

Predicting user behavior using data profiling and hidden Markov model

Bahaa Eddine Elbaghazaoui, Mohamed Amnai, Youssef Fakhri

Laboratory of Computer Sciences, Faculty of Sciences, Ibn Tofail University, Kenitra, Morocco

Article Info

Article history:

Received Oct 16, 2022

Revised Jan 19, 2023

Accepted Mar 9, 2023

Keywords:

Data profiling

Hidden Markov model

Machine learning

User behavior

User profiling

ABSTRACT

Mental health disorders affect many aspects of patient's lives, including emotions, cognition, and especially behaviors. E-health technology helps to collect information wealth in a non-invasive manner, which represents a promising opportunity to construct health behavior markers. Combining such user behavior data can provide a more comprehensive and contextual view than questionnaire data. Due to behavioral data, we can train machine learning models to understand the data pattern and also use prediction algorithms to know the next state of a person's behavior. The remaining challenges for this issue are how to apply mathematical formulations to textual datasets and find metadata that aids to identify the person's life pattern and also predict the next state of his comportment. The main idea of this work is to use a hidden Markov model (HMM) to predict user behavior from social media applications by analyzing and detecting states and symbols from the user behavior dataset. To achieve this goal, we need to analyze and detect the states and symbols from the user behavior dataset, then convert the textual data to mathematical and numerical matrices. Finally, apply the HMM model to predict the hidden user behavior states. We tested our program and identified that the log-likelihood was higher and better when the model fits the data. In any case, the results of the study indicated that the program was suitable for the purpose and yielded valuable data.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Bahaa Eddine Elbaghazaoui

Laboratory of Computer Sciences, Faculty of Sciences, Ibn Tofail University

Kenitra, Morocco

Email: bahaaeddine.elbaghazaoui@uit.ac.ma

1. INTRODUCTION

Personal data is getting more difficult to obtain [1], and the most recent legislation allows users to connect with a service without providing personal information. Furthermore, a trustworthy personal profile must be constructed utilizing a variety of attributes that are frequently hidden and can only be deduced using predictive models. Implicit feedback, on the other hand, is easy to collect but highly rare. Although strategies based on factorization of the user-item matrix [2] are simple to implement (given the ability to use parallelized algorithms), the gap between the number of ratings and products is usually too large to allow reliable estimation.

Various probabilistic methods have been devised by researchers to understand user behavior patterns in social data [3]. These models often incorporate a hidden parameter, representing a user's interest in a particular topic, estimated based on their interactions with messages, views, and other related content. The user data, such as their frequency of visits, number of messages inspected and received, play a crucial role in recommending services through web applications. Therefore, it is imperative to consider these factors while describing user behavior [4].

Thoughts are sometimes associated with user behavior [5]. We can use innate behavioral information to better comprehend a person's psychological and health behavior. It can also provide a method for predicting future states such as mood, which has a wide range of applications in psychology [6], medicine, and other economic transactions [7], where user mood is a major factor in decision-making.

Online user behavior is creating an unprecedented amount of data that is implicitly public. First, the amount of data is expanding as users spend more time on media applications and engage more through postings, comments, and other means [8]. Second, users almost always contribute information. However, what is less clear is that the capacity to swiftly collect, combine, and analyze that information can provide extremely detailed information about the user. Individual pieces of information are harmless, but when joined with others, they might lead to unexpected revelations.

Psychological profiling is a technique used by psychologists and law enforcement agencies to understand the behavior and motivation of criminals or suspects [9]. Behavioral profiling, on the other hand, involves the use of machine learning and sophisticated analytics to generate profiles of user behavior in various fields such as marketing, healthcare, and finance. The use of behavioral profiling can help organizations better understand and anticipate user needs. However, it is important to balance the benefits of behavioral profiling with privacy concerns and ethical considerations.

The term "prediction" is on everyone's lips. In the corporate sector, predictive analytics is exploding, and every company wants to know what their customers will do next. Predictive analysis can deliver great business value even when individual predictions are not always accurate [10].

The majority of prediction analyses are based on typical machine learning (ML) models such as linear regression, random forest (RF), and natural language processing (NLP). The difficulty with these models is that they use the entire dataset as input and provide a single column as a result [11]. The goal is to figure out the data pattern. Regrettably, the outcomes of those models differ from one database to the next. This article aims to predict user behavior using data profiling and the hidden Markov model (HMM) technique [12]. However, HMM allows profiling the most likely sequence of states. This paper will begin by profiling the dataset to better comprehend the data's worth and identify common user behavior states. Then, for easy application of the HMM model, we need to convert all textual data to a mathematical matrix.

The primary goal of our paper is to identify the users' behavioral directions, as well as the states of their behavior and the necessary actions that they must perform. The HMM model is one of the most applicable models for dealing with this problem. Indeed, after profiling the data, we can extract information from our database (min, max, repetition of an activity, ...). Using this metadata, we can then easily apply the HMM model. However, due to the Daylio application, we collected a user lifelog dataset that contains every action this user took in time as well as his mood at that time. After cleaning and analyzing the dataset, we apply the HMM model and predict the moods of the user based on his activities.

This paper outlines its structure, which aims to summarize previously discussed ideas. It starts by discussing pertinent attempts and significant challenges related to its approach in the first section. The second section covers data profiling and the HMM method. The paper's primary contribution is detailed in the third section, followed by a discussion of the implementation strategy. Lastly, the study finishes with a thorough discussion of the results and recommendations for potential future research.

