behavior of the system. The complete algorithm to learn the HMM of the system is provided in Table 1. These probabilities formulate the system to predict the synchronous state of the system that can give the converging point of sentiments. The complete process model of this study can be seen in Fig. 4.

4. EXPERIMENTS AND RESULTS

The data was downloaded from Twitter using the NodeXL tool. The hashtag used is a political party of Pakistan which has a mixed opinion in Pakistan. The reason to choose this hashtag is that we want to collect the Tweets about a topic where we can get a good mixture of positive, negative and neutral sentiments so that the imbalance class problem will not arise. For this research, we used eleven users having twenty-five Tweets over a period of Seventy-Two hours. Each user Tweets are divided in the group of twelve hours giving total six time stamps for each user. The sixth time stamp had no Tweet from any user, so we ignored this time stamp which left us with five time stamps. After data preparation and Sentiment tagging, we had five-time stamp observations for each user so that we can now apply HMM model on the data.

To learn parameters of HMM model the Baum-Welch algorithm has been used which is provided with the initial

TABLE 1. SYNCHSENTIMENTS ALGORITHM TO BUILD THE MODEL BY HMM AND K-MEANS CLUSTERING FOR PREDICTING THE COLLECTIVE SYNCHRONOUS STATE

[A*, π *', W_M] = SynchSentiments (x_M, A_0, W_0, π_0)

Required Parameters:
x_M = Observation sequences of all M persons
Initial guess matrices for:
W_0 = Emission matrix,
A_0 = Transition matrix,
π_0 = System stationary probability distribution
N ← Number of hidden states in A_0
All A_M ← A_0
All W_M ← W_0
π* ← π_0
form ← M **do**
[A_m' W_m'] ← **LearnHMM Parameters**(A_m, W_m, X_m)
$\pi_m' = \pi_m . A_m'$
End for
Initialize Clusters {C_1,…,C_K}
K ← N
{C_1,…,C_K} = **K-Means** (π_1', π_2',….,π_M', K)
C* ! {C_g | Max(|C_1|,|C_2|,… ,|C_g|, … , C_K|)}
For d = 1 to |C*| do

$$A^{*} = \left[a_{ij}\right] \leftarrow \sum_{d-1}^{|C^{*}|} \frac{a_{ij}^{d}}{|C^{*}|}$$

End for
π *' ← π*. A*
All A_M ← A*
Return A*, π *', W_M

transition matrix, initial emission matrix and the sequence of observation. Since we did not have the initial parameters, so we estimated them. The initial transition matrix was assigned value of 1/3 as a prior probability and the initial emission matrix was learned by using the observation sequence. The stationary probability vector for each user can be found by eigenvector of the state transition matrix of that user as shown in Equation (11).

$$\pi' = \pi \times A \tag{11}$$

where π is the initial stationary probability vector which is assigned prior value of 1/3 for each entry and A is the state transition matrix learned by Baum-Welch. K-Means Clustering is applied on the stationary probability matrices of each user. The clustering schemes are analyzed for different number of clusters (K) where K was set from 1-6. To evaluate and find the optimum clustering scheme, SSE (Sum of Squared Error) is calculated. A graph of SSE vs. K (number of clusters) is plotted to find the best K for our data distribution as shown in Fig. 5. The ideal number of clusters should be picked in a way that adding another cluster does not give better modeling of dataset.

