Received: April 20, 2026. Revised: June 4, 2026.

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

Knowledge-sharing Behavior on Facebook Using Theory of Planned Behavior

Yuliazmi[1,4] Diana Purwitasari[3] Cornelius Bagus Purnama Putra[3] Agus Budi Raharjo[3] Safitri Juanita[4] Surya Sumpeno[1,2] Mauridhi Hery Purnomo[1,2]*

[1]Department of Electrical Engineering, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia
[2]Department of Computer Engineering, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia
[3]Department of Informatics, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia
[4]Department of Information System, Universitas Budi Luhur, Jakarta, Indonesia
* Corresponding author's Email: hery@ee.its.ac.id

Abstract: Knowledge-sharing behavior on Facebook gained new relevance during the COVID-19 pandemic, when face-to-face interaction was restricted and much of this activity moved online. This study analyzes such behavior using the Theory of Planned Behavior (TPB), pairing survey-based attributes with observed Facebook activity from 1,841 lengthy posts by 81 users. Twenty-one TPB-based features were built from demographics, social interactions, and post characteristics. Regression confirms that demographics, mutual relationships, social-media use, the COVID-19 context, and post characteristics each predict knowledge sharing; audience response does not. Among six classification models, tree-based classifiers Random Forest and CatBoost gave comparable AUC-ROC of 0.7805 and 0.7801 respectively; the other four ranged from 0.6689 to 0.7657. Ablation pointed to post characteristics and attitudinal factors as the strongest contributors. Intrinsic motivation and content drive who shares, not how others react.

Keywords: Knowledge sharing, Social media, Theory of planned behavior (TPB), Machine learning, Facebook.

1. Introduction

Social media users share knowledge on Facebook through comments, replies, likes, and shares [1, 2]. During the COVID-19 pandemic, face-to-face interaction was restricted, and knowledge-sharing activity increasingly moved to social media (SM) platforms like Facebook [3]. Privacy concerns had previously limited Facebook's role as a formal knowledge repository [4]. A study of publicly posted status updates from consenting respondents found that older adults in particular favor privacy-preserving behavior [5].

Facebook (FB) enables knowledge sharing (KS) across interest groups and mutual friendship networks, while supporting media attachments such as images, videos, and URL links, as well as interactive responses including likes and comments. This platform offers a practical channel for sharing information, particularly when face-to-face interactions are restricted, such as during the COVID-19 pandemic.

Platform availability alone, however, does not explain why some users choose to share while others do not. Prior work has identified ten SM information-sharing motivators, including enjoyment, self-efficacy, learning, altruism, and reciprocity [6]. This study emphasizes three of them. Enjoyment drives sharing as a hobby or leisure activity. Self-efficacy [7] shapes users' motivation for knowledge sharing, specifically in contributing practical COVID-19 safety guidance on SM. Learning leads users to seek and circulate topic updates [8], which may also strengthen knowledge sharing within online communities [9, 10]. Researchers have examined these motivations through various theoretical lenses.

Multiple frameworks have been applied to knowledge sharing: SET (Social Exchange Theory) with SEM (Structural Equation Modeling) in Pakistani software SMEs (Small-Medium Enterprises) [11], SEM-partial least squares (SEM-PLS) in enterprise SM [12], the Theory of Planned Behavior with gratifications theory for COVID-19 FB sharing [13], AI-mediated exchange theory (AI-

MET) in IT multinationals [14], and Strategic Practices for informal SM use [15]. In addition, personality-and motivation-based approaches such as the Big Five personality traits [16], Self-Motivation combined with Affordance Theory [17], and Self-Determination paired with Social Dilemma Theory [18] have also been investigated, reflecting the varied nature of knowledge sharing research. Among these, TPB is particularly suited to SM contexts because its three components map directly onto observable user attributes and posting behavior, rather than relying on self-reported intentions alone.

To understand knowledge-sharing behavior (KSB) on personal FB timelines [13] this study adopts the Theory of Planned Behavior (TPB). TPB posits that behavioral intention is influenced by Attitude (AT), Subjective Norm (SN), and Perceived Behavioral Control (PBC) [19, 20], and this framework has been widely applied in knowledge-sharing research contexts [12].

Unlike prior studies using academic attributes as TPB proxies [19], this study maps 21 observable features from our SM dataset to the TPB model, shown in the red-dashed box in Fig. 1. Demographic attributes and post characteristics represent AT; dominant emotion, sentiment, post type, time, date, and response counts (likes/comments) represent SN; and COVID-19 status, media availability, topics, mutual friendship, and tagged friends represent PBC.

Despite extensive research on knowledge sharing in SM, several limitations remain. First, most existing studies rely on self-reported data, which may not accurately reflect actual user behavior on SM platforms [21]. The discrepancy between reported intentions and observed actions has been widely documented. Second, limited studies have examined the relative contribution of each TPB component in predicting KSB, leaving behavioral mechanisms underexplored [22]. Third, prior work predominantly applies single analytical techniques; although machine learning (ML) methods such as Random Forest, Gradient Boosting, and SVM have shown strong performance in SM classification tasks [23-25], their application to KSB remains limited. To address these gaps, this study integrates survey-based attributes with real FB data and combines regression analysis for hypothesis testing with multiple machine learning classifiers for predictive modeling.

Few studies have combined behavioral theory with ML to identify KSB from observable FB posts and user attributes. This study proposes such a framework. We pair the Theory of Planned Behavior with social-media behavioral features and machine-learning classification. Survey responses and FB activity are mapped into TPB constructs, then analyzed two ways: regression for hypothesis testing, and classification models for prediction. The combination supports theoretical examination of what drives KS and, at the same time, predictive identification from real FB data.

The previous literature review led to the proposal of six hypotheses concerning KSB, which were then tested using regression models:

1. H1: KSB is influenced by demographic factors.
2. H2: Individual interests and mutual friendships influence KSB.
3. H3: The intensity of SM usage influences KSB.
4. H4: KSB is influenced by the COVID-19 pandemic.
5. H5: The responses of others significantly influence KSB.
6. H6: KSB is influenced by the characteristics of posts.

The contributions of this study to identifying KSB from FB posts and user attributes using TPB and ML techniques are summarized as follows:

1. This study presents a systematic operationalization of TPB constructs as observable FB behavioral features, so that analysis draws on actual SM activity rather than self-reported data.
2. A TPB-based feature framework is developed by mapping demographic attributes, SM usage, posting characteristics, and interaction features into Attitude, Subjective Norm, and Perceived Behavioral Control.
3. The proposed framework combines regression analysis and ML classification, enabling both hypothesis testing and predictive modeling of KSB.
4. We compared six ML classifiers, with CatBoost added as a benchmark. Tree-based methods scored highest. Random Forest reached AUC 0.7805 and CatBoost 0.7801; other algorithms ranged from 0.6689 to 0.7657. The consistent performance across all six models indicates that TPB-based feature representation contributes substantially to prediction.
5. A hypothesis-based ablation analysis is introduced to quantify the contribution of each TPB-derived feature group.

The remainder of this paper is organized as follows. Section 2 presents the methodology, including data collection, preprocessing, and analytical methods. Section 3 presents the results, including regression-based hypothesis testing, ML classification, and ablation analysis. Section 4

Received: April 20, 2026. Revised: June 4, 2026.

presents the discussion. Section 5 presents the conclusion.

2. Methodology

This section describes the methodology summarized in Fig. 1. The pipeline has four stages: data collection, preprocessing, feature selection and mapping, and modeling. We collected data via an online survey paired with FB post extraction, then preprocessed the text and manually labeled posts for knowledge-sharing content. Feature selection picks 21 observable indicators which are then mapped to TPB constructs. Modeling uses regression for hypothesis testing and ML for classification. Subsections that follow expand on each stage.

2.1 Data collection

We did not find existing Indonesian FB datasets that include the knowledge-sharing labels and TPB-relevant metadata needed for this analysis. The authors and team built the dataset themselves. We acknowledge the limitations of working with a self-constructed dataset, particularly around validity and generalizability; the preprocessing and annotation steps described below are intended to mitigate these concerns.

Data collection was conducted using a two-stage approach, as illustrated in Fig. 1. In the first stage, an online Social Media Survey (OSMS), shown in Table 1, was distributed to SM users in Indonesia. Age categories spanned from 15 to ≥65 years old. Table 6 presents the demographic composition of respondents obtained from OSMS, with no respondents in the 15-20 or ≥65 brackets. The survey was conducted between January 27 and April 2, 2021. After removing incomplete responses, 345 valid responses were retained. Of the 44 OSMS questions in Table 1, the 5 demographic items and the FB-specific items within the 29 activities and behavior questions contribute to the 21 TPB features S1-S21 in Table 2, together with features derived from each respondent's actual FB post data.

The respondents were geographically distributed across 36 cities: 30 in Southeast Asia, 3 in Europe, 2 in West Asia, and 1 in North America. This distribution reflects the global mobility of Indonesian SM users who participated in the survey.

In the second stage, respondents were invited to voluntarily provide access to their publicly available FB data. Among the 345 survey participants, 184 respondents consented to share access to their FB activity. FB posts dating between December 1, 2019 and April 8, 2021 were retrieved between March and June 2021, yielding 12,888 posts and 90,901 comments. After applying filtering criteria based on post length and data completeness, 81 users had posts that satisfied the dataset requirements. This process resulted in 1,841 lengthy FB posts (LFP), which were used as the final dataset for subsequent analysis. This dataset has been deposited at Mendeley Data with permanent DOI https://doi.org/10.17632/6y6fgjcysj.1.