2. BACKGROUND

Lifelogging is the practice of digitally recording a user's everyday activities for a variety of objectives and at various levels of detail [13]. Such data can be saved to track daily activities and improve the human experience. Lifelogging is already being used in a variety of settings. It has, for example, been used to recall human memories [14], anticipate and diagnose physical health issues [15], detect chronic diseases [16], create themed and digitized diaries, record, identify, and review everyday activities, and gather and analyze data on aged health. Researchers in both the lifelogging and physical health fields have shown great interest in studying physical health due to its potential to enhance the quality of life for individuals.

Self-revelation behavior plays an essential role in various social media networks, affecting self-presentation and social desirability. This conduct caters to people's demands and influences their actions and attitudes [17]. Recently, there has been an increasing interest in exploring the role of emotions in self-disclosure behavior [18]. By evaluating online encounters that can be shared through an emotional perspective, a more comprehensive understanding of individuals' capacity to convey, comprehend, manage, and exchange their self-reports to gain social recognition is achievable [19]. According to present studies, online communication facilitates the sharing of opinions, perceptions, emotions, and experiences more than in-person interactions [20].

Our preliminary research findings on comprehending consumers psychologically and tracking their mental health in four psychological categories using life logs were published in Dang-Nguyen *et al.* [21]. These categories consist of measuring the big five personality "BIG5" traits [22], predicting mood and sleep

quality [23], and detecting music type and mood [24]. We concluded that using lifelog data, we can psychologically analyze and model the person [25].

According to past psychological research, various elements influence a person's behavior. Environmental factors [26], exercise and physical activities [27], [28], weather and air pollution [29], [30], sleep duration and quality, working hours, heart rate, blood pressure, and an individual's personality, particularly in terms of extraversion and neuroticism [31], are some of these factors. Temperature, wind speed, sunshine, precipitation, air pressure, and photoperiod are all factors that affect mood, according to one study by Denissen *et al.* [29]. Many studies on the relationship between body measurements and mood have shown that biometrics are mood markers and are influenced by mood [32], [33].

For researchers that are all heading in the same direction, some of them tried to anticipate and identify moods by gathering information from trial users' profiles, such as social participation, gender, linguistic style, and a variety of psychological data [34]. According to this study, user behaviors and postings can be used as behavioral clues to characterize Twitter users' tweets as positive, negative, or neutral [30]. Roshanaei *et al.* [35] the same group of researchers looked at predicting user emotions based on their mobile phone habits. To collect valuable and significant amounts of user data, they created an Android app that captures user emotions as well as certain physical data about their lives, such as activities they engage in, places they visit, and apps they are currently using. They offer correlations between user moods and these characteristics in their work, as well as create algorithms to predict user moods with promising accuracy.

ML involves several internal states that are challenging to identify or observe. An alternative approach is to infer these states from observable external factors [36]. This is precisely what HMM does. For instance, in speech recognition, we listen to a spoken utterance (the observable) to deduce the underlying text (the internal state representing the speech).

Our method would be to apply a mathematical approach using historical lifelog data to forecast future user behavior. We can forecast the next and hidden states of user behavior using the HMM. This strategy will be realized when we can convert all of our textual data into mathematical form.

3. PROBLEMATIC ISSUE

Several researchers have focused on collecting metadata from web applications to gain a comprehensive understanding of user behavior, which is commonly referred to as data profiling [37]. Studying user behavior can help to enhance the caliber of different services and goods [38] and offer them in a way that satisfies consumers' needs. To effectively market their products and services, large corporations are constantly trying to predict and comprehend their clients behavior. However, determining users' behavioral patterns, their states, and the crucial actions they must take [39] can be challenging. According to our research, the HMM model is the most suitable method for addressing this issue. We can extract information from our database after profiling the data (such as minimum, maximum, and activity frequency), and then use this metadata to apply the HMM model easily.

4. PRELIMINARIES

This section will focus on two methods that will be used for the project: data profiling application and HMM approach. The data profiling application will be utilized to extract meaningful insights from the given dataset. On the other hand, the HMM approach will be used to build a model to predict the next possible event based on the previous sequence of events.

4.1. Data profiling

Profiling is a procedure that involves gathering data from various data sources (databases and files) and compiling statistics and information on such data [40]. As a result, the data analysis is quite close [37]. The process of data profiling is useful in various scenarios where maintaining data quality is crucial. It can help with data warehousing, business intelligence, and linked data by assisting in the identification of potential problems and the implementation of necessary fixes. Moreover, data profiling is crucial for data migration and conversion since it reveals data quality problems that could be overlooked during translation or adaption.

It may come as a surprise that we engage in more data profiling while cleaning and preparing the data than during the preview stage. During this stage, we often handle, clean, remove, or repair various data elements such as NULL values, errors, missing values, noise, or unexpected data artifacts. Some activities necessitate expertise in a particular field or domain. For example, converting an attribute to a specific physical unit, generating a new explanatory variable by calculating the ratio of two specific attributes, or converting an IP address to a geographic location [41].

Accurate and effective data science models rely on a comprehensive understanding of the data being analyzed [42]. Therefore, data profiling is considered the most reliable and efficient way to “get to know the data” for analytics and machine learning initiatives [43]. Data profiling involves examining and evaluating data from multiple sources to identify patterns, quality, completeness, and consistency of the data. This process enables data scientists to gain insights into the data and determine the necessary steps required to clean, transform, and prepare the data for modeling.