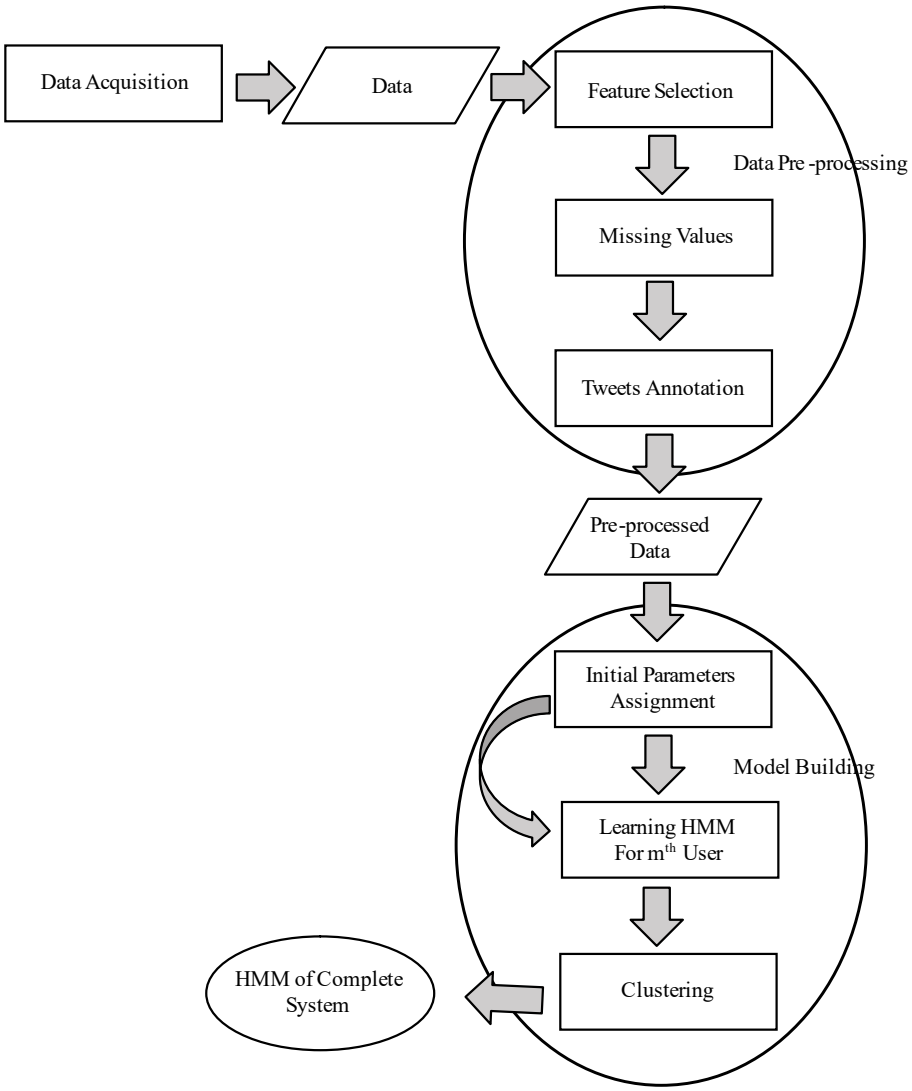

FIG. 4. COMPLETE PROCESS DIAGRAM

From the graph in Fig. 5, it could be seen that points K=3 and K=4 are two points which can be considered as the more appropriate number of clusters for the data under consideration. The detail of each cluster is given in Table 2. We analyze data by both keeping K=3 and by keeping K=4. Both revealed interesting insights of data which are discussed next in this section. The clustering performed on stationary probability matrices of each user are shown in Table 3.

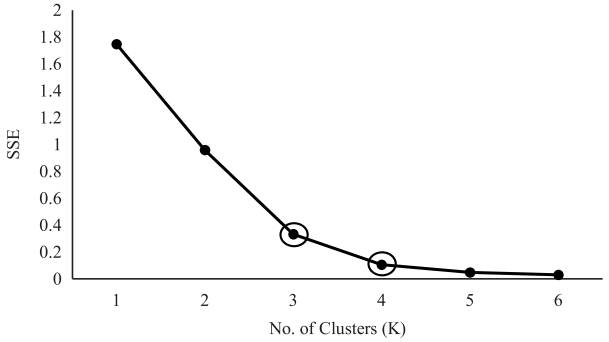

FIG. 5. PLOT OF SUM OF SQUARED ERROR V NUMBER OF CLUSTERS (K) TO FIND OPTIMUM CLUSTER SIZE

TABLE 2. CLUSTERS PROPORTION WITH K=3 AND K=4

Clusters	K = 3	K = 4
C1	81.82%	36.36%
C2	9.1%	9.1%
C3	9.1%	9.1%
C4	Ñ-	45.45%

4.1 CLUSTERING SCHEME (K = 3)

When the number of clusters are kept three then we have clusters having following proportion of data points: C1 has almost 82% of users, C2 and C3 have 9% of users in each as shown in Fig. 6. The largest size cluster is C1 having the largest number of users. The average Transition table of this cluster is shown in Table 4 and the Stationary probability of this group is given in Table 5.