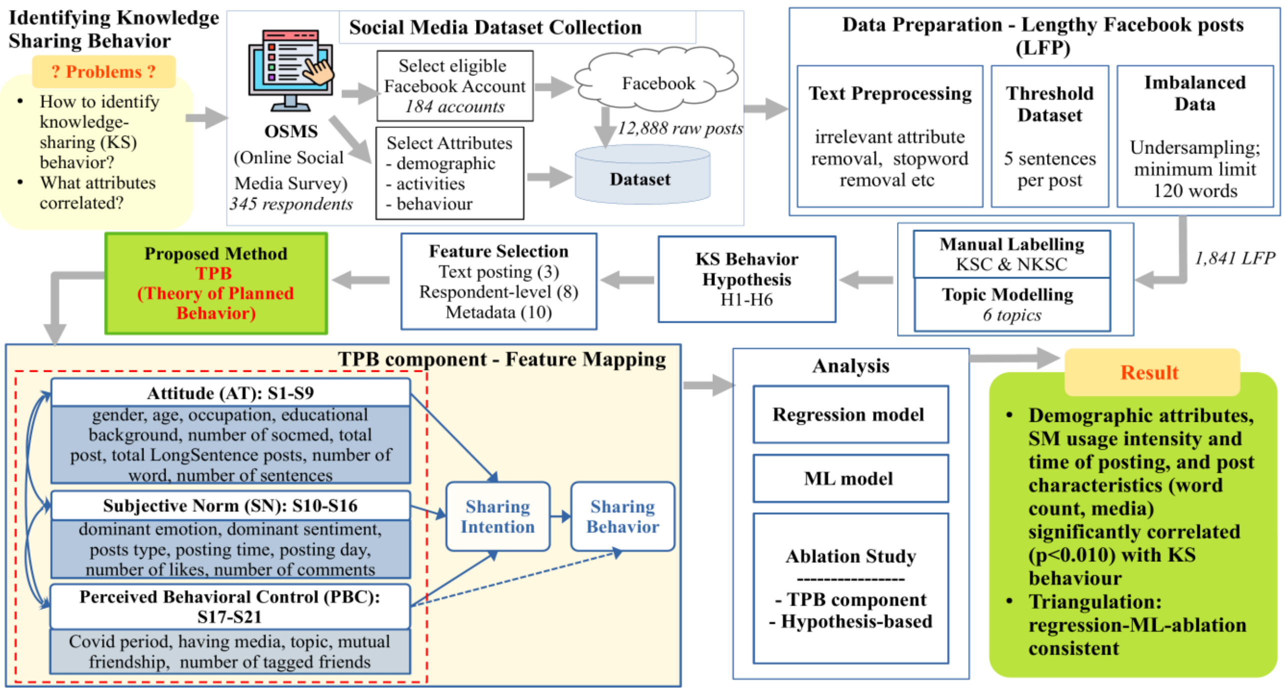

Figure. 1 Proposed Knowledge-Sharing Behavior on Facebook Using Theory of Planned Behavior

Table 1. OSMS question structure

Category	No. Q
Demographic item (age, domicile, gender, occupation and educational background)	5
Activities & behavior item - SM types and SM usage - Frequency - the purpose of sign-up - accepting and follow friends - favorite choices - willingness to share	29
Private message choices - email - message application - SM direct message choices	2
Further contact (option for further contact when needed)	8
Total of question (Q)	44

Table 2. TPB-Based feature representation from the SM dataset

TPB Component	Code	Attribute
Attitude (AT)	S1	Gender (Male, Female)
	S2	Age group (15-20, 21-25, 26-35, 36-40, 41-50, 51-64, ≥65)
	S3	Occupation (Teacher/Lecturer, Housewife, Government Employee, Private Employee, Student, Entrepreneur, Others)
	S4	Educational background (Economics, Educational Science, Non-ICT Science, ICT Science, Arts/Culture/Linguistics, Social Science)
	S5	Number of social media platforms
	S6	Total number of public FB posts
	S7	Number of lengthy Facebook posts
	S8	Number of words
	S9	Number of sentences
Subjective Norm (SN)	S10	Emotion (anger, fear, happy, love, neutral, sadness)
	S11	Sentiment (neg., neu., pos., none)
	S12	Post type (shared, friend-share, original)
	S13	Posting time (morning-late night)
	S14	Posting day (weekday/weekend)
	S15	Number of likes
	S16	Number of comments
Perceived Behavioral Control (PBC)	S17	COVID-19 period(before/after)
	S18	Media presence (text/text+media)
	S19	Topic category (from LDA)
	S20	Mutual friendship type
	S21	Number of tagged friends
Row color of dataset feature selected: Green: Text posting (3) Yellow: Respondent-level features (8): S1-S7, S20 White: Metadata (10)		

2.2 Data preprocessing

Text preprocessing was conducted to standardize FB posts prior to analysis. The preprocessing pipeline consisted of several steps. First, non-textual elements such as URLs, user mentions, hashtags, line breaks, and special symbols were removed. The text was then normalized through case folding and space normalization, followed by tokenization.

To address linguistic variations commonly found in Indonesian SM text, slang words and informal abbreviations were normalized using a manually compiled dictionary. Stopword removal was performed using the Sastrawi library, and lemmatization was conducted using NLP-ID to reduce words to their base forms.

Because SM posts are typically short and may contain limited informational value, a minimum threshold of five sentences per post was applied to retain posts that were more likely to contain meaningful knowledge-sharing information. This filtering step excluded casual updates and short reactions commonly found on SM platforms. After applying this criterion, the dataset consisted of 1,841 LFP generated by 81 users. The complete data preprocessing and feature construction pipeline is described in Pseudocode 1.

Following preprocessing and filtering, the labeling process was conducted manually based on predefined criteria for knowledge-sharing content. Ambiguous cases were reviewed to ensure labeling consistency. The posts were manually labeled into two categories: Knowledge Sharing Content (KSC) and Non-Knowledge Sharing Content (NKSC). The example ilustrated in Fig. 2. Categorical variables were encoded using appropriate encoding techniques, while numerical features were retained in their original form.

Pseudocode 1: Data Preprocessing and Feature Construction

Input:

D_{survey}: Raw survey dataset (345 OSMS respondents; 44 OSMS questions including 5 demographic and 29 activity/behavior items)
D_{posts}: Raw Facebook post dataset (posts dated Dec 2019 - Apr 2021, retrieved Mar-Jun 2021)

Output:

$X_{processed}$: Feature matrix (1,841 × 21) representing TPB attributes per post
$Y_{processed}$: Binary target label (KSC = 1, NKSC = 0)

Steps:

1. Load D_{survey}; filter by active Facebook accounts → $D_{matched}$ (184 accounts)
2. Retrieve Facebook posts for $D_{matched}$ → D_{raw} (12,888 posts, 90,901 comments)
3. Text preprocessing; FOR each post p in D_{raw}:
   a. Clean(p): remove URLs, special characters, punctuation → p_{clean}
   b. Normalize(p_{clean}): lowercase, slang dictionary → p_{norm}
   c. Tokenize + Lemmatize(p_{norm}) using NLP-ID + Sastrawi → p_{tokens}
4. Filter D_{raw} by sentence count ≥ 5 → D_{lfp} (1,841 LFP)
5. Label each post in D_{lfp} by manual annotation (three categories):
   - Category 0: non-educational content → NKSC (p.label = 0)
   - Category 1 (less-educational) + Category 2 (educational) → KSC (p.label = 1) → Final: 534 KSC, 1,307 NKSC [Note: BiLSTM was tested as a preliminary experiment on 836 posts (≥120 words), accuracy = 66.67%; not used for main labels]
6. Apply LDA (6 topics, coherence = 0.353) to D_{lfp} → S19 per post
7. FOR each post p, map survey responses to TPB features:
   a. AT_{score} = μ(S1-S9)
   b. SN_{score} = μ(S10-S16)
   c. PBC_{score} = μ(S17-S21)
8. Construct $X_{processed}$ = [S1, S2, …, S21] per post; $Y_{processed}$ = p.label
9. Stratified split (80:20): $X_{processed}$, $Y_{processed}$→ X_{train}, X_{test}, Y_{train}, Y_{test}

2.3 Topic modelling

To explore the thematic structure of the FB posts, topic modeling was performed using Latent Dirichlet Allocation (LDA) [26]. The LDA model was trained on the preprocessed textual corpus obtained from the preprocessing stage described in Section 2.2. The implementation was conducted using the Gensim library.

To determine the optimal number of topics, hyperparameter tuning was performed across several parameters, including the number of topics, iterations, alpha (α), eta (η), and gamma (γ). Model quality was evaluated using the Coherence score, which measures the semantic interpretability of generated topics. The configuration with the highest coherence score was selected as the final topic model.

2.4 Feature selection and mapping based on the theory of planned behavior

TPB explains behavior through three components: Attitude (AT), Subjective Norm (SN), and Perceived Behavioral Control (PBC). These constructs are abstract. We drew indicators from two sources: OSMS survey responses (44 questions per respondent) and metadata extracted from each respondent's FB posts.

From these two sources, we selected 21 indicators that align with the TPB constructs. Selection followed two criteria: the indicator had to be observable in the data rather than self-reported, and it had to plausibly proxy for AT, SN, or PBC based on prior TPB literature.

Post ID: 10158205704438360 User: vr001 [KSC]

Trying to estimate how much Google Cloud charges for a stopped instance. According to the documentation: "A stopped instance does not incur charges, but all resources attached to it continue to incur charges, such as persistent disks and external IP addresses." So, for one day: $0.010 × 24 + ($0.040/30) = $0.2413, roughly Rp3,800. Quite affordable for a large server that is not used around the clock. Just shut it down when not in use, then restart when needed. Worth considering :) Starting to explore Google Cloud :D Also looking at other solutions such as preemptible VM instances.