4.2. HMM

The HMM is a modified version of the Markov chain, which is a mathematical model that provides information about the likelihood of sets of random variables or states, such as words, tags, or symbols representing various things such as the weather. The Markov chain assumes that the current state is the only one that matters when predicting future events in a sequence, and that previous states have no influence on the future state. This is like to forecasting the weather for tomorrow without considering the weather from yesterday.

$$P(q_i = a | q_1 \dots q_{i-1}) = P(q_i = a | q_{i-1}) \quad (1)$$

HMM provides a more probabilistic approach to modeling time-series data. Unlike traditional methods, which are based around mathematical equations, HMM models a set of discrete states and transitions between them to capture the underlying patterns in the data. This approach allows for more accurate and robust predictions, as it can capture complex relationships and patterns between the data points. In addition, the use of stochastic models provides a means to model uncertainty and variability in the data, which can be beneficial in many contexts. Finally, HMM models are typically more computationally efficient than traditional methods, as they require fewer parameters and can be scaled up or down to adjust the complexity of the model.

Consider a list of state variables, denoted as q_i , can be modeled using a Markov model. This model is based on the assumption that the future probability of the sequence depends only on the present state, and not on the past. Thus, the Markov model simplifies the prediction of future states by disregarding the influence of the past states.

The HMM is a probabilistic model that uses observable data to infer unobserved information [44]. It is a statistical modeling technique that's utilized in speech recognition, handwriting recognition, as well as other applications [45]. In many circumstances, what is observed is not reality, and the HMM is one method for recovering the hidden truth. However, it is not magic for everything, and to utilize it, one must meet certain assumptions, which are the foundations of the HMM.

The focus is on the observations sequence rather than the sequence of states in HMM, a variation of Markov models where the states generating observable data are not explicitly observable [46]. Given the value of the hidden variable at moment $t - 1$, the hidden variable's value at time t dictates both the value of the observable variable at time t and the conditional probability distribution of the hidden variable at instant t . As a result, HMMs are particularly useful for imitating situations when a system can only be partially seen.

The HMM assumes a discrete state space for the hidden variables and a continuous Gaussian distribution for the observations. The HMM has two parameters: the probabilities of state transitions and the probabilities of output. Given a hidden state at time $t - 1$, the state transition probabilities determine the selection of the hidden state at time t . This yields one of N possible values for a set of hidden states modeled by a categorical distribution. Figure 1 depicts the HMM process, where the current state and transition probability matrix A determine the Markov process, which is hidden beneath the dashed line.

The higher-level Markov model can be transformed into a hidden Markov model, as depicted in Figure 2, for a better understanding. In this model, the state variables are hidden, and only observable outputs, or emissions, are visible. The model assumes that the probability distribution of each output depends only on the current hidden state, which can be inferred using the observable emissions.

Markov chain model is clearly visible: the “start” state, the two “state 1” and “state 2” states, as well as the transitions and the corresponding probability. Two new observations have been added: “A” and “B”. Each state has a given probability of emitting each observation, which we call “probability of emission. This probability could be null, indicating that the state is unable to issue the requested observation. In our example, state 1 has a 50% probability (0.5) of producing a “A” and a 50% chance of producing a “B” whereas state 2 has a 90% chance (0.9) of producing a “A” and a 10% chance (0.1) of producing a “B”. The sum of the state's emissions probabilities must always be equal to one, just as it must be for transitions leaving that state.

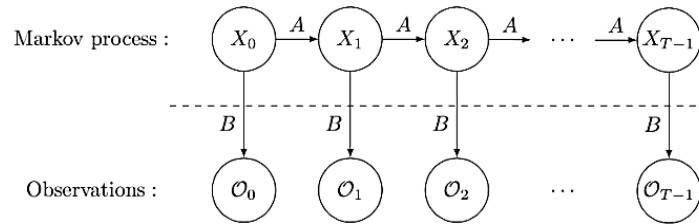


Figure 1. Hidden Markov model

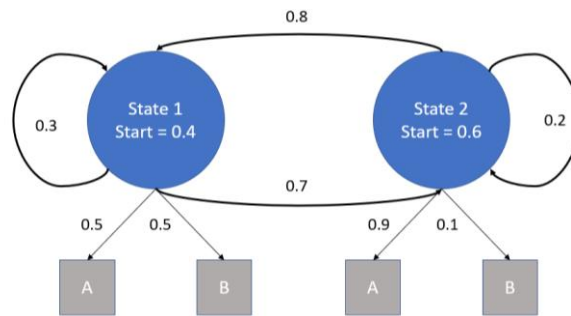


Figure 2. Example of HMM

The basic components of the HMM model is as: i) hidden states, ii) observation symbols the return to the initial hidden state when the initial state has changed, iii) transition to the terminal state probability distribution, iv) distribution of state transition probabilities, and v) distribution of state emission probabilities.

HMM is divided into two sections: hidden and observable. The hidden component is made up of hidden states that cannot be seen directly but are detected by observation symbols that hidden states emit. For example, we do not know what mood our user is in (mood is a hidden state), but we may see their actions (observable symbols) and infer what hidden state he is in from those actions.

There are three common problems that an HMM can be used to solve: i) calculate the probability of a specific sequence using an automated system (using the forward algorithm); ii) finding the most likely state (hidden) that led to the generation of a certain sortie sequence (using the Viterbi algorithm); and iii) given a flight sequence, determine the most likely set of states as well as the likelihood of each state's sorties. The Baum-Welch algorithm, often known as the forward-backward algorithm, is used.