The analysis of cluster C2 shows that it contains the members who possess neutral sentiments in the long run making this group of users undecided about their emotion towards the concerned political party. Cluster C3 contains users having extreme negative emotions in the long run for the concerned political party. Its analysis reveals that the users in this group initially had neutral opinion which changed into negative and stayed negative while never showed any positive emotions till the end of observation. The clusters C2 and C3 are the clusters having very small proportion i.e. 9.1% which is not significant and hence cannot be considered as the popular behavior in this network. But cluster C1 holds the largest proportion i.e. more than 81% representing the largest group and popular behavior in this cluster modeling scheme. Hence, the collective Synchronous behavior of this cluster shows

TABLE 3. STATIONARY PROBABILITY MATRICES WITH K=3 AND K=4

Users	Stationary Probability Matrices			K = 3	K = 4
1.	0.25	0.38	0.38	C1	C4
2.	0.6	0.2	0.2	C1	C1
3.	0	0	1	C3	C3
4.	0.6	0.2	0.2	C1	C1
5.	0	1	0	C2	C2
6.	0.25	0.5	0.25	C1	C4
7.	0.51	0.3	0.2	C1	C1
8.	0.38	0.38	0.25	C1	C4
9.	0.38	0.38	0.25	C1	C4
10.	0.22	0.3	0.48	C1	C4
11.	0.51	0.3	0.2	C1	C1

that at any given time 42% of the users would belong to the Positive state, 29% of the users would have neutral sentiments and 29% of the users would have negative sentiments. The probability of a random user belonging to the positive state is greater than the probabilities of belonging to negative and neutral states. So, the concerned political party can focus on the members of this cluster to design political campaign which is tailored particularly to target the users of this group because their tendency towards positive state is greater than negative and neutral state.

4.2.1 Clustering Scheme (K = 4)

When number of clusters are kept four then we have following proportion of data points: C1 has 36.4% of users, C2 and C3 have 9% of users and C4 has 45.45% of users as shown in Fig. 7. Cluster C4 has the highest number of users hence it is the cluster which is the representative of maximum group size when K = 4. The cluster C1 is the second largest cluster giving us the notion that C4 is not the only cluster which represent this data set. Hence, in this clustering scheme, we have two clusters possessing the representative behavior of majority of the users, therefore, both clusters are analyzed. The average Transition table of both (largest and second largest) clusters are given in Tables 6-7 respectively. The Stationary probability of these clusters are given in Tables 8-9 respectively.

In this clustering scheme, the clusters C2 and C3 are same as we had in the clustering scheme when we kept K = 3. These two clusters are exactly the same showing the same behavior i.e. C2 is the group of users showing undecided