*Sedang menduga-duga berapa harga Google Cloud untuk Instance yang di stop (shutdown). Jika melirik dokumentasinya, tertulis "A stopped instance does not incur charges, but all of the resources that are attached to the instance continue to incur charges. For example, you are charged for persistent disks and external IP addresses according to the price sheet, even if an instance is stopped." Jadi kalau tidak salah hitung, harga untuk sebuah instance yang di stop dalam 1 hari adalah: $0.010 * 24 + ($0.040/30) = $0.2413 atau sekitar Rp3,800. Murah juga ya untuk sebuah server besar yang tidak dipakai setiap waktu. Jadi shutdown jika tidak digunakan, lalu hidupkan kembali saat digunakan. Boleh juga :) Mulai tertarik menggunakan Google Cloud :D Sedang lihat-lihat solusi lain seperti preemptible VM instances.*

Post ID: 10222646930215244 User: vr291 [NKSC]

Everyone is bullying [me], no one is cheering [me] on, "Come on, Dad! You can definitely pass the certification! Go for it! Don't be scared!" Education means motivating, not bullying people who want to learn. #uhuk

Semua orang membully nggak ada yang menyemangati, "Ayo Ayah Naen.. Kamu pasti bisa lulus sertifikasi! Cemungudh!! Jangan atut!!" Pendidikan itu memotivasi bukan membully orang yang mau belajar. #uhuk

English translation (regular); Indonesian original (italic).

Figure. 2 Example of LFP labeled as KSC and NKSC. The original data is represented in italics

The 21 indicators (S1-S21) are organized by source into three groups: three text-content features (S8, S9, S19), eight respondent-level features (S1-S7, S20), and ten post-metadata features (S10-S18, S21). This grouping appears in the Feature Selection box of Fig. 1 and in the green, yellow, and white color coding of Table 2. Educational background (S4) was already associated with occupation (S3), so it was not included as a separate H1 predictor.

Table 2 summarizes the mapping. AT contains nine features: demographic attributes (S1-S4), respondent posting volume (S5-S7), and content-effort indicators (S8-S9). SN covers social-interaction signals from S10 to S16: emotion, sentiment, post type, posting time and day, likes, and comments. PBC contains contextual factors S17-S21: COVID-19 period, media presence, LDA topic, mutual friendship type, and tagged friends.

Mutual friendship maps to PBC, not AT. Why? It captures perceived control over who reaches the post, not the user's general attitude. Audience choice (broad versus selective) is itself an enabling condition that supports or limits sharing [22, 27]. These 21 indicators serve as inputs for both regression analysis and ML classification.

2.5 Statistical analysis using regression

To examine the relationship between the selected features and KSB, binary logistic regression was employed. In this study, the KSC/NKSC binary label was treated as the dependent variable, while the TPB-based features (S1-S21) served as independent variables.

The logistic regression model estimates the probability of a post being classified as KSC given the TPB-based feature values. The general form of the model is expressed as Eq. (1), where p denotes the probability that a post is KSC, β_0 is the intercept, $\beta_1 \ldots \beta_n$ are the logistic regression coefficients for each predictor, and $X_1 \ldots X_n$ are the TPB feature values (S1-S21). Each coefficient represents the change in the log-odds of KSC per unit increase in the corresponding feature. This regression analysis was used to evaluate the statistical significance of the TPB-based features in explaining KSB in FB posts.

$$log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \ldots + \beta_n X_n \quad (1)$$

Regression analysis was also used to test the proposed hypotheses (H1-H6) by evaluating the statistical significance of each feature group, namely demographic attributes, social interaction factors, and posting characteristics, on KSB. Each hypothesis was examined through a separate regression model, allowing for a focused assessment of the contribution of each feature group to the prediction of KSB.

2.6 Machine learning models

The dataset consisting of 1,841 posts was divided into training and testing sets using an 80/20 split, with stratified sampling applied to preserve the class distribution between KSC and NKSC. For models sensitive to feature scaling, such as Logistic Regression and SVM, the input features were normalized using StandardScaler. Hyperparameter optimization was performed using grid search combined with 5-fold cross-validation. The search space for each model is summarized in Table 3.

Model performance was evaluated using several classification metrics, including accuracy, precision, recall, F1-score, and the area under the receiver operating characteristic curve (AUC-ROC). Among these metrics, AUC-ROC was used as the primary evaluation metric. The complete model training and evaluation procedure is described in Pseudocode 2.

Pseudocode 2: Model Training and Evaluation

Input:
- X_{train}, Y_{train}: Training feature set and labels
- X_{test}, Y_{test}: Test feature set and labels
- Model set M = {Logistic Regression, Random Forest, Gradient Boosting, SVM, Neural Network}

Output:
- M_{best}: Best-performing model
- $Performance_{Summary}$: Accuracy, Precision, Recall, F1-Score, AUC-ROC, Confusion Matrix, CV error, stddev (per model)

Steps:
1. FOR each model M_i in set M:
   a. Apply class weighting (inverse class frequency) to M_i
   b. Train M_i (X_{train}, Y_{train}) with stratified 5-fold CV → $M_{i,trained}$
   c. $M_{i,trained}$.predict(X_{test}) → $Y_{pred,i}$
   d. Evaluate(Y_{test}, $Y_{pred,i}$) → Accuracy, Precision, Recall, F1-Score, AUC-ROC
   e. ComputeConfusionMatrix(Y_{test}, $Y_{pred,i}$) → TP, TN, FP, FN
   f. ComputeCVError(M_i, X_{train}, Y_{train}) → CV_{error}, stddev
   g. Store metrics in $Performance_{Summary}[M_i]$
2. Select M_{best} = argmax AUC-ROC from $Performance_{Summary}$
3. Finalize M_{best}: retrain with full X_{train}, Y_{train} → $M_{best,final}$
4. Output $M_{best,final}$ and $Performance_{Summary}$

To assess cross-user generalizability, a supplementary evaluation was performed using StratifiedGroupKFold (k=5) with user ID (id_user) as the grouping variable, ensuring all posts from the same user appear exclusively in either the training or validation set across each fold. Both within-cohort AUC (stratified 80/20 split, random_state=42) and cross-user AUC (StratifiedGroupKFold k=5, mean ± SD) are reported in Table 9 to distinguish within-dataset discriminability from cross-user generalizability. The complete experimental configuration for reproducibility is summarized in Table 4. The cross-user evaluation procedure is detailed in Pseudocode 3.

2.7 Hypothesis-based ablation analysis

To further examine the contribution of different feature groups to the prediction of KSB, a hypothesis-based ablation analysis was conducted. This analysis evaluates how model performance changes when specific groups of features associated with the proposed hypotheses (H1-H6) are removed from the full feature set.

The full model includes 21 TPB-based features representing demographic attributes, social interaction indicators, and contextual posting characteristics.

Table 3. Hyperparameter Search Space for ML models

Model	Hyperparameter	Search Space
Logistic Regression	Regularization (C)	{0.01, 0.1, 1, 10}
	Class weight	{balanced, none}
Random Forest	Number of trees	{100, 200}
	Maximum depth	{10, 20, none}
	Class weight	{balanced, none}
Gradient Boosting	Number of estimators	{100, 200}
	Learning rate	{0.01, 0.1}
	Maximum depth	{3, 5}
SVM	Kernel function	{rbf, linear}
	Regularization (C)	{0.1, 1, 10}
	Class weight	{balanced, none}
Neural Network	Hidden layer size	{50, 100}
	L2 regularization (α)	{0.0001, 0.001}

Random state was fixed at 42 for all experiments.

Received: April 20, 2026. Revised: June 4, 2026.

Table 4. Reproducibility summary: hyperparameter settings and experimental configuration

Item	Setting	Value / Detail
Logistic Regression	C	10
Logistic Regression	class_weight	balanced
Random Forest	n_estimators	200
Random Forest	max_depth	10
Random Forest	class_weight	None
Gradient Boosting	n_estimators	100
Gradient Boosting	learning_rate	0.1
Gradient Boosting	max_depth	3
SVM	C	1
SVM	kernel	rbf
SVM	class_weight	balanced
Neural Network	hidden_layer_sizes	(100,)
Neural Network	alpha	0.0001
CatBoost	iterations / depth / lr	200 / 6 / 0.01
CatBoost	auto_class_weights	SqrtBalanced
Preprocessing	Age encoding	Range → midpoint (e.g., 26-35 → 30.5)
Preprocessing	Categorical encoding	Ordinal encoding (Gender, Job, Education, etc.)
Preprocessing	Missing values	Column median imputation
Preprocessing	Infinite values	Replaced with column median
Preprocessing	Feature scaling	StandardScaler (LR, SVM, NN only; not RF, GB)
Preprocessing	Class imbalance	class_weight tuned via GridSearchCV (not fixed)
Validation	Inner CV	5-fold GridSearchCV
Validation	Outer evaluation (within-cohort)	Stratified 80/20 split
Validation	Outer evaluation (cross-user)	StratifiedGroupKFold k=5, groups=id_user
Validation	Random seed	42 (all operations: np.random.seed(42), random_state=42)
Validation	Feature matrix	21 features S1-S21 ; see Table 2

In the ablation procedure, feature groups corresponding to each hypothesis were systematically removed from the model while keeping the remaining features unchanged. The resulting performance degradation was used to estimate the contribution of each hypothesis group.

The ablation experiments were performed using the Random Forest classifier, which achieved the highest AUC-ROC of 0.7805 on the 80/20 stratified test split. For the ablation procedure, the baseline AUC of 0.7613 was computed via 5-fold cross-validation on the training set, so the two figures are not directly comparable. Model performance was measured using AUC-ROC, and the contribution of each hypothesis group was quantified based on the difference between the full model performance and the performance of the reduced model. This approach allows the identification of which TPB-related feature groups contribute most to the predictive performance of the classification model. This analysis provides a quantitative assessment of the contribution of each hypothesis to model performance. The complete ablation procedure is outlined in Pseudocode 4.

Hypothesis-based ablation tests each proposed group of hypotheses (H1-H6) by removing it and reading the change in AUC-ROC. This is leave-one-group-out feature selection applied to a single classifier. The purpose is hypothesis testing. Rankings under other selection methods could differ.

Pseudocode 3: Cross-User Validation via StratifiedGroupKFold

Input:
X: Feature matrix (1,841 × 21 TPB features, S1-S21)
Y: Target labels (KSC = 1, NKSC = 0)
groups: anon_user_id (81 unique anonymized users)
K: Number of folds = 5
M: Set of classifiers = {LR, RF, GB, SVM, NN, CatBoost}
Output:
$AUC_{user}[M_i]$: Mean ± SD AUC-ROC across user-grouped folds per classifier
$Diff[M_i]$: $AUC_{within}[M_i] - AUC_{user}[M_i]$
Steps:

Received: April 20, 2026. Revised: June 4, 2026.