Let's take the on-screen keyboard on a mobile phone as an example [47]. We may occasionally mistype the character next to what we meant to type. The observed data is the character you mistyped, while the unobserved data is the one you wanted to enter in your head. Another example is that owing to random noise, your global positioning system (GPS) measurements (observed) may jump into your actual location (unobserved).

A hidden Markov model can be evaluated using one of two approaches HMM.

- Likelihood of test data: In this strategy, some test data should be kept and the likelihood of the test sequences computed using the forward algorithm.
- Predicting data parts based on other data parts: The application determines whether or not a prediction task is significant. You might be interested in foreseeing the future, for example. You can utilize the forward method to follow the state of the hmm in this scenario.

5. METHOD AND CONTRIBUTION

This study presents our workflow approach, which is illustrated in Figure 3. The initial stage is to gather information about user behavior and activities. In the second step, we use data cleaning to identify relevant and powerful data, followed by data profiling to gather metadata for our HMM model. The final step will be to apply HMM to our metadata.

To make things clearer. In the first phase, we must choose a dataset or an application programming interface (API) that contains information about user behavior as well as its daily activities. These details will make it easier for us to put our strategy into action. In the second section, we must clean up our database of incorrect, incomplete, and null information [48]. As a result, for each state, calculate the average, minimum,

and maximum values to get a broad picture of the data. In the final section, you must calculate the matrix of state transitions and the emission matrix between each state and activity. The initial value of each state can be taken as 1 divided by the number of states. Finally, we'll apply our HMM model.



Figure 3. Approach workflow

In this case, our issue is based on the second problem, as mentioned in the preliminary portion of the HMM section. The Viterbi algorithm's goal is to draw an inference based on some observed data and a trained model [49], [50]. It works by posing the following question: given the trained parameter matrices and data, what are the states that maximize the joint probability? In other words, given the data and the trained model, what is the most likely option? The following algorithm 1 can be used to represent this statement, and the answer is dependent on the facts.

Algorithm 1. Viterbi

```

Input   Initial probabilities vector ( $\pi$ ), Observation sequence ( $O$ ), State transition
         probabilities matrix ( $A$ ), Emission probabilities matrix ( $B$ )
Output  The likely sequence of hidden states ( $Q$ )
1      Begin
2      //Initialization
3      Let  $N$  be the number of states
4      Let  $T$  be the length of observation sequence
5      Initialize a 2D array to store the probabilities of most likely path:
6       $V[1..N, 1..T]$ 
7      //Recursion
8      For each state  $1$  to  $N$ , do
9           $V[state, 1] = \pi[state] * B[state, O[1]]$ 
10         For each time  $2$  to  $T$ , do
11             For each state  $1$  to  $N$ , do
12                  $V[state, t] = \max(V[1..N, t-1] * A[1..N, state] * B[state, O[t]])$ 
13             //Termination
14         Let  $prob$  be the maximum of  $V[1..N, T]$ 
15         Let  $Q[1..T]$  be an array to store the most likely sequence of states
16          $Q[T] = \text{argmax}(V[1..N, T])$ 
17         For  $t=T-1$  to  $1$ , do
18              $Q[t] = \text{argmax}(V[1..N, t] * A[1..N, Q[t+1]])$ 
19         Return  $Q$ 
20     End
  
```

The algorithm 1 is a dynamic programming method that is used to identify the sequence of states with the highest probability in an HMM. The purpose of this algorithm is to maximize the probability of the most likely sequence of states given a sequence of observations. To achieve this, the algorithm initializes a two-dimensional array called V , which is used to store the probabilities of the most likely path. Then, for each state, the algorithm assigns probabilities of the observed sequence. Next, the algorithm iterates over each time step and, for each state, finds the maximum probability of the most likely sequence of states given the previous probabilities. The algorithm then utilizes the array with the maximum probability to identify the most likely set of states. The program then produces the most probable series of states.

The accuracy of an HMM can be determined using Python by calculating the model log-likelihood. The log-likelihood is a measure of how well the model fits the data. To calculate the log-likelihood of an HMM, you need to first fit the model to the data using the `fit()` method. This will create an instance of the HMM with its estimated parameters. Then, you can use the `score()` method to compute the log-likelihood. The better the model fits the data, the higher the log-likelihood.

6. IMPLEMENTATION AND RESULTS

Based on the workflow that we provided in our contribution. The steps for our implementation will be as follows. In the first instance, we need to gather a set of user-relevant data from a donation database or an API. Following whatever search one conducts, one discovers a database collected by the application Daylio that is tailored to our needs. The dataset may be found on the Kaggle website [51]. The dataset is Abid Ali Awan's lifelogs with goals, and it comprises full data about time, emotions, and activities that affect mood from 03/02/2018 to 16/04/2021 as represented in Table 1.

The dataset has 940 data rows. It has many columns as shown in the previous Table 1, but the last two columns (activities and mood) are the ones that interest us in our approach. The user "Abid Ali Awan" frequently alternates between five moods (good, normal, amazing, awful, and bad), and there are numerous activities to choose from, such as reading, learning, and praying...).

Following the data quality step, we applied a Python script to clean the dataset. However, we take the columns that we are interested in and also delete the rows that have null or empty values. After that, we started using data profiling. We used the pandas-profiling library to get a description of metadata such as min, max, and the most frequent values, as seen in Table 2. Data profiling in our work is based on extracting metadata that will help us predict emotional states based on activities. To accomplish so, we must define the transition matrix between states (between moods) as well as the transmission matrix between states and symbols (moods and activities).

