# A Conceptual Framework for Classifying Users' Perception of Data Privacy on Social Media Networks Using β k-means

***Omowumi Hafsat Aliu[1,2,*], Paschal Uchenna Chinedu[3], Israel Uzuazor Siloko[4], Daniel Aliu[3], Samuel Omaji[1], Kingsley Eghonghon Ukhurebor[5]***

[1]Department of Computer Science, Edo State University, Uzairue, Edo State, Nigeria
[2]Department of Computer Science, Auchi Polytechnic, Auchi, Edo State, Nigeria
[3]Department of Computer Engineering, Edo State University, Uzairue, Edo State, Nigeria
[4]Department of Mathematics, Edo State University, Uzairue, Edo State, Nigeria
[5]Department of Physics, Edo State University, Uzairue, Edo State, Nigeria

[*]***Corresponding author: aliu20.omowumi@edouniversity.edu.ng***

**ABSTRACT**

*Communication and information sharing have been changed by the emergence of social media networks (SMNs), it has brought new dynamics to the way people interact and share personal information and random content such as videos, images, sound, and text to convey insights, humor, opinion, gossip, and news on the internet. SMNs give people, companies, and organizations a platform for real-time communication and the promotion of their goods and services. However, the use of SMNs poses threats to users' privacy. The existing research work identifies problems with users' perception of data privacy without offering a plan to address or resolve these issues. In order to categorize users' perception of data privacy on SMNs, this research proposes a system model using a β k-means algorithm to categorize users' perceptions based on data privacy and assess their perceptions in the context of SMNs. A consistency factor β is incorporated in k-means to ensure the consistency of the classified users' perception of data privacy. The consistency factor is used to regular the degree to which users have a consistent view of their privacy perception. This proposed system will provide valuable feedback to both users and developers of SMNs.*

***Keyword: Data Privacy, Machine Learning, Social Media Networks, User Perception, β K-means***

## 1. Introduction:

Today, privacy remains one of the most significant and easily identifiable problems associated with social media networks (SMNs); the consequences of revealing private information online are still unknown to many users (Polakis *et al*., 2010; Saravanakumar & Deepa, 2016; Jain *et al*., 2021). The work of Reynol Junco (2012) discovered that people are reading the agreement policy of social media less frequently every day. These agreement policies are thought to be tiresome, time-consuming, and needless since social media has become increasingly popular. Social media sites have access to a ton of information about us, sometimes much more than we are aware of; therefore, it is important to talk about the considerable risk that comes with using social media. With billion active users online, these platforms draw users with bad intentions in addition to family members and friends who would like to keep in contact (Jain *et al*., 2021; Kurdi

*et al*., 2022; Bocar & Jocson, 2022; Hien Thu Bui, 2022; Ou *et al*., 2022). One of the common risks associated with utilizing social media is cybercrime (Babu & Siddik, 2022; Hire *et al*., 2020). The majority of social networks ask users to provide personal information like their email accounts, birthday dates and addresses, among others and with such information provided, privacy becomes an issue (Kundu *et al*., 2022).

## 2. Related Works

To mitigate privacy issues, Onifade *et al*., (2018) in their studies reveal the extent to which SMN users are aware of the terms of service and confirm the users' perceptions of the information shared on social media. Information can be leaked due to the way social network privacy protections are designed (Conti *et al*., 2011). Similar findings were made by Asif & Mamuna (2012), who discovered that despite the policies being clearly written, people were still not aware of them. The study also revealed that consumers shared their private information with strangers since they were unaware of how their personal data may be used. They come to the conclusion that unintended information sharing is caused by the complexity of privacy settings and a lack of user control. Also, in Wu *et al*., (2010), emphasized that as all of the privacy regulations implicated lack precise terms, data retention may be an issue. This is because data retention brings about privacy risks, vulnerability of database to theft and accidental disclosure. Some people believe that privacy policies are written in unclear words and are too long for users to read. They also think that users should be informed in clear, easily available privacy statements.