1. Initialize StratifiedGroupKFold(n_splits = K, groups = anon_user_id)
2. FOR each classifier M_i in M:
   a. fold_metrics = [ ]
   b. FOR each fold (train_idx, test_idx) in StratifiedGroupKFold.split(X, Y, groups):
      i. X_{train}, X_{test} = X[train_idx], X[test_idx]
      ii. Y_{train}, Y_{test} = Y[train_idx], Y[test_idx]
      iii. Inner GridSearchCV(M_i, params, cv = 3, scoring = 'roc_auc') → $M_{i,best}$
      iv. $M_{i,best}$.fit(X_{train}, Y_{train})
      v. Y_{prob} = $M_{i,best}$.predict_proba(X_{test})[:, 1]
      vi. fold_auc = roc_auc_score(Y_{test}, Y_{prob})
      vii. fold_metrics.append(fold_auc)
   c. $AUC_{user}[M_i]$ = mean(fold_metrics) ± std(fold_metrics)
   d. $Diff[M_i] = AUC_{within}[M_i] - AUC_{user}[M_i]$
3. Output AUC_{user}, Diff for Table 9

Pseudocode 4: Hypothesis-Based Ablation Study

Input:
X_{full}: Complete feature set (21 TPB features, S1-S21)
Y: Target labels (KSC / NKSC)
$component_{map}$: {AT → S1-S9, SN → S10-S16, PBC → S17-S21}
$hypothesis_{map}$: {H1 → demographic, H2 → education & friendship, H3 → SM usage, H4 → COVID-19, H5 → audience response, H6 → post characteristics}
Output:
AUC_{full}: Baseline AUC with all 21 features
$\Delta AUC_{hypothesis}[H_i]$: AUC drop when H_i features are removed
$\Delta AUC_{component}[C]$: AUC drop when TPB component C is removed
AblationTable: Ranked contribution per hypothesis and component
Steps:
1. Train RF(X_{full}, Y, stratified 5-fold CV) → AUC_{full}
2. Hypothesis-level ablation:
   FOR each H_i in {H1, H2, H3, H4, H5, H6}:
   a. $X_{full} - F_{Hi} \rightarrow X_{h,ablated}$
   b. Train RF($X_{h,ablated}$, Y) → $AUC_{ablated,Hi}$
   c. $AUC_{full} - AUC_{ablated,Hi} \rightarrow \Delta AUC_{hypothesis}[H_i]$
   d. Store (H_i, $\Delta AUC_{hypothesis}[H_i]$) → AblationTable
3. Component-level ablation:
   FOR each C in {AT, SN, PBC}:
   a. X_{full} − features(C) → $X_{ablated}$
   b. Train LR($X_{ablated}$, Y) → $AUC_{ablated,C}$
   c. $AUC_{full} - AUC_{ablated,C} \rightarrow \Delta AUC_{component}[C]$
   d. Store (C, $\Delta AUC_{component}[C]$) → AblationTable
4. Sort AblationTable by ΔAUC descending → ranked contribution
5. Output AUC_{full}, AblationTable

3. Result

3.1 Topic modeling results

The best LDA configuration reached a Coherence score of 0.353. The settings: six topics, 150 iterations, and $\alpha = 0.100$, $\eta = 0.100$, $\gamma = 0.001$, as shown in Fig. 3. Six dominant topics emerged. The labels are “life guides”, “new normal COVID-19”, “health”, “invitation”, “family”, and “hobby/interest”. About 25.4% of posts belong to “new normal COVID-19” alone, showing how heavily pandemic themes drove user discussions.

A closer look at the topic distributions points to “life guides” and “new normal COVID-19” as the most educational and informational themes. Longer FB posts often served as guidance: how to live with the pandemic, how to handle daily life, what to do next. This is an initial mapping of themes. Regression and ML take over from here.

3.2 Descriptive analysis of dataset

This section presents the descriptive characteristics of the dataset used in this study. The distribution of FB posts is analyzed based on demographic attributes, posting characteristics, and content-related features.

Table 5 presents the descriptive statistics for the 1,841 lengthy FB posts. The summary spans four categories. These are demographic attributes, posting characteristics, emotional expressions, and sentiment patterns. Each variable is shown with counts and percentages. As noted in the Table 5 footnote, categories with frequency below 5% are omitted, so the percentages reflect dominant subsets, not the full distribution. Regression and ML analyses follow.

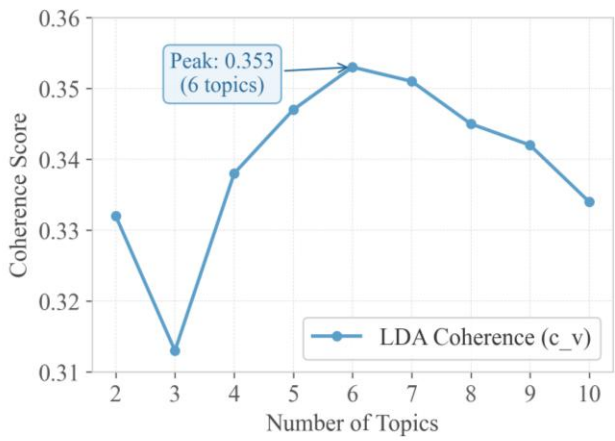

Figure. 3 Coherence Score Diagram

Table 5. Summary of lengthy FB posts showing categories with frequency ≥ 5% of N (N=1,841)

Attribute	Category	Frequency	Percentage (%)
Gender	Male	674	36.60
	Female	1,167	63.40
Age	26-35	328	17.80
	36-40	163	8.90
	41-50	1,161	**63.10**
	51-64	167	9.10
Occupation	Teacher/Lecturer	691	**37.50**
	Housewife	129	7
	Private employee	247	13.40
	Entrepreneur	402	21.80
	Others	302	16.40
Posting time	Morning	409	22.20
	Noon	534	29
	Afternoon	317	17.20
	Night	505	27.40
COVID-19 period[a]	Before	274	14.90
	After	1,567	**85.10**
Day category	Weekday	1,349	73.30
	Weekend	492	26.70
Posts type	Shared with friends	291	15.80
	Own posts	1,542	83.80
Dominant emotion	Anger	293	15.90
	Happy	731	39.70
	No emotion	571	31
	Sadness	205	11.10
Dominant sentiment	No sentiment	571	31
	Negative	479	26
	Neutral	145	7.90
	Positive	646	35.10

[a] First confirmed COVID-19 case in Indonesia on March 2, 2020.
This table presents a truncated summary. Categories with frequency < 5% of N are omitted from the main rows. Omitted categories: posting time, Wee hour, n=76, 4.1%; post type, Shared from other platform, n=2, 0.1%; Posts from friends, n=6, 0.3%; age group 21-25, n=22, 1.2%.

Table 6 reports gender distribution. Female users dominated both classes. They contributed 53.9% of KSC and 67.3% of NKSC. Male users contributed 46.1% and 32.7% respectively.

The 41-50 group led age distribution. They contributed 54.9% of KSC and 66.4% of NKSC posts. The 26-35 group followed at 22.8% of KSC. The 36-40 group showed higher KSC than NKSC. Most KSC came from users aged 26-50.

Fig. 4 splits the data into two views. The upper panel shows the 81 users by age group and gender. The 41-50 group forms the largest cohort. The lower panel shows the 1,841 posts by age group, split into NKSC and KSC.

Respondents' educational backgrounds were grouped into six academic fields commonly found in Indonesian higher education. The largest proportion of users producing LFP came from ICT-related disciplines (46.91%), followed by social sciences (16.05%), economics (13.58%), and non-ICT sciences (13.58%). Smaller proportions were observed for arts, culture, and linguistics (6.17%) and educational sciences (3.70%).

Respondents' occupations were categorized into seven groups, as presented in Table 6 and the upper panel of Fig. 5. Private employees represented the largest occupational group (27%), followed by teachers/lecturers (22%) and entrepreneurs (16%). In the lower panel of Fig. 5, although teachers and lecturers generated the highest number of posts (n = 691, of which 119 were KSC), they exhibited the lowest KSC proportion (119/691 = 17.2%). In contrast, students showed the highest KSC proportion (2 of 3 posts, 66.7%).

Table 6. Summary of respondent's demographic and knowledge-sharing content of lengthy FB post

Attibutes	81 Users Producing LFP (%)	Labeled LFP (N=1841)	
		KSC (%) (N=534)	NKSC (%) (N=1,307)
Gender (p=0.001)			
Male	53.090	46.100	32.700
Female	46.910	53.900	67.300
Age (p=0.001)			
15-20	-	-	-
21-25	4.940	0.700	1.400
26-35	18.520	22.800	15.800
36-40	23.460	13.500	7.000
41-50	44.440	54.900	66.400
51-64	8.640	8.100	9.500
>=65	-	-	-
Occupation (p = 0.001)			
Teacher/Lecturer	22.220	22.300	43.800
Housewife	11.110	7.500	6.800
Government employee	12.350	4.900	3.100
Private employee	27.160	15.500	12.500
Student	2.470	0.400	0.100
Entrepreneur	16.050	25.100	20.500
Other	8.640	24.300	13.200

Percentages in KSC and NKSC columns represent the distribution within each class
Highest percentage values per column shown in bold