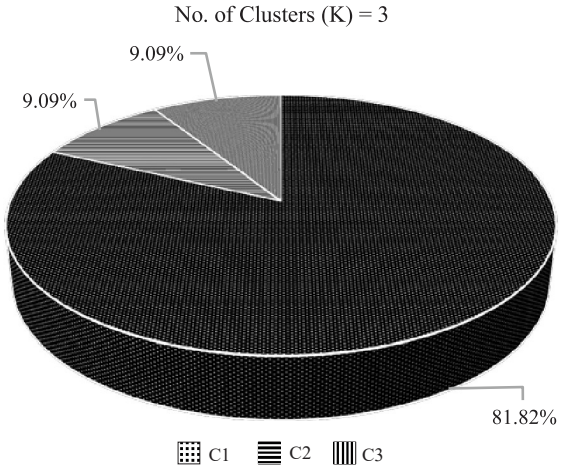

FIG. 6. PROPORTION OF CLUSTERS WHEN K=3

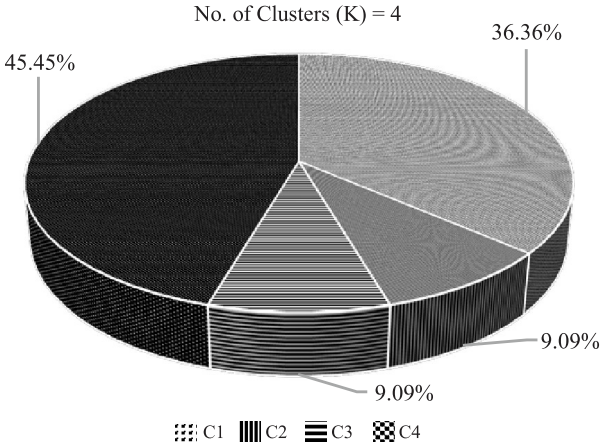

FIG. 7. PROPORTION OF CLUSTERS WHEN K=4

TABLE 4. AVERAGE TRANSITION TABLE OF C1 WHEN K=3

	P	Nu	Ng
P	0.24	0.36	0.39
Nu	0.45	0.22	0.33
Ng	0.65	0.27	0.09

TABLE 5. STATIONARY PROBABILITY MATRIX OF C1 WHEN K=3

	P	Nu	Ng
π *'	0.42	0.29	0.29

behavior and C3 is the group having users showing negative behavior or can be seen that they have extreme negative behavior. Since their respective proportions are very small in the comparison of C1 and C4, so they do not exhibit the representative behavior of the population. The cluster C1 which we had in the previous clustering scheme i.e. K = 3, is now broken down into two clusters C4 and C1 when we did clustering keeping K = 4. This gives us greater insights about C4 and C1 that they must possess some different behavior which causes them to lie in different clusters. The steady state behavior of the individuals belonging to this cluster shows that there is 30% chance that users belonging to this group will have positive state, 35% that they will belong to neutral or negative state. Their converged behavior shows that the users in this group are more inclined towards negative and neutral state than positive state. While the steady state probabilities of second largest cluster C1 shows, there is 54% chance that users in this group will have positive state, 26% chance they will have neutral state or 20% chance that they will have negative state. The converged behavior of this group shows clear inclination towards positive state suggesting that this group could be taken as a group of party supporters. But the party should deal with this group with care because there is also a chance although little, that they might have gone towards the negative state for the party. The combined analysis clusters C1 and C4 by both clustering schemes is shown in Fig. 8.

