10.48047/jocaaa.2023.31.04.51

NLP-BASED INTER AND INTRA-SENTENCE RELATIONSHIP ANALYSIS-AWARE BANK CUSTOMER BEHAVIOR ANALYSIS AND PREFERENCE DETECTION USING GLSNSTM

Rajesh Kumar Kanji1, Vinodkumar Reddy Surasani2 Naveen Kumar Kotha3 and Uday Kiran Chilakalapalli4

1Individual researcher

2 Sr Software Engineer, RBC Wealth Management, MN, USA

3 Senior Staff Software Engineer

4 Department of Data Science and Analytics. Georgia State University, Atlanta, GA, USA

Abstract: Email and Chat logs are crucial information for examining customer transactions and communication patterns to understand their preferences. The existing studies didn't analyze inter and intra-sentence relationship analysis between the chat logs/email for accurate bank customer preference detection. Therefore, this paper presents an NLP-enabled bank customer behavior analysis and preference detection system using the Gating adaptive blending unit Long Schatten-p Norm Short Term Memory (GLSNSTM). Initially, the unstructured chat logs/email data are gathered, followed by structure conversion. Then, the structured data are pre-processed based on steps like abbreviation expansion, tokenization, atop word removal, spelling correction, stemming, and PoS tagging by HKNSMM. Next, the code-mixed languages are translated to English based on LangId. Thereafter, the RST is employed to analyze inter and intra-sentence relationships between language-translated data. Afterward, the grammatical roles of the words are identified using DP. From the DP outcomes, the essential keywords are extracted. Then, the bank customer preference is detected using Yule Jeffreys K-Means (YJK-Means). Subsequently, the behavior of bank customers is analyzed employing GLSNSTM for deciding whether preference access should be provided to the customer or not. If the behavior is normal, then preferences are provided; otherwise, access is denied. Also, Zak's SHapley Additive exPlanation (Za-SHAP)-based deepxplainer is employed to provide explanations. The results proved that the proposed methodology achieved a high accuracy of 98.76%.

Keywords: *Natural Language Processing (NLP), Client data analysis, Part of Speech (PoS) tagging, Dependency Parsing (DP), Rhetorical Structure Theory (RST), Hidden Kneser–Ney Smoothing Markov Model (HKNSMM), Inter and Intra-sentence relationship analysis, and Syntactic structure analysis.*

1. INTRODUCTION

In the digital banking era, understanding customer preferences and identifying behavioral patterns are significant for providing personalized services, ensuring security, enhancing customer satisfaction, and improving operational efficiency (Zouari & Abdelhedi, 2021). Banks can offer services to customers by identifying their needs, including intent in credit cards, debit cards, loans, digital banking, and so on (Wang et al., 2021). Likewise, behavior analysis aids in identifying abnormal patterns that represent fraudulent activities or financial distress. In the U.S., banks are progressively interacting with customers via digital platforms like email and chat log systems (Venkateswararao et al., 2023). Chat logs express emotional prompts, linguistic anomalies, or abrupt topic shifts. Likewise, emails and chat logs consist of early signs of unusual activity, including repeated complaints, contradictory statements, and sudden changes in tone (Katsafados et al., 2021). U.S. banks employ these platforms for customer service, transaction queries, product inquiries, and complaint handling. Analyzing these data gives an important opportunity to understand customer needs, preferences, and behaviors more accurately (Kaur et al., 2020).

10.48047/jocaaa.2023.31.04.51

Nevertheless, the unstructured nature of email and chat logs leads to many challenges in data analysis. Therefore, NLP techniques are established to detect these patterns in real-time and take appropriate actions (Olujimi & Ade-Ibijola, 2023). NLP techniques extract meaningful information from email/chat logs by recognizing keywords, emotions, intents, and conversational patterns for predicting customer preferences (Mashaabi et al., 2022). However, the traditional models poorly analyzed the syntactic structure of sentences, leading to suboptimal outcomes. In Banks of America, customer satisfaction, security, and personalization are important. Commonly, Artificial Intelligence (AI) capabilities, such as NLP, predictive analytics, machine learning, and deep learning algorithms, empower financial firms to obtain deep insights into customer behavior and give proactive solutions (Egbuhuzor et al., 2021). Nowadays, NLP approaches, such as Named Entity Recognition (NER), sentiment analysis, and Bidirectional Encoder Representation from Transformers (BERT), capture both meaning and intent behind bank client comments (Örpek et al., 2023).