It is time to talk about transition matrixes. We use the term "transition" from one specific state to another. In fact, a transition between "Good" and "Bad" is a day when the user is "Good" and the next day is "Bad". The transition value between these two states is the value of transition between these two states divided by the total number of transitions. Table 3 shows the values of the transition matrix in our dataset.

Now, we must generate the emission matrix, which contains the probabilities for each state of emitting each of the potential observations. In fact, if we choose the state "Good" and the symbol "designing", the emission value is the number of days that the user is "Good" as well as performing the activity "designing" divided by the number of days that the user is "Good". Following the extraction of all the necessary data. For our preferred language "Python", we used the hmmlearn library with Gaussian mixture emissions. The hmmlearn is a collection of techniques for unsupervised learning and inferring HMM.

Table 1. Abid Ali Awan's lifelog in the Daylio app

Index	Full date	Date	Weekday	Time	Activities	Mood
0	16/04/2021	Apr 16	Friday	8:00 pm	Reading, Art, Prayer...	Good
1	15/04/2021	Apr 15	Thursday	2:37 am	Reading, Learning, Art...	Good
2	14/04/2021	Apr 14	Wednesday	2:39 am	Reading, Learning, Prayer...	Normal
...	...	...	...	...	...	...
938	03/02/2018	Feb 03	Saturday	7:52 pm	Write, Dota 2, Streaming ...	Awful
939	03/02/2018	Feb 03	Saturday	3:12 pm	Walk, Meditation, Dota 2...	Normal

Table 2. Profiling Daylio dataset

	Activities	Mood
Count	12,614	940
Unique	58	5
Top	YouTube	Good
Freq	770	487

Table 3. Transition matrix result

	Good	Normal	Amazing	Awful	Bad
Good	0.6057	0.2032	0.127	0.0266	0.0369
Normal	0.4756	0.2540	0.1027	0.0918	0.0756
Amazing	0.4071	0.0838	0.4790	0.0179	0.0119
Awful	0.4117	0.1764	0.0392	0.1960	0.1764
Bad	0.2857	0.3469	0.0816	0.1632	0.1224

7. DISCUSSION

The following experience is used as an application. We give our program the following lists of activities. This set of activities represents an example, if a user does these activities, what states can be obtained according to our program? Indeed, we tried in the first tests, random activities that were different and not repeated, for example:

[‘Write’, ‘Prayer’, ‘Shower’, ‘Dota’, ‘Good’, ‘Streaming’, ‘YouTube’, ‘Research’, ‘Power’]

According to those activities, the program predicts the following hidden states:

[‘Bad’, ‘Awful’, ‘Bad’, ‘Awful’, ‘Amazing’, ‘Good’, ‘Normal’, ‘Amazing’, ‘Good’]

After that, we gave our program a list of activities that it repeated as follows:

[‘Streaming’, ‘Reading’, ‘Streaming’, ‘YouTube’, ‘YouTube’, ‘YouTube’, ‘YouTube’, ‘YouTube’, ‘YouTube’]

Based on those activities, the program predicts the following hidden states.

[‘Good’, ‘Normal’, ‘Bad’, ‘Awful’, ‘Awful’, ‘Awful’, ‘Awful’, ‘Awful’, ‘Awful’]

According to our tests, if we repeat some activities that the user must do, we predict that the mood states will repeat too (i.e., if the user watches YouTube several times, the mood states will be “Awful”). For this, we find that user activities influence the future state of mood. This validates the HMM hypothesis and enhances the results of our program.

The evaluation of our system can also be based on the Likelihood results. The log-likelihood value of a regression model is a measure of how well the model fits the data. A higher log-likelihood value indicates a better fit to the dataset. The log-likelihood method is useful when comparing multiple models. In practice, multiple regression models are often fitted to a dataset, and the model with the highest log-likelihood value is chosen as the best fit.

A degenerate solution will result from fitting a model with 34 free scalar parameters with just 5 data points and a $\log(\text{likelihood}) = 165.2432$. If we take 9 activities from our dataset, we get $\log(\text{likelihood}) = 14.3623$. If we take 100 activities from our dataset, we get $\log(\text{likelihood}) = -238.5843$.

The results indicate that the model is overfitted when using a small data set of 5 data points with 34 free scalar parameters. The log-likelihood score is much higher than when using larger datasets of 9 and 100 activities. This suggests that the model is better at predicting outcomes when more data points are available. Additionally, the log-likelihood score is much lower when using a dataset of 100 activities as opposed to 9, indicating that the model is more accurate when more data points are used.

8. CONCLUSION AND FUTURE DIRECTION

HMM has the ability to capture the underlying structure of user behavior, allowing for more accurate predictions. This has been used to great success in a variety of different applications, from predicting user behavior on websites to predicting stock market movements. By leveraging the power of HMM, we can better understand user behavior, predict their next state, and ultimately improve user experience. The objective of this project is to combine basic duties. However, storage and data collection, state and symbol profiling from our data, and the extraction of the transition matrix and emission matrix are all steps in the process. Then, using the HMM model, we forecast hidden states based on observable symbols. This method will aid us in comprehending and anticipating customer behavior, and the current product will be tailored to meet their needs.