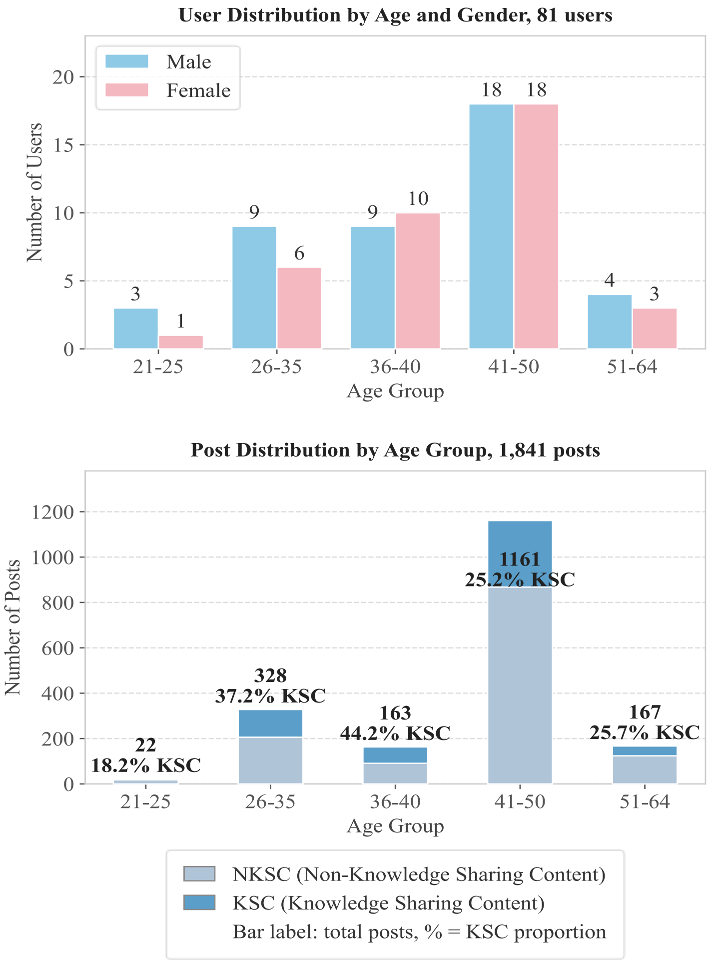

Figure. 4 Respondent's demographic of Lengthy Facebook post (LFP)

Received: April 20, 2026. Revised: June 4, 2026.

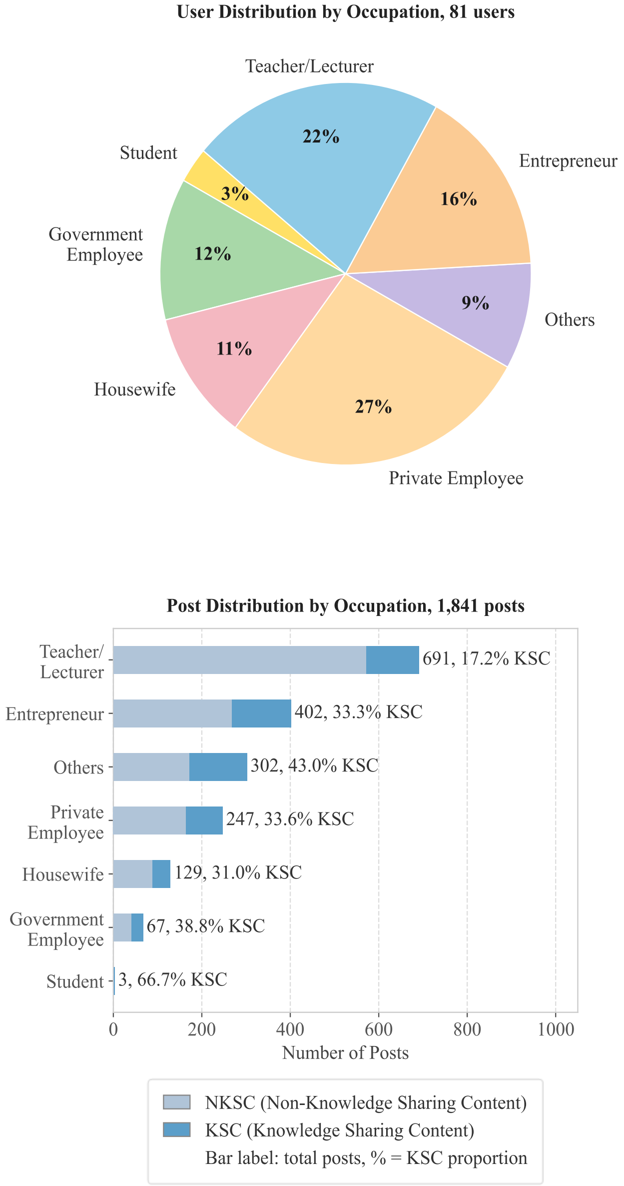

Figure. 5 Dataset distribution based on respondents' occupation

These per-occupation KSC proportions were derived by cross-referencing the post totals in Table 5 with the within-class distributions in Table 6. Educational background also relates to sharing behavior through its association with occupational categories.

3.3 Regression analysis for hypothesis testing

The regression models use the TPB-based features (S1-S21) defined in Table 2 as predictor variables. Binary logistic regression was applied to evaluate the six proposed hypotheses, with KSC/NKSC as the binary outcome. Model fit is evaluated using model chi-square (χ^2), Nagelkerke pseudo-R^2, and Wald-test p-values. Nagelkerke pseudo-R^2 is a logistic-regression goodness-of-fit measure scaled from 0 to 1, while the Wald-test p-value indicates whether each predictor coefficient differs significantly from zero. Six regression models were constructed using feature groups corresponding to the hypotheses (H1-H6), allowing the influence of demographic attributes, social interaction factors, and post characteristics to be examined separately. The regression results are summarized in Table 7. The findings of each regression model are discussed below according to the corresponding hypotheses.

- H1: Demographic factors.

  Model 1 examines whether demographic attributes influence KSB. The results show that gender, age, and occupational category are significant predictors of KSC ($\chi^2(3)$ =

Received: April 20, 2026. Revised: June 4, 2026.

50.16, $p < 0.001$, Nagelkerke $R^2 = 0.038$). Most KS posts were generated by users aged 26-50 years, indicating that middle-aged users are more active in sharing educational content on FB. This pattern matches prior work on demographic effects in online KSB. H1 is supported.

- H2: Individual interests and mutual friendships.

  Model 2 evaluates the effect of educational background and mutual friendship types. The regression results indicate that educational background and friendship relationships with workmates and business partners significantly influence KSB ($\chi^2(3) = 14.32$, $p = 0.003$, Nagelkerke $R^2 = 0.011$). Professional and interest-based networks matter for knowledge-sharing activity. H2 is supported.

  The five friendship categories do not behave uniformly. Business partner ties were the strongest predictor ($p < 0.001$), while workmate ties showed a negative bivariate association despite remaining significant in the combined model. Family and school friend connections fell near zero. The divergence in both direction and magnitude suggests that which network a user shares with matters more than whether mutual friendships exist at all, a pattern more consistent with behavioral control than with a fixed personal attribute [27,28].

- H3: SM usage intensity.

  Model 3 examines whether the intensity of SM usage influences KSB. The results show that the number of SM platforms (S5) and total posting activity (S6) significantly predict KSC ($\chi^2(4) = 154.18$, $p < 0.001$, Nagelkerke $R^2 = 0.115$), while posting time and day category do not show significant effects. These findings indicate that users with higher posting activity tend to share more knowledge-related content. H3 is supported.

- H4: COVID-19 pandemic context.

  Model 4 examines the influence of the COVID-19 pandemic period on KSB. The model is statistically significant ($\chi^2(4) = 33.94$, $p < 0.001$, Nagelkerke $R^2 = 0.026$), indicating that the contextual features in this group collectively account for some variation in KSC. The modest effect size reflects the indirect nature of pandemic context as a predictor; nonetheless, the overall result supports H4.

- H5: Responses from others.

  Model 5 examines whether audience responses influence KSB. The results indicate that variables related to user responses (including likes, comments, tagged friends, and sentiment indicators) do not significantly predict KSC behavior ($\chi^2(6) = 3.83$, $p = 0.700$, Nagelkerke $R^2 = 0.003$). KSC posts come from intrinsic motivations, not audience feedback. H5 is rejected.

- H6: Post characteristics.

  Model 6 evaluates whether the characteristics of posts influence KSB. The results show that word count, sentence count, and the presence of additional media significantly predict KSC ($\chi^2(4) = 58.46$, $p < 0.001$, Nagelkerke $R^2 = 0.045$), indicating that longer and richer posts are more likely to contain knowledge-sharing content. However, topic category does not show a significant effect. H6 is supported.

Table 7. Summary of logistic regression models for hypothesis testing (H1-H6)

Logistic Regression Model	Hypothesis	Features (n)	χ^2(df)	p-value	Nagelkerke R^2
Model 1	H1: Demographics	3	50.16 (3)	< 0.001	0.038
Model 2	H2: Mutual friendship	3	14.32 (3)	0.003	0.011
Model 3	H3: SM intensity	4	154.18 (4)	< 0.001	0.115
Model 3a	H3: SM intensity (alt.)	4	127.62 (4)	< 0.001	0.096
Model 4	H4: COVID context	4	33.94 (4)	< 0.001	0.026
Model 5	H5: Others' response	6	3.83 (6)	0.700	0.003
Model 6	H6: Post characteristics	4	58.46 (4)	< 0.001	0.045

χ^2 indicates the overall model significance in binary logistic regression. Nagelkerke R^2 is reported as a pseudo-R^2 measure for model fit. H5 is not supported ($p = 0.700$); H1-H4 and H6 are supported ($p < 0.05$).
Model 3a is an alternative operationalization of H3 using S7 (lengthy posts) instead of S6 (total posts).

3.4 Machine learning classification results

In addition to regression analysis, six ML classifiers, Logistic Regression, Random Forest, Gradient Boosting, Support Vector Machine (SVM), Neural Network, and CatBoost, were applied to assess the predictive capability of the TPB-based features. Model performance was evaluated using accuracy, precision, recall, F1-score, and AUC-ROC.

Table 8 lists the classification performance, cross-validation stability, and recommended use case for each model. Random Forest and CatBoost achieved comparable AUC-ROC: 0.7805 for Random Forest and 0.7801 for CatBoost. We used AUC-ROC for model selection. CatBoost also reached stronger accuracy of 0.7588 and F1 of 0.5436; pick it when balanced precision and recall matter most. Ensembles work. The other models trade differently. Gradient Boosting reached AUC-ROC = 0.7657. SVM had the highest recall at 0.7477. Logistic Regression gave a steady baseline. Neural Network was the weakest at AUC = 0.6689, likely from the small dataset.