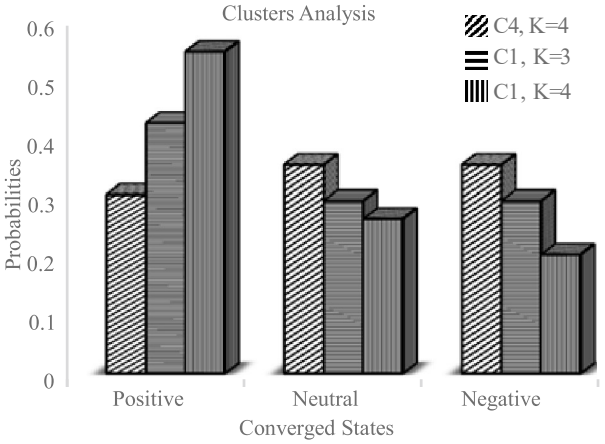

FIG. 8. STEADY STATE PROBABILITIES OF CONVERGED BEHAVIOR OF POPULAR BEHAVIORS WHEN K =3 AND K =4C

TABLE 6. AVERAGE TRANSITION TABLE OF C4 WHEN K=4

	P	Nu	Ng
P	0.18	0.34	0.47
Nu	0.17	0.31	0.52
Ng	0.52	0.4	0.08

TABLE 7. STATIONARY PROBABILITY MATRIX OF C4 WHEN K=4

	P	Nu	Ng
π *'	0.3	0.35	0.35

TABLE 8. AVERAGE TRANSITION TABLE OF C1 WHEN K=4

	P	Nu	Ng
P	0.32	0.39	0.29
Nu	0.8	0.1	0.1
Ng	0.8	0.1	0.1

TABLE 9. STATIONARY PROBABILITY MATRIX OF C1 WHEN K=4

	P	Nu	Ng
π *'	0.54	0.26	0.2

As it has already been stated that the cluster C1 with K=3 is equal to the sum of users in C4 and C1 with K = 4. The graph in Fig. 8 shows that the cluster C1 with K = 3 is giving a sort of averaged behavior for every state (Positive, Neutral, Negative). It is the collection of users having mixed opinions about the concerned political party that can easily be seen through their converged behavior (steady state behavior). The clusters C4 and C1 with K = 4, separate these people into two groups giving us better information about their polarities of sentiments by identifying one group with somehow neutral behavior (C4) and one group with clear tendency towards positive state (C1). This way we can say that clustering scheme with clustering size (K) = 4 is better approach to find the groups having distinct behavior from our data.

5. FINDINGS

To get the maximum advantage from coupling HMM model we need to have a larger number of users and for each user, we should have sufficient sequence of observations so that the algorithm can understand the behavior of each user and can anticipate the influence of other users. For this research, we have used a real-world problem data downloaded from a social media site unlike the work in [13] where they have applied coupling HMM on the synthesized dataset which they created themselves. In our dataset, we faced some very critical issues. For example, there were some time stamps where we had multiple Tweets, so we had to find a way to represent the polarity of that time stamp which we did by using Mode. Similarly, there were some cases when we did not have any Tweets for a particular time stamp, so we had to find a way to assign a polarity to this time stamp which we accomplished by assigning the polarity of its previous time. The real contribution of theoretical concepts can only be seen by applying them on a real-world problem. In this study, we have analyzed the Synchronous and the most popular behaviors in a network in terms of sentiments. Our analysis reveals that the political party under consideration will have a Positive state after the representative group converges with higher probability. It results in the most popular behavior among people of this network. It is to be noted that the findings of this research will not only assist the political party under consideration but also other parties. This is because the users with negative sentiments can be used by other party to get attention of such users towards their party. Therefore, this analysis could benefit all types of parties to devise their political campaign in a way to increase their vote bank while decreasing the vote bank of their opponent party.

6. CONCLUSION

This work attempted to develop a model based on HMC for analyzing the sentiments of social media users. The data for this research was collected from Twitter, which has been selected because of the recent increasing population and usage of social media in general and Twitter in particular. Tweets about a famous Pakistani political party which has a mixed reputation lately in the news were studied. We used a coupling HMM for finding collective synchronous behavior. We used this technique for modeling and predicting collective state of sentiments for users in a network. Rather than using this model on a synthesized dataset as done by past researchers, we have applied it on real world data. In this process we

encountered several issues including varied frequency of individual tweets and had to assign values based on past trends. Additionally, in each time stamp, we had to use statistical mode to assign representative polarity. In conclusion we were successful in finding the collective synchronous state of people towards the concerned political party. In the future we recommend that this technique can be further improved by using a larger dataset.

ACKNOWLEDGMENT

Authors would like to acknowledge the funding support provided by Research Funding & Publication Committee, Institute of Business Administration, Karachi, Pakistan.

REFERENCES

[1] Pawar, A.B., Jawale, M., and Kyatanavar, D.N., "Fundamentals of Sentiment Analysis: Concepts and Methodology", Sentiment Analysis and Ontology Engineering, Volume 639, pp. 25-48, 2016.