In our future directions, we will explore the potential of combining HMMs with machine learning techniques to create more efficient and powerful solutions. We will investigate the use of HMMs to improve the accuracy of machine learning algorithms, as well as explore their potential to enable more intelligent and dynamic machine learning models. We will also investigate the potential of HMMs to act as an intermediary between machine learning algorithms, potentially allowing for more efficient and interactive learning systems.

REFERENCES

- [1] M. Naeem *et al.*, “Trends and future perspective challenges in big data,” in *Smart Innovation, Systems and Technologies*, vol. 253, Springer Singapore, 2022, pp. 309–325, doi: 10.1007/978-981-16-5036-9_30.
- [2] T.-H. Pham *et al.*, “A novel machine learning framework for automated detection of arrhythmias in ECG segments,” *Journal of Ambient Intelligence and Humanized Computing*, vol. 12, no. 11, pp. 10145–10162, Nov. 2021, doi: 10.1007/s12652-020-02779-1.
- [3] J.-H. Kang, K. Lerman, and L. Getoor, “LA-LDA: A limited attention topic model for social recommendation,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 7812, Springer Berlin Heidelberg, 2013, pp. 211–220, doi: 10.1007/978-3-642-37210-0_23.
- [4] H. Hao, F. Feng, W. Shao, and W. Huang, “Understanding the influence of contextual factors and individual social capital on American public mask wearing in response to COVID-19,” *Health Place*, vol. 68, 2021, doi: 10.1016/j.healthplace.2021.102537.

Predicting user behavior using data profiling and hidden Markov model (Bahaa Eddine Elbaghazaoui)