To assess generalizability across users, an additional evaluation was conducted using StratifiedGroupKFold (k=5) with user ID as the grouping variable. This setup ensures that all posts from the same user appear in either the training or validation set, not both. Table 9 shows that cross-user AUC is substantially lower than within-cohort AUC for all models, with Random Forest dropping from 0.7805 to 0.5583 ± 0.0301. CatBoost shows a similar drop, from 0.7801 to 0.5463 ± 0.0448, indicating that the user-level reduction affects all six classifiers. This reduction is expected: demographic features such as Gender, Age, Job, and Education are constant across all posts from the same user, so within-cohort evaluation inadvertently allows the model to recognize individual users rather than generalizing behavioral patterns. This finding is discussed further in Section 5. The within-cohort results are retained as a measure of within-dataset discriminability of the 21 TPB features. Fig. 6 presents ROC curves for all six classifiers; Random Forest reached the highest AUC-ROC of 0.7805 and Neural Network the lowest of 0.6689.

Table 8. Performance comparison of ML classifiers on the within-cohort 80/20 split with post-level 5-fold cross-validation

Model	Accuracy	Precision	Recall	F1-score	AUC-ROC	Post-level 5-fold CV AUC (Mean±SD)	Stability	Recommended Use
Random Forest	0.7534	0.6481	0.3271	0.4348	**0.7805**	0.7633±0.0344	High	Balanced performance
Gradient Boosting	0.7425	0.6034	0.3271	0.4242	0.7657	0.7527±0.0355	High	Alternative to RF
SVM	0.6531	0.4420	0.7477	0.5556	0.7285	0.7230±0.0227	Very High	Maximize KSC detection
Logistic Regression	0.6341	0.4213	0.7009	0.5263	0.7166	0.6764±0.0190	Very High	Baseline model
Neural Network	0.6748	0.4343	0.4019	0.4175	0.6689	0.6750±0.0311	High	Limited dataset
CatBoost	0.7588	0.6023	0.4953	0.5436	0.7801	0.7614±0.0271	High	Best F1 score

Post-level 5-fold CV AUC = mean ± SD of AUC-ROC across ordinary 5-fold cross-validation at post level (within-dataset stability). These are not user-grouped folds; see Table 9 for cross-user evaluation.

Table 9. Comparison of within-cohort and cross-user AUC for each classifier

Model	Within-cohort AUC	Cross-user AUC (mean ± SD)	Difference
Logistic Regression (LR)	0.7166	0.4464 ± 0.0950	−0.2702
Random Forest (RF)	0.7805	0.5583 ± 0.0301	−0.2222
Gradient Boosting (GB)	0.7657	0.5192 ± 0.0387	−0.2465
Support Vector Machine (SVM)	0.7285	0.5625 ± 0.1137	−0.1660
Neural Network (NN)	0.6689	0.5332 ± 0.0196	−0.1357
CatBoost	0.7801	0.5463 ± 0.0448	−0.2338

Within-cohort AUC: stratified 80/20 split (random_state=42). Cross-user AUC: StratifiedGroupKFold k=5 with user ID as grouping variable. Negative Difference = reduction in AUC when generalizing to unseen users.

Received: April 20, 2026. Revised: June 4, 2026.

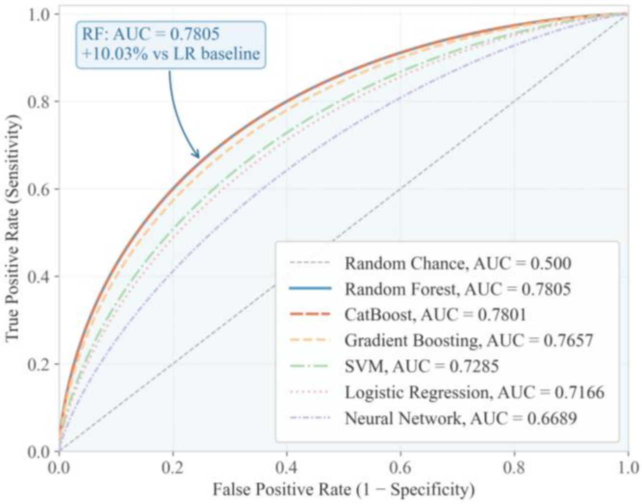

Figure. 6 ROC curves of six ML classifiers for KSC identification

To further validate the regression findings, each hypothesis model was evaluated using ML classification. Table 10 presents the best-performing algorithm for each regression model along with its corresponding AUC-ROC score. The results show that Random Forest consistently provides strong predictive performance, particularly for demographic attributes (H1) and the full feature model. Gradient Boosting performs best for the SM intensity models (H3) and mutual friendship (H2). For the audience-response model (H5), Neural Network obtains the highest AUC-ROC, but the value is only 0.5481, which is close to random prediction. This result is consistent with the logistic regression finding that H5 is not supported. For the COVID-19 context (H4), SVM is the best model, though with a modest AUC of 0.6379.

The ML results confirm the regression findings. Model 5 (Others' Response) achieved the lowest performance at AUC-ROC 0.5481, which aligns with the regression result where this hypothesis was rejected with R^2 = 0.003. The statistical and ML analyses agree on H5, adding confidence to the rejection.

Recent SM classification studies offer a useful reference frame in Table 11. Khan et al. [24] compared SVM and Random Forest for sentiment classification with accuracies of 0.8039 (SVM) and 0.7856 (RF); their study did not report AUC-ROC. Wu et al. [25] benchmarked Random Forest, XGBoost, LightGBM, CatBoost, and SVM for topic popularity prediction on Facebook data under a 1:22 class imbalance. CatBoost achieved the highest AUC-ROC of 0.82 and accuracy of 0.93. The present study's Random Forest result, AUC = 0.7805, falls within this range, though direct comparison is constrained by domain differences.

Table 10. Machine learning validation for regression models

Regression Model	Hypothesis	Features	Nagelkerke R^2	Best ML Model	AUC-ROC
Model 1	H1: Demographics	3	0.038	Random Forest	0.7318
Model 2	H2: Mutual friendship	3	0.011	Gradient Boosting	0.6481
Model 3	H3: SM intensity	4	0.115	Gradient Boosting	0.7331
Model 3a	H3: SM intensity (alt.)	4	0.096	Gradient Boosting	0.7461
Model 4	H4: COVID context	4	0.026	SVM	0.6379
Model 5	H5: Others' response	6	0.003	Neural Network	0.5481
Model 6	H6: Post characteristics	4	0.045	Neural Network	0.6868
Full Model	All features	21	-	Random Forest	**0.7805**

Nagelkerke R^2 from binary logistic regression in Table 7. AUC-ROC from standalone ML model using only features of each hypothesis group (best algorithm per group). Bold data indicates best AUC per column.

Table 11. Contextual comparison with related studies

Study	Year	Task	Platform	N	Best Model	AUC-ROC	Accuracy	Theory	Imbalance	Interpretability
This study	2026	KSC classification	Facebook	1,841	Random Forest²	0.7805	0.7534	TPB	Stratified + class weight	Hypothesis ablation
[25]	2025	Topic popularity	Facebook	3,570	CatBoost	0.82	0.93*	None	SMOTE	None
[23]	2025	Academic performance	Institutional records	922	Random Forest	0.87	0.7143	None	None	None
[24]	2024	Sentiment classification	Facebook / Twitter	-	SVM	-	0.8039	None	None	None

*Accuracy inflated by class imbalance ratio 22:1 in [25]

²Random Forest is selected as best by AUC-ROC (0.7805); CatBoost achieves the highest Accuracy (0.7588) and F1-score (0.5436).

The primary contribution extends beyond classifier performance: the hypothesis-based ablation quantifies which TPB constructs drive classification gain, an interpretive dimension absent in content-only approaches. Transformer-based models and LLM embeddings, while effective for raw text classification, operate on text representations and are not directly applicable to the tabular TPB-based behavioral features used here; the feature type and task definition differ substantially, making a direct performance benchmark methodologically inappropriate.

Direct comparison across studies is hard. Task definition, domain, and dataset scale all differ. Even so, our Random Forest at AUC = 0.7805 exceeds the RF baseline reported by Wu et al. [25] on FB data at AUC = 0.77. Their best model (CatBoost, AUC = 0.82) exceeds ours, but addresses a different task. They predict topic popularity with 38 engagement features, while we classify individual knowledge sharing with 21 TPB-derived features. TPB-based feature selection delivered competitive accuracy on a substantially smaller dataset. We add something the comparison studies lack: hypothesis-based ablation that quantifies what each behavioral construct contributes. Content-only approaches cannot do this

3.5 Hypothesis-based ablation analysis result

To measure each hypothesis group's contribution to predictive performance, we ran a hypothesis-based ablation analysis with Random Forest. The full model uses all 21 TPB-based features. Ablation variants drop one hypothesis group at a time. Performance is read off the AUC-ROC. Each group's contribution is the change in AUC relative to the full model, as shown in Table 12.

The results show that removing post characteristics (H6) produces the largest performance degradation with ΔAUC = 0.0488, indicating that textual attributes such as word count, sentence count, and media presence play the most important role in predicting knowledge-sharing content. This finding is consistent with the regression results where post characteristics significantly influence KSB.

In contrast, removing demographic attributes (H1) produces only a modest performance reduction with ΔAUC = 0.0142, suggesting that demographic information contributes to prediction but is less critical than content-based features. Similarly, mutual friendship and SM intensity (H2 and H3) show relatively small contributions. Removing the COVID-19 context variable (H4) slightly increases model performance by ΔAUC = +0.0024, suggesting that this variable adds noise rather than predictive signal.

Beyond the Without Hx removal effects above, the Only Hx setting offers a complementary view by training each model on the features of a single hypothesis group. This reveals which groups carry independent predictive signal and which depend on context from other features.