As at the time of this research work, available scientific methods for classification of user's perceptions on SMNs are limited. The available literatures (Lin & Liu, 2012; Bur, 2022; Kundu *et al*., 2022) have highlighted the data privacy challenges without providing a framework for the development of a system for classification of users' perceptions of these data privacy on social media networks. In this work, a model based on machine learning to classify users' perceptions of data privacy on SMNs is proposed. The proposed model provides the necessary feedback for enabling them precise decisions.

## 3. Research Methodology

This methodology's goal is to give a thorough description of the methods and strategies used.

### 3.1 The System Model

Figure 1 presents the system model. The system model is divided into three essential phases: data collection, data preprocessing and data modelling. Detail of each phase is discussed in subsequent subsections.

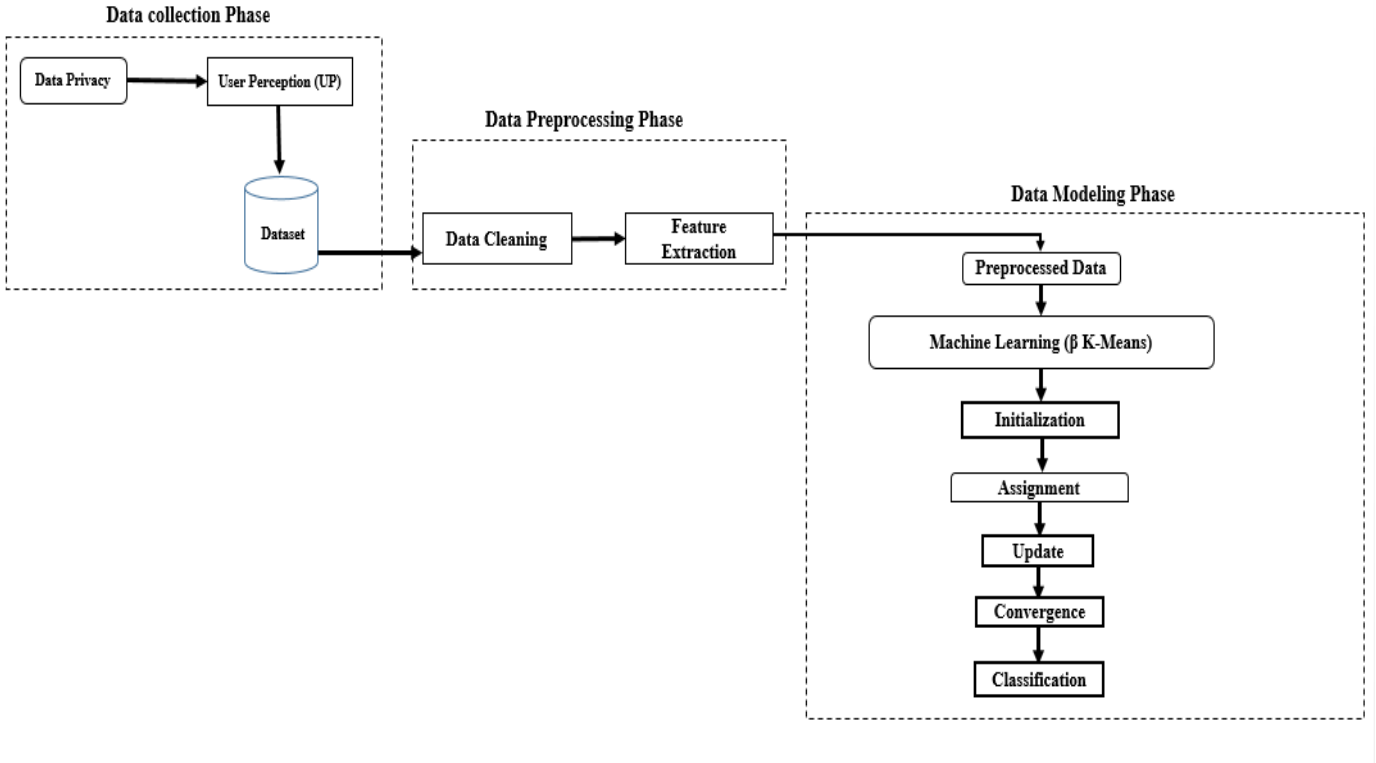


**Figure 1: The system model**

### 3.2 Data Collection phase

In the data collection phase, users' perceptions datasets of data privacy on social media network are collected. Users' perception dataset is a collection of users' perceptions of social media privacy data that is organized in a way that enables efficient storage, retrieval, and management of that data. This dataset is used to perform operations on that data, such as sorting, filtering, and searching. It will also be used to create reports, and other visualizations to help users analyze and interpret the data. A dataset is a collection of data. In the case of tabular data, a dataset relates to one or more database tables, where each row refers to a specific record of the corresponding data set and each column to a specific variable. While databases are an essential tool for managing large amounts of data.

### 3.3 Data Pre-processing Phase:

Data preprocessing is an essential step in machine learning that involves transforming raw data into a form that can be easily understood and analyzed by machine learning algorithms. The main goal of data preprocessing in this research work is to improve the quality of data gather from SMNs users' perception, reduce noise, and prepare the data for machine learning algorithms (i.e., β k-means) to make accurate predictions and classifications. Techniques used in data preprocessing are as follows.

### 3.3.1 Data Cleaning

Data cleaning is the process of identifying and handling missing or erroneous values, duplicates, and outliers in the dataset. The primary goal of data cleaning in this research work is to improve the quality of the SMNs dataset, making it suitable for analysis by machine learning algorithms (β k-means). It is essential to clean the data to remove inconsistencies that could lead to inaccurate models. The data cleanings process includes: identifying missing data, handling missing data, identifying /handling duplicates and handling outliers.