In existing studies, document term frequency analysis-based key terminology extractions and clustering method-based subdomain detection were performed for effective customer behavior analysis (Chao et al., 2021). Similarly, the Latent Dirichlet Allocation (LDA) technique was used to perform topic modeling for client data (Papadia et al., 2022). Likewise, Machine Learning techniques, including Naïve Bayes (NB), K-Nearest Neighbor (KNN), decision tree, random forest, and support vector machine, were employed to analyze customer behavior (Maheswari et al., 2021) (Elrefai et al., 2021). Also, some prevailing works used Back Propagation Neural Network (BPNN) to analyze the behavioral patterns of customers (Xiong et al., 2021). Likewise, the Deep Neural Network (DNN) algorithm was employed by conventional methods for analyzing customer behavior/ customer churn prediction (Domingos et al., 2021). However, in existing studies, the absence of inter and intra-sentence relationship analysis led to inaccurate customer preference detection. To conquer this problem, a novel NLP-based inter and intra-sentence relationship analysis-aware bank customer behavior analysis and preference detection using GLSNSTM is proposed in this article.

1.1 Problem Statement

- None of the existing works identified inter and intra-sentence relationships between the chat log/emails, leading to inaccurate customer preference detection.
- Conventional (Kumar et al., 2022) didn't effectively handle multi-lingual texts in conversations. Global and local bank customers (Bank of America) might mix languages; thus, the models failed to generalize across such data.
- In existing (Rustamov et al., 2021), the lack of syntactic structure analysis resulted in shallow keyword extraction, missing the right intent behind flexible sentence patterns in chat logs. Due to the poor understanding of grammatical roles, the model misclassified phrases by concentrating only on the presence of words. Thus, the accuracy of customer preference detection was reduced.
- The conventional (Vo et al., 2021) failed to identify the correlation between the high-impact behavioral features, leading to suboptimal generalization and reduced effectiveness for customer behavior analysis.
- In existing works, the accuracy of customer preference detection was affected due to misspelling words and poor pre-processing.

10.48047/jocaaa.2023.31.04.51

- Most of the DL models for customer behavior analysis had black-boxes (i.e., they had no transparency in why decisions were made), leading to poor model trust.

1.2 Objective

- ❖ RST is employed to analyze inter and intra-sentence relationships between the chat log/emails for accurate customer preference detection.
- ❖ Language Identification (LangId)-based predefined Python library is used to handle the code-mixed languages.
- ❖ DP is introduced to analyze the syntactic structure of the sentences.
- ❖ Pearson Correlation Coefficient (PCC) is utilized to identify the correlation between the high-impact behavioral features.
- ❖ Essential pre-processing steps like tokenization, spelling correction, POS tagging, and so on are performed for accurate customer preference detection.
- ❖ Za-SHAP is established to provide explanations about the prediction outcomes.

This paper is structured as: Section 2 conveys the literature survey, Section 3 illustrates the proposed research model, Section 4 elucidates the result and discussion, and finally, Section 5 concludes the proposed model with future scope.

2. LITERATURE SURVEY

(Kumar et al., 2022) presented a model for detecting financial crimes regarding customer behavior. In this work, machine learning techniques, including decision tree, random forest, and k-nearest neighbour, were employed to analyze customer behavior. The presented model achieved the highest accuracy score, f-score, precision, and recall than the prevailing methods. Also, it aided as the best model for bankers to better understand the customer's behavior. However, the model failed to handle the multi-lingual texts in conversations, making the model poorly generalizable.

(Rustamov et al., 2021) developed a dialogue management model for the banking sector. In this work, the languages of the dialogue were identified by using the LangId component. Afterward, Named Entity Recognition (NER) was employed to extract essential information from the messages. Subsequently, user intention was detected by using logistic regression, neural networks, and support vector machine. The model excellently identified the preferences of bank customers. Owing to the poor understanding of grammatical roles, the model misclassified phrases by concentrating only on the presence of words.