[2] Ceron, A., Curini, L., Iacus, S.M., and Porro, G., "Every Tweet Counts? How Sentiment Analysis of Social Media Can Improve Our Knowledge of Citizens' Political Preferences with an Application to Italy and France", New Media & Society, Volume 16, No. 2, pp. 340-358, 2014.

[3] Singh, V., Piryani, R., Uddin, A., and Waila, P., "Sentiment Analysis of Movie Reviews: A New Feature-Based Heuristic for Aspect-Level Sentiment Classification", International Multi-Conference on Automation, Computing, Communication, Control and Compressed Sensing, Kottayam, India, 2013.

[4] Salathé, M., and Khandelwal, S., "Assessing Vaccination Sentiments with Online Social Media: Implications for Infectious Disease Dynamics And Control", PLoS Computational Biology, Volume 7, No. 10, 2011.

[5] Dey, L., Chakraborty, S., Biswas, A., Bose, B., and Tiwari, S., "Sentiment Analysis of Review Datasets using Naïve Bayes' and K-NN Classifier", arXiv preprint arXiv:1610.09982, 2016.

[6] Hasan, A., Moin, S., Karim, A., and Shamshirband, S., "Machine Learning-Based Sentiment Analysis for Twitter Accounts", Mathematical and Computational Applications, Volume 23, No. 1, pp. 11, 2018.

[7] Coletta, L.F.S., Silva, N.F.F., Hruschka, E.R., and Hruschka, E.R., Jr., "Combining Classification and Clustering for Tweet Sentiment Analysis", Brazilian Conference on Intelligent Systems, 2014.

[8] Pouransari, H., and Ghili, S., "Deep Learning for Sentiment Analysis of Movie Reviews", Technical Report, Stanford University, 2014.

[9] Zucchini, W., MacDonald, I.L., and Langrock, R., "Hidden Markov Models for Time Series: An Introduction Using R", Boca Raton, Chapman & Hall/CRC, 2016.

[10] Tumasjan, A., Sprenger, T.O., Sandner, P.G., and Welpe, I.M., "Predicting Elections With Twitter: What 140 Characters Reveal About Political Sentiment", ICWSM, Volume 10, No. 1, pp. 178-185, 2010.

[11] Acebrón, J.A., Bonilla, L.L., Vicente, C.J.P., Ritort, F., and Spigler, R., "The Kuramoto Model: A Simple Paradigm For Synchronization Phenomena", Reviews of Modern Physics, Volume 77, No. 1, pp. 137, January, 2005.

[12] Weidlich, W., "Physics and Social Science—The Approach of Synergetics", Physics Reports, Volume 204, No. 1, pp. 1-163, May, 1991.

[13] Liang V.C., and Ng, V.T., "Modeling of Collective Synchronous Behavior on Social Media", IEEE 12th International Conference on Data Mining Workshops, pp. 945 - 952,2012.

[14] Haken, H., "SYNERGETICS-An Introduction: Nonequilibrium Phase Transition and Self-Organization in Physics", Chemistry and Biology, 1978.

[15] Caton, S., Hall, M., and Weinhardt, C., "How do Politicians Use Facebook? An Applied Social Observatory", Big Data & Society, Volume 2, No. 2, 2015.

[16] Oliveira, D.J.S., Bermejo, P.H.D.S., and Santos, P.A.D., "Can Social Media Reveal the Preferences of Voters? A Comparison between Sentiment Analysis and Traditional Opinion Polls", Journal of Information Technology & Politics, Volume 14, No. 1, pp. 34-45, 2017.

[17] Carvalho, C.M., Nagano, H., and Barros, A.K., "A Comparative Study for Sentiment Analysis on Election", Proceedings of 11th Brazilian Symposium in Information and Human Language Technology, pp. 103-111, Uberlandia, MG, 2017.

[18] Medhat, W., Hassan, A., and Korashy, H., "Sentiment Analysis Algorithms and Applications: A Survey", Ain Shams Engineering Journal, Volume 5, No. 4, pp. 1093-1113, December, 2014.

[19] Indiran, M., "A Survey on Sentiment Analysis on Social Network Data", International Journal of Scientific Research in Computer Science, Engineering and Information Technology, Volume 2, No. 2, 2017.