Data cleaning is a crucial step in data preprocessing that ensures the data is suitable for analysis by machine learning algorithms. It involves identifying and handling missing or erroneous values, duplicates, and outliers, and ensuring the data is consistent and in the correct format.

### 3.3.2. Feature Selection

This involves selecting the most relevant features from the dataset of SMNs that are most likely to have a significant impact on the outcome of the machine learning model (β k-means), that can be used for machine learning tasks such as classification, regression, or clustering. The choice of method for feature extraction depends on several factors, including the type of data, the analysis or model being used, and the goals of the analysis or model.

In extracting features in the SMNs, dimensionality reduction technique such as entropy for reducing the number of variables or features in a dataset is considered. Entropy is a measure of the randomness or uncertainty in a dataset or probability distribution. In feature extraction, entropy can be used as a criterion for selecting features that are most informative for a given task. The basic idea is to select features that maximize the amount of information gained about the target variable.

## 4. Data Modelling Phase

The main goal of this research is to apply machine learning techniques to understand how people perceive data privacy. Machine Learning algorithms are perfect for simulating the complex behavior of arbitrary parameters and spotting emerging behavioral trends by utilizing data gathered from various tests. A machine learning (ML) method, which is a subset of AI, employs a variety of precise, probabilistic, and advanced techniques to enable computers to learn from the past and recognize difficult-to-perceive patterns from large, noisy, or complex datasets. With these understanding, machine learning algorithms can be used to classify problems that are qualitative in nature (i.e., not discrete). Although, researchers on other field of studies have used other methods like the fuzzy co-joint analysis (FCA), the mean and standard deviation, classic probability concept to solve questions that are not discrete in nature. Furthermore, Rafaelof & Schroeder (2018) used bagged decision trees/random forests (BDTRF) to investigate machine learning algorithms to model perception of sound and used only one variant on a five Likert scale. For this research

work, β k-means model is better used on multivariate data gather from a seven Likert scale. The classification model is evaluated with an expert survey to compare several iterations of the model, together with an analytical method to draw inferences from key entities of the users' perception data. In addition, this attribute analysis of the method is necessary to determine the accuracy of the guidelines and relationships, which is proffered by the method. The successful completion and evaluation of the main outputs (model) of the research lead to the final stage of theorization of the research outcome.