(Vo et al., 2021) employed unstructured call logs for customer churn prediction. In this work, Naïve Bayes, logistic regression, random forest, and extreme gradient boosting along with the SHAP algorithm were employed for predicting the customer churn by analyzing their behavior. The research accurately predicted the client churn risks and generated meaningful insights. Nevertheless, the model failed to identify the correlation between the high-impact behavioral features, leading to reduced effectiveness.

(Andrian et al., 2022) established a model for sentiment analysis on customer satisfaction with digital banking. In this work, nine standalone classifiers like k-nearest neighbours, random forest, Naïve Bayes, decision tree, adaptive boosting, logistic regression, extreme gradient boosting, support vector machines, and light gradient boosting machine were

10.48047/jocaaa.2023.31.04.51

employed for sentiment analysis. The model excellently identified the sentiment as positive, negative, and neutral. Yet, it had a lack of transparency, leading to poor model trust.

(Ibitoye & Onifade, 2022) suggested a framework named an improved customer churn prediction system. In this work, the word order contextualized semantic technique was employed for predicting customer churn. The experimental outcomes proved that the model achieved high accuracy, precision, and recall values. However, in this work, the accuracy of customer churn prediction was hindered due to the misspelling of words and poor pre-processing.

(Zhang et al., 2020) propounded a model for personalized digital customer services. In this work, the multi-task neural network was employed to identify the behavior of bank customers**.** In this model, the banks provided a more effective customer service experience. Yet, the research struggled to handle the diverse accents, languages, background noise, and informal speech. Thus, the performance of the model was hindered.

(Mishev et al., 2020) investigated the effectiveness of several sentiment analysis models based on combinations of text representation algorithms and machine-learning techniques. Here, a specific lexicon was employed to perform sentiment analysis in finance. Furthermore, the Natural Language Processing (NLP) transformers were introduced to encode words and sentences. This framework had high preciseness. Nevertheless, this model struggled to handle the large datasets owing to the lack of effective NLP schemes.

(Ogunleye et al., 2023) introduced the topic modeling process in the banking context. In this work, Kernel Principal Component Analysis (KernelPCA) and K-means Clustering in the BERTopic were employed to perform topic modelling for tweets of customers. Hence, the model achieved a high coherence score and accuracy. However, the model struggled to perform topic modelling for the short texts.

(Gavval & Ravi, 2020) presented a model for clustering bank customer complaints for customer relationship management. In this work, the Multi-objective Particle Swarm Optimization along with heuristics of K-means was employed to cluster the bank customer complaints for improving the banking services. The model achieved promising results in analyzing customer behaviors. Yet, the model consumed more time and had computational complexity for large-scale social media data.

(Abbasimehr & Shabani, 2019) developed a model for customer behavior analysis. For extracting the dominant behavioral patterns of customers over time, the time series clustering approach was employed. Regarding the monetary attribute, the clients were split into high-value growing customers, middle-value growing customers, and prone to churn. The model attained improved performance outcomes. However, the model eliminated the external events that affected customer behavior.

3. PROPOSED NLP-BASED INTER AND INTRA-SENTENCE RELATIONSHIP ANALYSIS-AWARE BANK CUSTOMER BEHAVIOR ANALYSIS AND PREFERENCE DETECTION METHODOLOGY

Here, the proposed model excellently detected the preferences and analyzed the behavior of United States (US) bank clients from chat logs/email data using NLP techniques. The proposed YJK-Means is introduced to detect the U.S. bank customer preferences. Also, for analyzing inter and intra-sentence relationships, the RST is employed. Likewise, the proposed GLSNSTM is established to analyze the U.S. bank customer behavior. For performing PoS

10.48047/jocaaa.2023.31.04.51

tagging, the proposed HKNSMM is used. Similarly, a deep explanation of the prediction outcomes is provided by the proposed Za-SHAP. Also, the DP is utilized to analyze the syntactic structure of the sentences. The pictorial representation of the proposed model is shown in Figure 1,