Among the Only Hx models, Only H1 (demographics) reaches the highest AUC = 0.7318 with Random Forest, confirming that demographic features alone carry substantial predictive information.

Table 12. Hypothesis-based ablation analysis using Random Forest

Model Variant	Features	AUC-ROC	ΔAUC	Contribution (%)	Nagelkerke R^2
Full model (RF 5-fold CV)[1]	**21**	0.7613	-	-	-
Without H1 (Demographics)	18	0.7471	−0.0142	1.86	0.038
Without H2 (Mutual friendship)	18	0.7533	−0.0080	1.05	0.011
Without H3 (SM intensity)	17	0.7418	−0.0195	2.56	0.115
Without H4 (COVID-19 context)	17	0.7637	+0.0024	−0.32	0.026
Without H5 (Others' response)	15	0.7497	−0.0116	1.52	0.003
Without H6 (Post characteristics)	17	0.7125	−0.0488	6.41	0.045
Only H1: Demographics (RF)	3	0.7318	-	-	0.038
Only H2: Mutual Friendship (GB)	3	0.6481	-	-	0.011
Only H3: SM Intensity (GB)	4	0.7331	-	-	0.115
Only H4: COVID-19 Context (SVM)	4	0.6379	-	-	0.026
Only H5: Others' Response (NN)	6	0.5481	-	-	0.003
Only H6: Post Characteristics (NN)	4	0.6868	-	-	0.045

[1]Baseline AUC = 0.7613 (RF 5-fold CV mean). Removal rows (Without Hx): ΔAUC = AUC_variant − AUC_full; negative = removing Hx hurts performance. Standalone rows (Only Hx): AUC of model trained on Hx features only; best algorithm shown in parentheses. Nagelkerke R^2 from binary logistic regression in Table 7.
Positive contribution values indicate that the removed feature group contributes positively to model performance.

Only H6 (post characteristics) follows at AUC = 0.6868 with Neural Network, consistent with the Without Hx result where removing H6 produced the largest drop. In contrast, the Only H5 model (others' response) produces the lowest performance with AUC = 0.5481, which is close to random prediction and consistent with the regression result where H5 was rejected. Overall, the ablation results confirm that post characteristics and demographic attributes are the most important contributors to knowledge-sharing prediction, while audience response indicators have minimal predictive value.

3.6 TPB component validation

To further examine the role of the TPB components in KSB, we conducted additional statistical and ML validation analyses.

Table 13 presents the Pearson correlation matrix between TPB components and KSB. Attitude shows a weak negative correlation with KSB ($r = -0.182$, $p < 0.001$), while perceived behavioral control (PBC) shows a very weak positive correlation ($r = +0.076$, $p < 0.01$). Subjective Norm shows negligible correlation with KSB. The low inter-component correlations ($\max |r| = 0.122$) indicate that the TPB constructs are relatively independent.

To further examine group differences, Table 14 compares TPB component scores between KSC and NKSC posts using independent-samples t-tests. The results show that KSC posts have significantly lower Attitude scores ($p < 0.001$) and higher PBC scores ($p = 0.001$), while Subjective Norm does not differ significantly between groups.

Finally, a component-level ablation analysis was performed using logistic regression to evaluate the predictive contribution of each TPB component. Removing the Attitude component results in the largest performance degradation with $\Delta AUC = 0.1694$ or 23.66%, confirming that Attitude is the dominant predictor. Removing Subjective Norm yields a marginal performance gain at $\Delta AUC = -0.0146$, suggesting it contributes minimally as an independent component. Removing PBC results in nearly no change at $\Delta AUC = 0.0011$ or 0.15%. These results are consistent with the regression and ML findings, confirming that attitudinal characteristics play the primary role in shaping KSB on SM.

4. Discussion

This study investigated KSC on FB based on the assumption that lengthy posts are more likely to contain educational and informational content. The findings from regression and ML analyses provide consistent evidence supporting this assumption.

The regression results show that demographic factors (H1), mutual interests and friendships (H2), SM usage intensity (H3), COVID-19 context (H4), and post characteristics (H6) significantly influence KSB. In contrast, audience response (H5) was not found to be a significant predictor. This finding may be explained by the distribution of emotional and sentiment features in the dataset, where a substantial number of posts showed no dominant emotion or sentiment, reducing the observable impact of audience feedback on posting behavior.

These results suggest that KSB on SM is primarily driven by intrinsic and content-related factors rather than external responses. This aligns with prior studies indicating that users are motivated to share knowledge based on personal intentions and perceived usefulness rather than social feedback mechanisms.

The strong influence of demographic and relational factors points to the role of social structure and user characteristics in shaping KSB. In particular, the role of professional connections (e.g., business partners) supports the principles of Weak Tie Theory [27, 28], where loosely connected networks facilitate broader information flow across different social groups.

The ML results back up the regression findings. Random Forest reached the highest AUC-ROC of 0.7805; ensemble methods captured the joint effect of the TPB-based features well. Ablation tells a similar story: removing post characteristics and attitudinal components causes the largest drops in performance, so those groups carry most of the signal. The comparison of Model 3 (S6, total post count) and Model 3a (S7, lengthy posts) for H3 splits two ways.

Table 13. Pearson correlation matrix of TPB components and KSB

Variable	Attitude	Subjective Norm	PBC	KSB
Attitude	1.000	0.048	−0.034	−0.182***
Subjective Norm	0.048	1.000	−0.122	−0.006
PBC	−0.034	−0.122	1.000	0.076**
KSB	−0.182***	−0.006	0.076**	1.000

*** p < 0.001; ** p < 0.01

Table 14. Comparison of TPB component scores between KSC and NKSC posts

Component	KSC (N = 534) Mean (SD)	NKSC (N = 1,307) Mean (SD)	Mean Diff	t	p	Cohen's d
Attitude	0.345 (0.096)	0.392 (0.122)	−0.047	−7.916	<0.001	−0.407
Subjective Norm	0.366 (0.133)	0.368 (0.137)	−0.002	−0.264	0.792	−0.014
PBC	0.484 (0.138)	0.461 (0.136)	0.023	3.271	0.001	+0.168

Independent samples t-test; Mean (SD) = mean and standard deviation.

Model 3 has higher Nagelkerke R^2 at 0.115 vs. 0.096; Model 3a has higher ML AUC at 0.7461 vs. 0.7331. In other words, total posts explain more variance in KSC across users, but lengthy posts are a sharper classifier. One caveat, under user-level cross-validation using StratifiedGroupKFold with k=5, AUC falls sharply for every model, with Random Forest at 0.5583 ± 0.0301. So within-cohort performance reflects discriminability inside this dataset, not generalization to unseen users. That fits the scope of the study, which is a behavioral analysis of one community rather than a deployable classifier.

The TPB component validation provides additional insights. The dominance of Attitude-related features, as reflected in both regression and ablation results, suggests that individual perception and internal motivation are the primary drivers of KSB. In contrast, the relatively weak contribution of Subjective Norm indicates that social pressure or external expectations have limited influence in this context.

From a theoretical perspective, these findings are consistent with multiple frameworks. The results align with Self-Determination Theory, which emphasizes intrinsic motivation as a key driver of behavior, as well as Social Capital Theory, where knowledge sharing is facilitated through network structures and relationships. The observed importance of professional and weak ties further supports Weak Tie Theory [27, 28], where bridging connections carry knowledge across communities.

Overall, this study suggests that online KSB differs from traditional face-to-face contexts, where social norms may play a stronger role. In SM environments, users are more influenced by personal attitudes, content characteristics, and network structures, while external responses and normative pressures play a less significant role.

5. Conclusion

This study investigated KSB on FB using a combination of TPB, regression analysis, ML, and hypothesis-based ablation techniques. By analyzing 1,841 lengthy FB posts from 81 users, the study provides both theoretical and predictive insights into KSB in SM contexts. The main contribution of this study is the systematic operationalization of TPB constructs as observable FB behavioral features, enabling empirical hypothesis testing on SM data without relying solely on self-reported surveys; the ML component serves as a complementary exploration of the discriminative capacity of the derived feature set.

The regression results confirm that demographic factors (H1), mutual interests and friendships (H2), SM usage intensity (H3), COVID-19 context (H4), and post characteristics (H6) significantly influence KSB. In contrast, audience response (H5) does not have a significant effect, suggesting that KSB is primarily driven by intrinsic motivation rather than external feedback mechanisms.

Random Forest reached AUC-ROC 0.7805 with CatBoost close behind at 0.7801. We used AUC-ROC for model selection. CatBoost also reached higher F1-score of 0.5436 and accuracy of 0.7588, so the right pick depends on what matters more. Both confirm the TPB feature set carries useful signal. SVM offered higher recall. Logistic Regression gave a steady baseline. User-level cross-validation gives a different reading. Under StratifiedGroupKFold (k=5), AUC drops sharply for every model. Take Random Forest: 0.5583 ± 0.0301. The within-cohort numbers reflect discriminability inside this dataset, not generalization to unseen users. That matches our scope: behavioral analysis of one FB group, not a deployable classifier.

The hypothesis-based ablation analysis reveals that post characteristics and attitudinal components are the most important contributors to predictive performance, while audience response features have minimal impact. Similarly, TPB component validation confirms that Attitude is the dominant factor, whereas Subjective Norm and Perceived Behavioral Control play relatively limited roles.

From a theoretical perspective, the findings support the relevance of TPB in SM contexts, while also aligning with Self-Determination Theory, Social Capital Theory, and Weak Tie Theory. The results suggest that online knowledge sharing is more strongly influenced by internal motivation and content characteristics than by social pressure or audience interaction, distinguishing it from traditional face-to-face knowledge-sharing environments.

From a practical standpoint, the findings suggest enhancing content quality and supporting users' intrinsic motivation to promote KSB on SM platforms. Future research may explore larger datasets, multimodal content (e.g., images and videos), and deeper analysis of user interaction dynamics, as well as investigate the relationship between KSB and user well-being.