### 4.1 The β k-means for Social Media Networks

The β k-means is described in this section. Today, most research considers how to initialize the centres for k-means with the intention of maximizing the efficiency of the algorithm. The β k-means is employed in this study to resolve the limitation of k-means clustering algorithm, as the existing k means is sensitive to initial centroid selection. The β k-means allows respondents to individually determine the number of clusters that fits their local data. A clustering method is used to partition elements in a dataset such that similar elements are assigned to the same cluster, while dissimilar elements are assigned to another cluster.

$y = \{y_1, y_2, \dots y_k\}$ $\mathbb{R}^k$ $C$ $Obj\ (k) = Min\left[\frac{1}{K}\sum_{k=1}^{K} || k - C ||^2\right]$ $k$ $k = 2$ $k$ $k$

Given of k-samples in the k-means finds the minimum variance clustering of the dataset into k-clusters with centroids , given the following minimization objective function as expressed in equation 1.

. (1) $N$

To identify the centroid point, we determine the cluster to which the respondents' privacy perception is assigned. Every perception is mapped to the cluster with the closest mean value. Note that as the value of increases, i.e., , a NP hard problem is considered. In a typical k-means algorithm, is selected randomly from the input space of the dataset. The forms the initial centroids of the clusters. Afterwards, all privacy perceptions are assigned to the centroid they are closest to. An average of all users' perceptions in the datasets are used to compute the centroid for each cluster. Then, the entire process of computing the centroids is repeated until a convergence is reached. The β K-means algorithm is an iterative method used to partition user's privacy perception into k-clusters. Here, clusters are derived from observations, where each observation belongs to the cluster with the closest mean value. The procedure for the β K-means algorithm is derived as follows.

$k$

1. is arbitrarily generated from the observations using heuristic $(dc)$ technique.
2. Assign each observation (users' privacy perception) to the cluster that maximized the distance between observation and cluster center (centroid).

3. Find the average of the observation in the cluster to re-compute the centroid.

4. Repeat steps two (2) and three (3) until convergence is reached.

$d_{ik}$ Note that users or da- $k$ ta members can be sim- $i$ $d_{jk}$ ultaneously assigned to $k$ one or more $j$ $\beta$ classes by a certain degree of member- ship. As progressive research, we hope in the future to analyze the degree of membership using the fuzzy logic approach. Thus, we define the degree of consistency of users' perception in equation 2.

$$dc\ (d) = \frac{1}{\sum_{i=1}^{c}\sum_{j=1}^{k}\left(\frac{d_{ik}}{d_{jk}}\right)}\beta \qquad (2)$$

$dc\ (d)$ where is distance from point to present cluster centre and is the distance from point to other clusters and regulates the degree of consistency. Note that β= [0,1].

The consistency of users' privacy implies that users have a consistent view of their privacy perception. Here, the uniqueness of users' responses is measured to ensure that integrity is maintained. Also, consistency of users' privacy perception ensures that the users share the same view of their privacy perception when using the same SMNs. Furthermore, it involves changes that was made on privacy perception and changes made by others. Once there is a privacy perception inconsistency, it means that there is a variation in the same privacy regulations for different SMNs. Equation (2) defines the degree of consistency of users' privacy perception. Note that lies within 0 and 1. If the degree of perception approaches 0, it implies that there is inconsistency of privacy perception. On the other hand, if the degree of consistency approaches 1, it means that there is consistency of users' privacy perception.