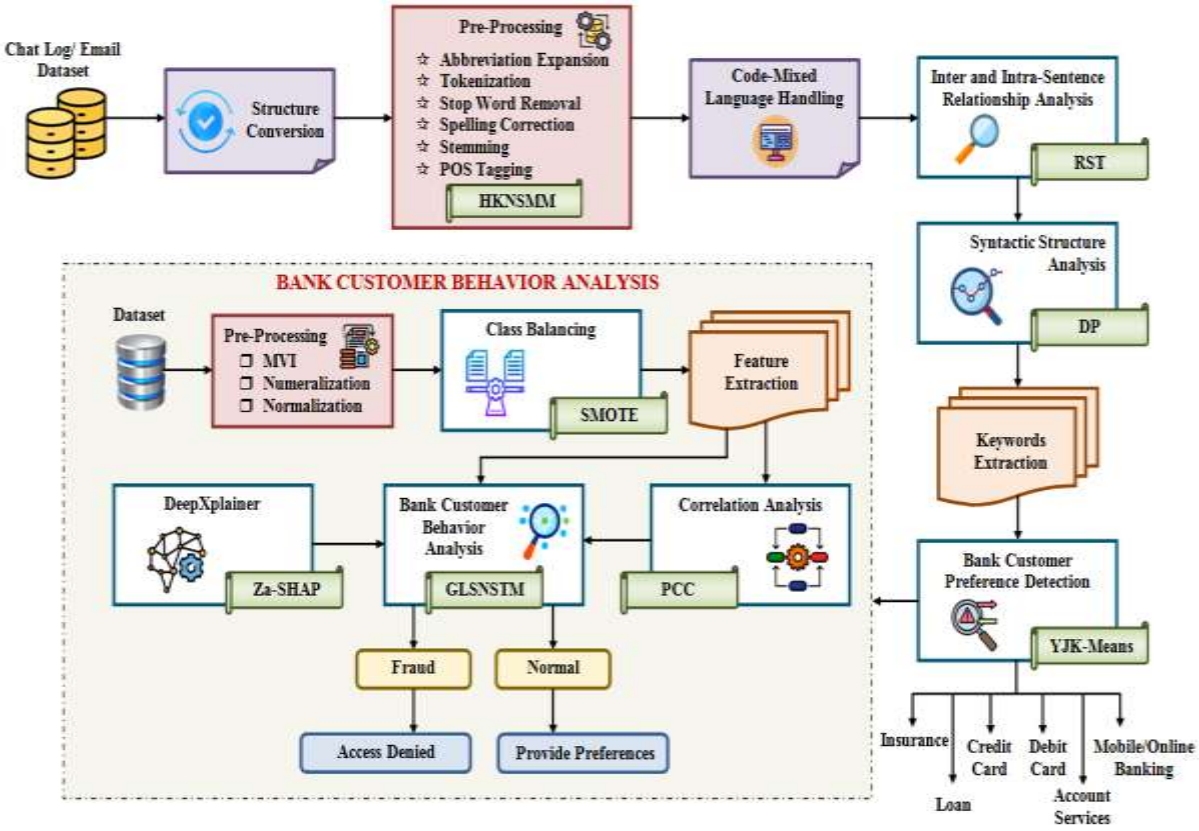


Figure 1: Pictorial representation of the proposed model

The proposed model comprises significant processes, such as pre-processing, code-mixed language handling, inter and intra-sentence relationship analysis, syntactic structure analysis, keyword extraction, customer preference detection, and ban customer behavior analysis. The working process of the proposed model is explained below,

3.1 Chat log/ Email Dataset

Primarily, the unstructured data like chat log/email datasets are gathered and analyzed for detecting the preferences of U.S bank clients. Actually, a chat log/email consists of linguistic and contextual cues that convey the client's preferences and sentiments. The chat log/email data are indicated as Q_i .

3.2 Structure Conversion

Afterward, for analyzing and processing the data effectively, the unstructured chat log/email data (Q_i) is converted into a structured format using a Python library named 'python-docx'. The structured chat log/email data are specified as Ω_κ .