Data and Code Availability

The raw Facebook text, usernames, profile links, and other potentially identifiable information are not publicly released due to privacy and platform restrictions. The anonymized derived-feature dataset containing the S1-S21 feature matrix, binary KSC/NKSC labels, and anonymized user IDs used for grouped validation, together with the 80/20 train/test split indices and StratifiedGroupKFold (k = 5) fold assignments, is available at Mendeley Data (https://doi.org/10.17632/6y6fgjcysj.1). The full preprocessing, feature-construction, machine-learning training, ablation, and cross-user validation procedures used to reproduce Tables 7-14 are described in Pseudocodes 1-4 (Sections 2.2, 2.6, 2.7). Selected hyperparameter values are listed in Table 4 (reproducibility summary), and the random seed used throughout is 42.

Conflicts of Interest

The authors declare no conflict of interest.

Author Contributions

Conceptualization, Yuliazmi; Methodology, Yuliazmi and Diana Purwitasari; Software, Yuliazmi and Cornelius Bagus Purnama Putra; Validation, Yuliazmi and Cornelius Bagus Purnama Putra; Data Curation, Yuliazmi and Cornelius Bagus Purnama Putra; Writing—Original Draft Preparation, Yuliazmi and Safitri Juanita; Writing—Review and Editing, Yuliazmi, Diana Purwitasari and Agus Budi Raharjo; Visualization, Yuliazmi; Supervision, Diana Purwitasari, Surya Sumpeno and Mauridhi Hery Purnomo;

Acknowledgments

This work was supported by Indonesian Ministry of Research and Technology under Research Grant No. 3/E1/KP.PTNBH/2021 with an institutional contract No 955/PKS/ITS/2021

References

[1] R. D. Kusumastuti, N. Nurmala, J. Rouli, and H. Herdiansyah, "Analyzing the factors that influence the seeking and sharing of information on the smart city digital platform: Empirical evidence from Indonesia", *Technology in Society*, Vol.68, p.101876, 2022, doi: 10.1016/j.techsoc.2022.101876

[2] H. Alshahrani and D. Pennington, "'Maybe we can work together': researchers' outcome expectations for sharing knowledge on social media", *Global Knowledge, Memory and Communication*, Vol.70, No.4–5, pp.377–398, 2020, doi: 10.1108/GKMC-07-2020-0093

[3] O. Oyinlola, "Social media usage among older adults: Insights from Nigeria", *Activities, Adaptation & Aging*, Vol.46, No.4, pp.343–373, 2022, doi: 10.1080/01924788.2022.2044975

[4] Y. Cai, S. Kamarudin, and S. Nujaimi, "Willingness to share information on social media: a systematic literature review (2020–2024)", *Frontiers in Psychology*, Vol.16, p.1567506, 2025, doi: 10.3389/fpsyg.2025.1567506

[5] O. L. Haimson, A. J. Carter, S. Corvite, B. Wheeler, L. Wang, T. Liu, and A. Lige, "The major life events taxonomy: Social readjustment, social media information sharing, and online network separation during times of life transition", *Journal of the Association for Information Science and Technology*, Vol.72, No.7, pp.933–947, 2021, doi: 10.1002/asi.24455

[6] S. Oh and S. Y. Syn, "Motivations for sharing information and social support in social media: A comparative analysis of Facebook, Twitter, Delicious, YouTube, and Flickr", *Journal of the Association for Information Science and Technology*, Vol.66, No.10, pp.2045–2060, 2015, doi: 10.1002/asi.23320

[7] S. Pov, N. Kawai, and R. Murakami, "Exploring factors influencing teacher self-efficacy in implementing inclusive education in Cambodia: A two-level hierarchical linear model", *Studies in Educational Evaluation*, Vol.83, p.101415, 2024, doi: 10.1016/j.stueduc.2024.101415

[8] L. Luo, Y. Wang, S. Duan, S. Shang, B. Ma, and X. Zhou, "Continuous knowledge contribution in social Q&A communities: the moderation effects of self-presentation and motivational affordances", *Information Technology & People*, Vol.37, No.5, pp.1950–1982, 2023, doi: 10.1108/ITP-02-2022-0128

[9] S. Kaiser, D. Klare, M. Gomez, N. Ceballos, S. Dailey, and K. Howard, “A comparison of social media behaviors between sexual minorities and heterosexual individuals”, *Computers in Human Behavior*, Vol.116, p.106638, 2021, doi: 10.1016/j.chb.2020.106638

[10] V. C. Gever, F. O. Talabi, O. Adelabu, B. O. Sanusi, and J. M. Talabi, “Modeling predictors of COVID-19 health behaviour adoption, sustenance and discontinuation among social media users in Nigeria”, *Telematics and Informatics*, Vol.60, p.101584, 2021, doi: 10.1016/j.tele.2021.101584

[11] Sherani, J. Zhang, M. Riaz, F. A. Boamah, and S. Ali, “Harnessing technological innovation capabilities by the mediating effect of willingness to share tacit knowledge: a case from Pakistani software SMEs”, *Kybernetes*, Vol.52, No.12, pp.6590–6616, 2023, doi: 10.1108/K-09-2021-0845

[12] X. Feng, L. Wang, Y. Yan, Q. Zhang, L. Sun, J. Chen, and Y. Wu, “The sustainability of knowledge-sharing behavior based on the theory of planned behavior in Q&A social network community”, *Complexity*, Vol.2021, pp.1–12, 2021, doi: 10.1155/2021/1526199

[13] A. Malik, K. Mahmood, and T. Islam, “Understanding the Facebook users' behavior towards COVID-19 information sharing by integrating the theory of planned behavior and gratifications”, *Information Development*, Vol.39, No.4, pp.750–763, 2023, doi: 10.1177/02666669211049383

[14] H. Al-Mawali and K. A. Al-Busaidi, “Knowledge sharing through enterprise social media in a telecommunications context”, *International Journal of Knowledge Management*, Vol.18, No.1, pp.1–27, 2022, doi: 10.4018/IJKM.291706

[15] A. Malik, M. T. T. De Silva, P. Budhwar, and N. R. Srikanth, “Elevating talents' experience through innovative artificial intelligence-mediated knowledge sharing: Evidence from an IT-multinational enterprise”, *Journal of International Management*, Vol.27, No.4, p.100871, 2021, doi: 10.1016/j.intman.2021.100871

[16] S. Kwayu, M. Abubakre, and B. Lal, “The influence of informal social media practices on knowledge sharing and work processes within organizations”, *International Journal of Information Management*, Vol.58, p.102280, 2021, doi: 10.1016/j.ijinfomgt.2020.102280

[17] H. Keshavarz, “Personality factors and knowledge sharing behavior in information services: the mediating role of information literacy competencies”, *VINE Journal of Information and Knowledge Management Systems*, Vol.52, No.2, pp.186–204, 2021, doi: 10.1108/VJIKMS-05-2020-0095

[18] D. Shwartz-Asher, S. Chun, N. R. Adam, and K. LG. Snider, “Knowledge sharing behaviors in social media”, *Technology in Society*, Vol.63, p.101426, 2020, doi: 10.1016/j.techsoc.2020.101426

[19] N. Rijati, D. Purwitasari, S. Sumpeno, and M. Purnomo, “A decision making and clustering method integration based on the theory of planned behavior for student entrepreneurial potential mapping in Indonesia”, *International Journal of Intelligent Engineering and Systems*, Vol.13, No.4, pp.129–144, 2020, doi: 10.22266/ijies2020.0831.12

[20] W. Qi, B. Shen, Y. Yang, and X. Qu, “Modeling drivers' scrambling behavior in China: An application of theory of planned behavior”, *Travel Behaviour and Society*, Vol.24, pp.164–171, 2021, doi: 10.1016/j.tbs.2021.03.008

[21] C.-H. Liao, “Exploring social media determinants in fostering pro-environmental behavior: insights from social impact theory and the theory of planned behavior”, *Frontiers in Psychology*, Vol.15, p.1445549, 2024, doi: 10.3389/fpsyg.2024.1445549

[22] L.-Y. Leong, T.-S. Hew, K.-B. Ooi, B. Metri, and Y. K. Dwivedi, “Extending the theory of planned behavior in the social commerce context: A meta-analytic SEM (MASEM) approach”, *Information Systems Frontiers*, Vol.25, No.5, pp.1847–1879, 2023, doi: 10.1007/s10796-022-10337-7

[23] A. F. A. Sayed, M. A. Arafa, N. A. El-Nimr, K. A. S. Banawan, and M. S. Abdou, “Comparison of machine learning classification and regression models for prediction of academic performance among postgraduate public health students”, *Scientific Reports*, Vol.15, No.1, p.44056, 2025, doi: 10.1038/s41598-025-31023-z

[24] T. A. Khan, R. Sadiq, Z. Shahid, M. M. Alam, and M. B. M. Su'ud, “Sentiment analysis using support vector machine and random forest”, *Journal of Informatics and Web Engineering*, Vol.3, No.1, pp.67–75, 2024, doi: 10.33093/jiwe.2024.3.1.5

[25] Z. Wu, Y. Liao, C. Luo, J. Shi, and Y. Yang, “Predicting emerging trends: a machine learning approach to topic popularity on social media”,

PeerJ Computer Science, Vol.11, p.e3245, 2025, doi: 10.7717/peerj-cs.3245

[26] H. Jelodar, Y. Wang, R. Orji, and S. Huang, "Deep sentiment classification and topic discovery on novel coronavirus or COVID-19 online discussions: NLP using LSTM recurrent neural network approach", *IEEE Journal of Biomedical and Health Informatics*, Vol.24, No.10, pp.2733–2742, 2020, doi: 10.1109/JBHI.2020.3001216

[27] M. Kim and R. M. Fernandez, "What makes weak ties strong?", *Annual Review of Sociology*, Vol.49, No.1, pp.177–193, 2023, doi: 10.1146/annurev-soc-030921-034152

[28] N. C. Krämer, V. Sauer, and N. Ellison, "The strength of weak ties revisited: Further evidence of the role of strong ties in the provision of online social support", *Social Media + Society*, Vol.7, No.2, 2021, doi: 10.1177/20563051211024958