## 5. Conclusion

The creation of this model is for classifying user perceptions and a method for drawing relevant inferences from the hidden attributes and relationships within the users' perception data. β k-means classifier proved better in the classification of users' perception of data privacy with regards to SMNs. Since β K-means algorithm considered the best classifier; hence, online privacy experts or researchers must focus attention on the use and application of the β K-means algorithm.

## REFERENCES

Asif, Z. & Mamuna, K. (2012). Users' perceptions on facebook's privacy policies. ARPN Journal of Systems and Software, 2(3), retrieved on 3rd April, 2022

Babu, K. E. K., & Siddik, M. A. B. (2022). Cybercrime in the social media of Bangladesh: an analysis of existing legal frameworks. *International Journal of Electronic Security and Digital Forensics*, *14* (1), 1-18.

Bocar, A. C., & Jocson, G. G. (2022). Understanding the Challenges of Social Media Users: Management Students' Perspectives in Two Asian Countries. Journal of Business, Communication & Technology, 1(1), 24-34.

Bui, H. T. (2022). Exploring and explaining older consumers' behaviour in the boom of social media. International journal of consumer studies, 46(2), 601-620.

Conti, M., Passarella, A., & Pezzoni, F. (2011). A model for the generation of social network graphs. In *2011 IEEE International Symposium on a World of Wireless, Mobile and Multimedia Networks,* 1-6. IEEE.

Hire, N., Patil, P., Nandwalkar, B., Mahale, D., & Patel, J. () Cybercrimes on Social Media using Machine Learning. http://nms.sagepub.com.

Jain, A. K (2010) "Data clustering: 50 years beyond k-means," Pattern Recognition Letters, 31, 651–666.

Junco, R. (2012). The relationship between frequency of Facebook use, participation in Facebook activities, and student engagement. *Computers & education*, *58*(1), 162-171.

Kundu, P., Kundu, P., Mallik, S., Bhowmick, S., Mandal, P., Banerjee, H., & Pal, S. B. (2022). Analysis of Security and Privacy in Social Media Platforms. *American Journal of Electronics & Communication*, *2*(3), 20-29

Kurdi, B., Alshurideh, M., Akour, I., Tariq, E., AlHamad, A., & Alzoubi, H. (2022). The effect of social media influencers' characteristics on consumer intention and attitude toward Keto products purchase intention. International Journal of Data and Network Science, 6(4), 1135-1146.

Lin, S.W. & Liu, Y.C. (2012). The effects of motivations, trust, and privacy concern in social networking. *Serv. Bus. 6*, 411–424.

Onifade O., Olomu M., Ajao B. F., Atoyebi M., & Ilevbare O. (2018): Social Media Users Perception on Privacy Issues in a Nigerian University. Journal of Digital Innovations & Contemp Res. In Sc., Eng & Tech. 6(2), 35-46

Ou, C. C., Chen, K. L., Tseng, W. K., & Lin, Y. Y. (2022). A study on the influence of conformity behaviors, perceived risks, and customer engagement on group buying intention: A case study of community e-commerce platforms. *Sustainability*, *14*(4), 1941.

Polakis, I., Kontaxis, G., Antonatos, S., Gessiou, E., Petsas, T., & Markatos, E. P. (2010). Using social networks to harvest email addresses. In *Proceedings of the 9th Annual ACM Workshop on Privacy in the Electronic Society,* 11-20.

Rafaelof, M., & Schroeder, A. (2018). Investigation of machine learning algorithms to model perception of sound. In *Proceedings of Meetings on Acoustics 175ASA,* Acoustical Society of America 33(1), 050004.

Saravanakumar, K., & Deepa, K. (2016). On privacy and security in social media–a comprehensive study. Procedia Computer Science, 78, 114-119.

Wu, X., Ying, X., Liu, K., & Chen, L. (2010). A survey of privacy-preservation of graphs and social networks. *Managing and mining graph data*, 421-453.