3.3 Pre-Processing

10.48047/jocaaa.2023.31.04.51

Then, Ω_κ are pre-processed to improve the quality of the data and prepare the data for further analysis. Firstly, the abbreviations present in the Ω_κ are expanded and are denoted as α. After that, in the tokenization process, α are broken into small units called tokens. The tokenized data are signified as τk. Subsequently, the stop words (i.e., 'is', 'and', 'for', 'of', etc) are removed from the τk. It is written as,

$$\tau k \xrightarrow{remove\ stop\ words} \varsigma\omega\gamma \tag{1}$$

Here, $\varsigma\omega\gamma$ implies the stop words removed data. Next, misspelled words in $\varsigma\omega\gamma$ are detected and corrected and are specified as Msp. Afterward, in the stemming process, Msp are reduced to their base or root form by eliminating the suffixes or prefixes. The stemming outcomes are given as,

$$Msp \xrightarrow{reduce\ to\ root} Rt \tag{2}$$

Where, Rt defines the stemming outcomes. Thereafter, the PoS tagging is performed for Rt to reveal the function of words in context, aiding NLP systems in understanding sentence structure. Here, the HKNSMM algorithm is employed for PoS tagging. In general, the Hidden Markov Model (HMM) excellently tags PoS owing to its probabilistic nature. Also, HMM assigns probabilities to sequences instead of relying on fixed rules, thus making the model more flexible in handling unseen sentences. Nevertheless, HMM depends on predefined probability distributions that don't adapt to new data. To solve this problem, the Kneser–Ney Smoothing technique is used, which computes the probability distribution of a language model and is useful for dealing with unseen word sequences. The working process of HKNSMM is explained below,

In the first step, the components of HMM, including hidden states and observations, are estimated. It is expressed as,

$$H_s \xrightarrow{Rt} \{H_1, H_2, H_3, \ldots\ldots, H_q\} \tag{3}$$

$$Ob_\nu \xrightarrow{Rt} \{Ob_1, Ob_2, Ob_3, \ldots\ldots, Ob_\kappa\} \tag{4}$$

Here, H_s signifies the hidden states (i.e., PoS tags), H_q is the total number of hidden states, Ob_ν denotes the observations (i.e., sequence of words), and Ob_κ exemplifies the κ^{th} observations. Thereafter, the probability at the initial state is calculated. Here, the Kneser–Ney Smoothing technique is employed to compute the probability of events. Kneser–Ney Smoothing technique effectively deals with unseen word sequences. It is mathematically expressed as,

$$I_{pb} = \frac{max\left(ct\left(\left(Ob_\nu\right)_{T-1}, \left(Ob_\nu\right)_T\right) - D, 0\right)}{ct\left(\left(Ob_\nu\right)_{T-1}\right)} + bf\left(\left(Ob_\nu\right)_{T-1}\right) Uq_{cont}\left(\left(Ob_\nu\right)_T\right) \tag{5}$$

Here, I_{pb} specifies the initial state probability, $ct\left(\left(Ob_\nu\right)_{T-1}, \left(Ob_\nu\right)_T\right)$ denotes the count of a bigram, D implies the discount constant, $bf\left(\left(Ob_\nu\right)_{T-1}\right)$ indicates the backoff weight, $Uq_{cont}\left(\left(Ob_\nu\right)_T\right)$ specifies the number of unique contexts that appear in total unique bigrams, and $ct\left(\left(Ob_\nu\right)_{T-1}\right)$ exemplifies the count of prefix unigram. After that, the emission probability (εm) and state transition probability $(\tau\rho)$ are computed and formulated as,

10.48047/jocaaa.2023.31.04.51

$$\tau\rho = \Xi\left(H_{k+1} = H_w \mid H_k = H_\mu\right) \tag{6}$$

$$\varepsilon m = \Xi\left(Ob_k = sp_\varphi \mid H_k = H_w\right) \tag{7}$$

Where, H_{k+1} implies the hidden states at $k+1^{th}$ time step, H_k exemplifies the hidden states at k^{th} time step, H_w denotes the w^{th} hidden state, H_μ demonstrates the μ^{th} hidden state, Ξ depicts the probability factor, Ob_k indicates the observation at k^{th} time step, and sp_φ denotes the specific observation. Then, the probability score of the word's tags is obtained. Afterward, the probability scores of each word's tags are compared. Eventually, the most effective PoS tag is computed and is denoted as $\wp_u$. The pseudocode for HKNSMM is given as follows,