- [5] D. A. Reed, D. B. Gannon, and J. R. Larus, "Imagining the future: thoughts on computing," *Computer*, vol. 45, no. 1, pp. 25–30, Jan. 2012, doi: 10.1109/MC.2011.327.
- [6] R. Buehler, C. McFarland, V. Spyropoulos, and K. C. H. Lam, "Motivated prediction of future feelings: effects of negative mood and mood orientation on affective forecasts," *Personality and Social Psychology Bulletin*, vol. 33, no. 9, pp. 1265–1278, Sep. 2007, doi: 10.1177/0146167207303014.
- [7] J. Bollen, H. Mao, and X. Zeng, "Twitter mood predicts the stock market," *Journal of Computational Science*, vol. 2, no. 1, pp. 1–8, Mar. 2011, doi: 10.1016/j.jocs.2010.12.007.
- [8] S. Khalid, "Motion-based behaviour learning, profiling and classification in the presence of anomalies," *Pattern Recognition*, vol. 43, no. 1, pp. 173–186, Jan. 2010, doi: 10.1016/j.patcog.2009.04.025.
- [9] R. N. Turco, "Psychological profiling," *International Journal of Offender Therapy and Comparative Criminology*, vol. 34, no. 2, pp. 147–154, Sep. 1990, doi: 10.1177/0306624X9003400207.
- [10] M. F. Zanuy, "Speaker recognition by means of a combination of linear and nonlinear predictive models," in *6th European Conference on Speech Communication and Technology (Eurospeech 1999)*, 1999, pp. 763–766, doi: 10.21437/Eurospeech.1999-185.
- [11] P. Petousis and V. Stylianou, "A big data COVID-19 literature pattern discovery using NLP," *BioRxiv*, Jun. 2022, doi: 10.1101/2022.06.01.494451.
- [12] S. Li and Y. Bai, "Deep learning and improved HMM training algorithm and its analysis in facial expression recognition of sports athletes," *Computational Intelligence and Neuroscience*, vol. 2022, pp. 1–12, Jan. 2022, doi: 10.1155/2022/1027735.
- [13] S. Ali, S. Khuroo, A. Khan, and H. Khan, "Smartphone-based lifelogging: toward realization of personal big data," in *EAI/Springer Innovations in Communication and Computing*, Springer International Publishing, 2022, pp. 249–309, doi: 10.1007/978-3-030-75123-4_12.
- [14] J. E. Kragel and J. L. Voss, "Looking for the neural basis of memory," *Trends in Cognitive Sciences*, vol. 26, no. 1, pp. 53–65, Jan. 2022, doi: 10.1016/j.tics.2021.10.010.
- [15] J. Holt-Lunstad and A. Steptoe, "Social isolation: An underappreciated determinant of physical health," *Current Opinion in Psychology*, vol. 43, pp. 232–237, Feb. 2022, doi: 10.1016/j.copsyc.2021.07.012.
- [16] V. Singh, V. K. Asari, and R. Rajasekaran, "A deep neural network for early detection and prediction of chronic kidney disease," *Diagnostics*, vol. 12, no. 1, Jan. 2022, doi: 10.3390/diagnostics12010116.
- [17] N. Andalibi, "Disclosure, privacy, and stigma on social media: Examining non-disclosure of distressing experiences," *ACM Transactions on Computer-Human Interaction*, vol. 27, no. 3, pp. 1–43, Jun. 2020, doi: 10.1145/3386600.
- [18] M. Burke, J. Cheng, and B. de Gant, "Social comparison and Facebook: feedback, positivity, and opportunities for comparison," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, Apr. 2020, pp. 1–13, doi: 10.1145/3313831.3376482.
- [19] O. L. Haimson and T. C. Veinot, "Coming out to doctors, coming out to 'everyone': understanding the average sequence of transgender identity disclosures using social media data," *Transgender Health*, vol. 5, no. 3, pp. 158–165, Sep. 2020, doi: 10.1089/trgh.2019.0045.
- [20] K. L. Hinderliter, S. R. Puhl, and L. A. Skinta, "Mental health disparities among transgender and gender-diverse adults: A review of the literature," *LGBT Health*, vol. 8, no. 6, pp. 341–353, 2021, doi: 10.1089/lgbt.2020.0353.
- [21] D.-T. Dang-Nguyen, L. Zhou, R. Gupta, M. Riegler, and C. Gurrin, "Building a disclosed lifelog dataset: Challenges, principles and processes," in *Proceedings of the 15th International Workshop on Content-Based Multimedia Indexing*, Jun. 2017, pp. 1–6, doi: 10.1145/3095713.3095736.
- [22] C. MacCann, Y. Jiang, L. E. R. Brown, K. S. Double, M. Bucich, and A. Minbashian, "Emotional intelligence predicts academic performance: a meta-analysis," *Psychological Bulletin*, vol. 146, no. 2, pp. 150–186, Feb. 2020, doi: 10.1037/bul0000219.
- [23] L. Wang, S. Huang, L. Huangfu, B. Liu, and X. Zhang, "Multi-label out-of-distribution detection via exploiting sparsity and co-occurrence of labels," *Image and Vision Computing*, vol. 126, p. 104548, 2022, doi: 10.1016/j.imavis.2022.104548.
- [24] A. Palazzi, B. Wagner Fritzen, and G. Gauer, "Music-induced emotion effects on decision-making," *Psychology of Music*, vol. 47, no. 5, pp. 621–643, Sep. 2019, doi: 10.1177/0305735618779224.
- [25] H. Tong et al., "Music mood classification based on lifelog," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11168, Springer International Publishing, 2018, pp. 55–66, doi: 10.1007/978-3-030-01012-6_5.
- [26] R. Küller, S. Ballal, T. Laike, B. Mikellides, and G. Tonello, "The impact of light and colour on psychological mood: A cross-cultural study of indoor work environments," *Ergonomics*, vol. 49, no. 14, pp. 1496–1507, Nov. 2006, doi: 10.1080/00140130600858142.
- [27] K. Scrivener, C. Sherrington, and K. Schurr, "Amount of exercise in the first week after stroke predicts walking speed and unassisted walking," *Neurorehabilitation and Neural Repair*, vol. 26, no. 8, pp. 932–938, 2012, doi: 10.1177/1545968312439628.
- [28] R. R. Yeung, "The acute effects of exercise on mood state," *Journal of Psychosomatic Research*, vol. 40, no. 2, pp. 123–141, Feb. 1996, doi: 10.1016/0022-3999(95)00554-4.
- [29] J. J. A. Denissen, L. Butalid, L. Penke, and M. A. G. van Aken, "The effects of weather on daily mood: A multilevel approach," *Emotion*, vol. 8, no. 5, pp. 662–667, 2008, doi: 10.1037/a0013497.
- [30] M. Roshanaei, R. Han, and S. Mishra, "Having fun?: Personalized activity-based mood prediction in social media," in *Lecture Notes in Social Networks*, 2017, pp. 1–18, doi: 10.1007/978-3-319-51049-1_1.
- [31] J. M. L. Poon, "Mood: a review of its antecedents and consequences," *International Journal of Organization Theory and Behavior*, vol. 4, no. 3/4, pp. 357–388, Mar. 2001, doi: 10.1108/IJOTB-04-03-04-2001-B008.
- [32] J. Warland, "Low blood pressure, low mood?," *BMC Pregnancy and Childbirth*, vol. 12, 2012, doi: 10.1186/1471-2393-12-S1-A9.
- [33] J. Bakker, M. Pechenizkiy, and N. Sidorova, "What's your current stress level? detection of stress patterns from GSR sensor data," in *2011 IEEE 11th International Conference on Data Mining Workshops*, 2011, pp. 573–580, doi: 10.1109/ICDMW.2011.178.
- [34] J. Burger et al., "Reporting standards for psychological network analyses in cross-sectional data," *Psychological Methods*, Apr. 2022, doi: 10.1037/met0000471.
- [35] M. Roshanaei, R. Han, and S. Mishra, "EmotionSensing: Predicting mobile user emotion," in *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2017*, Jul. 2017, pp. 325–330, doi: 10.1145/3110025.3110127.
- [36] D. R. K. Srivastava and D. Pandey, "Speech recognition using HMM and soft computing," *Materials Today: Proceedings*, 2022, vol. 51, pp. 1878–1883, doi: 10.1016/j.matpr.2021.10.097.
- [37] B. E. Elbaghazoui, M. Amnai, and A. Semmouri, "Data profiling over big data area," in *Advances in Intelligent Systems and Computing*, pp. 111–123, 2021, doi: 10.1007/978-3-030-72588-4_8.