[20] Dattu, B.S., and Gore, D.V., "A Survey on Sentiment Analysis on Twitter Data", International Journal of Computer Science and Information Technologies, Volume 6, No. 6, 2016.

[21] Nagy, A., and Stamberger, J., "Crowd Sentiment Detection During Disasters and Crises", Proceedings of 9th International ISCRAM Conference, pp. 1-9, Vancouver, Canada, 2012.

[22] Shamma, D.A., Yew, J., Kennedy, L., and Churchill, E.F., "Viral Actions: Predicting Video View Counts Using Synchronous Sharing Behaviors", 5th International AAAI Conference on Weblogs and Social Media, 2011.

[23] Gao, W., Kaji, N., Yoshinaga, N., and Kitsuregawa, M., "Collective Sentiment Classification Based on User Leniency and Product Popularity", Journal of Natural Language Processing, Volume 21, No. 3, pp. 541-561, 2014.

[24] Bhardwaj, S., Srivastava, S., and Gupta, J.R.P., "Pattern-Similarity-Based Model for Time Series Prediction", Computational Intelligence, Volume 31, No. 1, pp. 106-131, 2015.

[25] Nguyen, L.T., Wu, P., Chan, W., Peng, W., and Zhang, Y., "Predicting Collective Sentiment Dynamics from Time-Series Social Media", Proceedings of 1st International Workshop on Issues of Sentiment Discovery and Opinion Mining, Beijing, China, 2012.

[26] Chis, T., and Harrison, P.G., "Modeling Multi-User Behaviour in Social Networks", Modelling, Analysis & Simulation of Computer and Telecommunication Systems, Paris, France, 2014.

[27] Esmael, B., Arnaout, A., Fruhwirth, R.K., and Thonhauser, G., "Improving Time Series Classification Using Hidden Markov Models", 12th International Conference on Hybrid Intelligent Systems, Pune, India, 2012.

[28] Nguyen, N., "Hidden Markov Model for Stock Trading", International Journal of Financial Studies, Volume 6, No. 2, p. 36, 2018.

[29] Bobulski, J., "Face Recognition Method with Two-Dimensional HMM", Proceedings of 9th International Conference on Computer Recognition Systems, 2015.

[30] Bobulski, J., "Comparison of the Effectiveness of 1D and 2D Hmm in the Pattern Recognition", Image Processing & Communications, Volume 19, No. 1, pp. 5-11, 2014.

[31] Dennemont, Y., Bouyer, G., Otmane, S., and Mallem, M., "A Discrete Hidden Markov Models Recognition Module for Temporal Series: Application to Real-Time 3D Hand Gestures", 3rd International Conference on Image Processing Theory, Tools and Applications, Istanbul, Turkey, 2012.

[32] Ghassempour, S., Girosi, F., and Maeder, A., "Clustering Multivariate Time Series Using Hidden Markov Models", International Journal of Environmental Research and Public health, Volume 11, No. 3, pp. 2741-2763, 2014.

[33] Coviello, E., Chan, A.B., and Lanckriet, G.R., "Clustering Hidden Markov Models with Variational HEM", The Journal of Machine Learning Research, Volume 15, No. 1, pp. 697-747, 2014.

[34] Li, K., and Fu, Y., "Prediction of Human Activity by Discovering Temporal Sequence Patterns", IEEE Transactions on Pattern Analysis and Machine Intelligence, Volume 36, No. 8, pp. 1644 - 1657, January, 2014.

[35] Mikaeil, A.M., Guo, B., Bai, X., and Wang, Z., "Primary User Channel State Prediction Based on Time Series and Hidden Markov Model", 2nd International Conference on Systems and Informatics, Shanghai, China, 2014.

[36] Baum, L.E., Petrie, T., Soules, G., and Weiss, N., "A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains", The Annals of Mathematical Statistics, Volume 41, No. 1, pp. 164-171, February, 1970.