Pseudocode for HKNSMM

Input: Stemming outcomes $\left(Rt\right)$
Output: PoS tags $\left(\wp_u\right)$

Begin
Initialize $\left(Rt\right)$
For each $\left(Rt\right)$
Estimate hidden state and observations

$$H_s \xrightarrow{Rt} \left\{H_1, H_2, H_3, \ldots\ldots, H_q\right\}$$

$$Ob_v \xrightarrow{Rt} \left\{Ob_1, Ob_2, Ob_3, \ldots\ldots, Ob_\kappa\right\}$$

Discover initial state probability by Kneser–Ney Smoothing technique

$$I_{pb} = \frac{max\left(ct\left(\left(Ob_v\right)_{T-1}, \left(Ob_v\right)_T\right) - D, 0\right)}{ct\left(\left(Ob_v\right)_{T-1}\right)} + bf\left(\left(Ob_v\right)_{T-1}\right) Uq_{cont}\left(\left(Ob_v\right)_T\right)$$

Compute emission probability

$$\varepsilon m$$

Find state transition probability

$$\tau\rho = \Xi\left(H_{k+1} = H_w \mid H_k = H_\mu\right)$$

Compare probability scores
End For
Obtain PoS tags $\left(\wp_u\right)$
End

Thus, the PoS tags are effectively computed by the proposed HKNSMM. The pre-processed chat log/email data (PoS tags) is denoted as $\ddot{\chi}_\varsigma$.

3.4 Code-Mixed Language Handling

Next, the code-mixed languages in $\ddot{\chi}_\varsigma$ are detected by using a LangId-based predefined library for improving the U.S. bank customer preference detection. In general, LangId identifies the language of $\ddot{\chi}_\varsigma$, allowing seamless handling of multiple languages. If $\ddot{\chi}_\varsigma$ has words from more than one language, then it is translated into English language. The language-translated data is expressed as,

10.48047/jocaaa.2023.31.04.51

$$\ddot{\chi}_{\varsigma} \xrightarrow{translate} G \tag{8}$$

Where, G denotes the language-translated data.

3.5 Inter and Intra-Sentence Relationship Analysis

From the language-translated data G, the inter and intra-sentence relationship is identified for accurate U.S. bank client preference detection. Here, the RST method is utilized to analyze the inter and intra-sentence relationship between the G. RST enables discourse-level analysis by modeling relationships between multiple sentences or clauses.

Firstly, G are broken into Elementary Discourse Units (EDUs). Next, the rhetorical relations (e.g., cause, elaboration, contrast) are identified between the EDUs. In the next step, the nucleus (i.e., main idea) and satellite (i.e., supporting idea) in each rhetorical relation are identified. Then, a hierarchical tree structure that combines EDUs regarding the identified rhetorical relations is generated. It is expressed as,

$$G \xrightarrow{constructed} \hbar^{trr} \tag{9}$$

Where, $\hbar^{trr}$ indicates the hierarchical tree. From the obtained nucleus $(n\mu)$ and satellite words $(s\alpha)$, the nucleus is chosen for further processing.

3.6 Syntactic Structure Analysis

Next, by using DP, the syntactic structure (i.e., grammatical role) of the words in $(n\mu)$ is identified for extracting the depth keywords. In general, dependency parsing has the ability to provide word-to-word relations. Therefore, it is suitable for complex sentence structures. The working of DP is explained as,

Firstly, in DP, the grammatical representation is assigned to the words in $(n\mu)$. In the next step, for each word, the head and dependencies between the words are identified. Thereafter, the dependency tree is created in which the central node is the main verb (i.e., Root) and other words rely on it. The constructed dependency tree is given as,

$$n\mu \xrightarrow{to} \Re_{tree} \tag{10}$$

Where, $\Re_{tree}$ implies the constructed dependency tree. Then, from $\Re_{tree}$, the keywords essential for U.S. bank customer preference detection are extracted.