- [38] M. Giannakis, R. Dubey, S. Yan, K. Spanaki, and T. Papadopoulos, "Social media and sensemaking patterns in new product development: demystifying the customer sentiment," *Annals of Operations Research*, vol. 308, no. 1–2, pp. 145–175, Jan. 2022, doi: 10.1007/s10479-020-03775-6.
- [39] A. Vitetta, "Risk reduction in transport system in emergency conditions: a framework for network design problems," in *WIT Transactions on the Built Environment*, 2021, vol. 206, pp. 267–274, doi: 10.2495/SAFE210221.
- [40] L. Ehrlinger and W. Wöb, "A survey of data quality measurement and monitoring tools," *Frontiers in Big Data*, vol. 5, Mar. 2022, doi: 10.3389/fdata.2022.850611.
- [41] C. Gokhale, "Network analysis of dark web traffic through the geo-location of South African IP address space," in *EAI/Springer Innovations in Communication and Computing*, Springer International Publishing, 2020, pp. 201–219, doi: 10.1007/978-3-030-14718-1_10.
- [42] E. Peer, D. Rothschild, A. Gordon, Z. Evermden, and E. Damer, "Data quality of platforms and panels for online behavioral research," *Behavior Research Methods*, vol. 54, no. 4, pp. 1643–1662, Sep. 2021, doi: 10.3758/s13428-021-01694-3.
- [43] Y. Liu, Y. Li, and J. Wang, "Data profiling for big data analytics," *Journal of Intelligent and Fuzzy Systems*, vol. 38, no. 5, pp. 6081–6092, 2020, doi: 10.3233/JIFS-189726.
- [44] J. H. Jung, "Hand gesture recognition techniques: a review," *Journal of Imaging Science and Technology*, vol. 64, no. 2, pp. 20501–20511, 2020, doi: 10.2352/J.ImagingSci.Technol.2020.64.2.020501.
- [45] Q. Deng and D. Sofker, "A review of HMM-based approaches of driving behaviors recognition and prediction," *IEEE Transactions on Intelligent Vehicles*, vol. 7, no. 1, pp. 21–31, Mar. 2022, doi: 10.1109/TIV.2021.3065933.
- [46] P. J. Schweitzer, "A survey of aggregation-disaggregation in large Markov chains," in *Numerical Solution of Markov Chains*, Boca Raton: CRC Press, 2021, pp. 63–88, doi: 10.1201/9781003210160-4.
- [47] N. Kimura *et al.*, "SilentSpeller: Towards mobile, hands-free, silent speech text entry using electropalatography," in *CHI Conference on Human Factors in Computing Systems*, Apr. 2022, pp. 1–19, doi: 10.1145/3491102.3502015.
- [48] C. Chai, J. Wang, Y. Luo, Z. Niu, and G. Li, "Data management for machine learning: a survey," *IEEE Transactions on Knowledge and Data Engineering*, p. 1, 2022, doi: 10.1109/TKDE.2022.3148237.
- [49] G. Rathee, C. A. Kerrache, and M. A. Ferrag, "A blockchain-based intrusion detection system using viterbi algorithm and indirect trust for IIoT systems," *Journal of Sensor and Actuator Networks*, vol. 11, no. 4, 2022, doi: 10.3390/jsan11040071.
- [50] A. Jandera and T. Skovranek, "Customer behaviour hidden Markov model," *Mathematics*, vol. 10, no. 8, Apr. 2022, doi: 10.3390/math10081230.
- [51] A. A. Awan, "Daylio mood tracker," *Kaggle*, May 18, 2021. Accessed Oct. 16, 2022. [Online]. Available: <https://www.kaggle.com/datasets/kingabzpro/daylio-mood-tracker>

BIOGRAPHIES OF AUTHORS



Bahaa Eddine Elbaghazaoui    began his studies with a Mathematical Science Scientific baccalaureate option. In 2013, he passed the preparatory classes that were integrated into the National School of Applied Sciences of Khouribga, and then he progressed into the computer engineering sector, earning an engineering diploma in software engineering in 2019. Bahaa Eddine El. is Ph.D. student, he does research at the Laboratory of Computer Science. Ibn Tofail University in Kenitra, Morocco. He can be contacted at email bahaaeddine.elbaghazaoui@uit.ac.ma and elbaghazaoui.bahaa@gmail.com.



Mohamed Amnai    completed his master's degree in Computers, Electronics, Electrical, and Automation (IEEA) from Molay Ismail University in Errachidia in 2000. Later in 2007, he obtained his master's degree from Ibn Tofail University in Kenitra. In 2011, he received his Ph.D. in Computer Science and Telecommunications from Ibn Tofail University in Kenitra, Morocco. Currently, he serves as an assistant at the National School of Applied Sciences Khouribga of Settat University since March 2014. In 2018, he was appointed as an Associate Professor at the Department of Computer Science and Mathematics, Faculty of Sciences of Kenitra, Ibn Tofail University in Morocco. He is also a member of the Networks and Telecommunications Team at the Kenitra Faculty of Science and a member of the Research Laboratory in Computer Science and Telecommunications (LaRIT). He can be contacted at email mohamed.amnai@uit.ac.ma.



Youssef Fakhri    obtained a Bachelor of Science (B.S.) degree in Electronic Physics from the University Mohammed V's Faculty of Sciences in Rabat, Morocco, in 2001. He pursued his Master's Degree (DESA) in Computer and Telecommunication at the same university, where he completed his Master's Project with the Moroccan ICI Company in 2003. Later in 2007, he received his Ph.D. from the University Mohammed V-Agdal in Rabat, Morocco, in collaboration with the Polytechnic University of Catalonia (UPC) in Spain. Following his academic accomplishments, he was appointed as an Associate Professor at the Ibn Tofail University's Faculty of Sciences of Kenitra in Morocco in 2009, where he teaches in the Department of Computer Science and Mathematics. Additionally, he serves as an Associate Researcher at the Rabat Faculty of Sciences and is currently the Laboratory Head at LaRIT. He can be contacted at email fakhri@uit.ac.ma.