3.7 Keywords Extraction

After that, the keywords, including Head-dependent pairs (i.e., root and its main object), are extracted from the constructed dependency tree $(\Re_{tree})$. The extracted keywords are formulated as,

$$K_{\ell}(\Re_{tree}) \rightarrow \langle K_1, K_2, K_3, \ldots\ldots, K_{\lambda} \rangle \quad Here \quad \ell = (1,2,\ldots,\lambda) \tag{11}$$

10.48047/jocaaa.2023.31.04.51

Where, (K_ℓ) is the extracted keywords and $\ell=(1,2,\ldots,\lambda)$ denotes the total number of extracted keywords. Next, the extracted keywords (K_ℓ) are given for U.S. bank customer preference detection.

3.8 Bank Customer Preference Detection

Based on the extracted keywords (K_ℓ), the U.S. bank customer preference is detected by using YJK-Means. Normally, the K-Means algorithm has the capability to handle large datasets with high dimensionality. Likewise, K-Means make sure that each data is uniformly distributed within each cluster. Yet, K-Means struggled to cluster nodes, where clusters are of varying density. To address this problem, the Yule distance is introduced instead of Euclidean distance. Also, during the random initialization of centroids, K-Means can converge to different solutions. In order to overcome this problem, Jeffreys Entropy is utilized to initialize the centroid of K-Means. Here, firstly, based on the keywords related to bank customer preferences, the YJK-Means groups the preferences. Here, the keywords are set to a centroid. During the testing time, (K_ℓ) are grouped to relevant preferences. The step-by-step mathematical expression of the proposed YJK-Means is described below,

In the first step, the centroids are initialized by using Jeffreys Entropy $(J\varepsilon)$, which improves the clustering process.

$$J\varepsilon = Kl(K_1 \,||\, K_2) + Kl(K_2 \,||\, K_1) \tag{12}$$

$$\Phi_\varpi \xrightarrow{J\varepsilon} (\Phi_1 + \Phi_2 + \Phi_3 + \ldots\ldots + \Phi_g) \quad Here \quad \varpi = (1,2,\ldots,g) \tag{13}$$

Where, Kl depicts the Kullback–Leibler divergence, Φ_ϖ represents the initialized centroids, and $\varpi=(1,2,\ldots,g)$ denotes the number of Φ_ϖ. Next, the distance between the Φ_ϖ and (K_ℓ) is computed using Yule distance and is formulated as,

$$Y(\Phi_\varpi, K_\ell) = \frac{2\Phi_1 K_1}{\Phi_2 K_2 + \Phi_1 K_1} \tag{14}$$

Here, Y indicates Yule distance. According to the Y, the points are assigned to the cluster with the closest centroid. After that, based on the average position of all points in the cluster, the new centroid (F) is estimated.

$$F = \frac{\sum_{\varpi=1,\ell=1}^{g,\lambda} \gamma_{\varpi\ell} K_\ell}{\sum_{\varpi=1,\ell=1}^{g,\lambda} K_\ell} \tag{15}$$

10.48047/jocaaa.2023.31.04.51

Where, $\gamma_{\varpi\ell}$ indicates the average position. Thereafter, all nodes are reassigned to the new closest centroid of each cluster. The above steps are followed until convergence is achieved. The detected bank customer preference (BP) are signified as,

$$BP = (L, Cc, Dc, Is, Om, Ac) \tag{16}$$

Where, L indicates the loan service, Cc denotes the credit card service, Dc implies the debit card service, Is is the insurance service, Om exemplifies online banking/mobile banking services, and Ac signifies account services.

3.9 Bank Customer Behavior Analysis

After detecting the preferences of U.S. bank customers, the behavior of bank customers (i.e., normal behavior, fraud behavior) is analyzed to decide whether preference access should be provided to the customer or not.

3.9.1 Dataset

Initially, the "Bank Account Fraud Dataset Suite" is collected from publicly available sources for training the bank customer behavior analysis system. This dataset comprises the transaction and behavior patterns of bank customers. The total ρ number of data (β_{∂}) in the "Bank Account Fraud Dataset Suite" is defined as,

$$\beta_{\partial} \Rightarrow [\beta_1, \beta_2, \beta_3, \ldots\ldots, \beta_{mn}] \tag{17}$$

Where, β_{mn} indicates the mn^{th} data in the "Bank Account Fraud Dataset Suite".

3.9.2 Pre-Processing

Next, (β_{∂}) are pre-processed for preparing the data for further analysis. Here, three important processes, namely Missing Value Imputation (MVI), numeralization, and normalization are performed. Firstly, the missing values in (β_{∂}) are imputed using the mean formula. The missing value imputed data (M_{im}) are defined as,

$$M_{im} = \frac{\sum \beta_{\partial}}{mn} \tag{18}$$

Thereafter, M_{im} are converted into numerical values in the numeralization process. The numeralized data is specified as ϑ. After that, ϑ are normalized in the range of 0 to 1 using the min-max normalization technique. It is expressed as,

$$\aleph = \frac{\vartheta - \vartheta_{min}}{\vartheta_{max} - \vartheta_{min}} \tag{19}$$

Where, $\aleph$ indicates the normalized data. The pre-processed data is specified as Rn_h.

3.9.3 Class Balancing

After that, the Rn_h is balanced using the Synthetic Minority Oversampling Technique (SMOTE) technique for avoiding errors. Normally, SMOTE has the ability to solve class

10.48047/jocaaa.2023.31.04.51

imbalance issues. In SMOTE, for oversampling the minority classes, the synthetic instances are created. The minority and majority classes are denoted as $R(Rn_h)$ and $S(Rn_h)$, correspondingly. Next, from $R(Rn_h)$, the minority instances (δ_d) are randomly selected. It is expressed as,

$$\delta_d = \{\delta_1, \delta_2, \delta_3, \ldots\ldots, \delta_e\} \tag{20}$$

Where, δ_e defines the number of minority instances. Subsequently, the nearest neighbor of the minority instances (δ_d) is discovered and is given as,

$$\eta\eta(\delta_d, R(Rn_h)) = \left[\sum_{d=1}^{e}(\delta_d - S(Rn_h))^2\right]^{\frac{1}{2}} \tag{21}$$

Here, $\eta\eta$ defines the nearest neighbor of (δ_d). After that, the random neighbor $(\eta\eta')$ is chosen and is indicated as $\eta\eta' \in R(Rn_h)$. In the next step, the new instance (nI) is created and is equated as,

$$nI = \delta_d + (\eta\eta' - \delta_d) \times rand \tag{22}$$

Where, $rand$ depicts the random number. Thereafter, nI is added to Rn_h for minority class. The balanced data (Ba_f) is written as,

$$Ba_f \rightarrow \lfloor Ba_1 + Ba_2 + Ba_3 + \ldots\ldots + Ba_{pq} \rfloor \tag{23}$$

Where, Ba_{pq} implies the number of balanced data.

3.9.4 Feature Extraction

From Ba_{pq}, features like income, name_email_similarity, prev_address_months_count, current_address_months_count, velocity_4w, email_is_free, customer_agedays_since_request, intended_balcon_amount, payment_type, zip_count_4w, velocity_6h, velocity_24h, bank_branch_count_8w, date_of_birth_distinct_emails_4w, employment_status, device_os, credit_risk_score, housing_status, phone_home_valid, phone_mobile_valid, bank_months_count, has_other_cards, proposed_credit_limit, foreign_request, source, month, session_length_in_minutes, keep_alive_session, device_distinct_emails_8w, and device_fraud_count are extracted. The total $\exists$ number of extracted features is given as,

$$\zeta_\upsilon \rightarrow \{\zeta_1, \zeta_2, \zeta_3, \ldots\ldots, \zeta_\exists\} \quad where \quad \upsilon = (1\ to\ \exists) \tag{24}$$

Where, ζ_υ is the number of extracted features.

3.9.5 Correlation Analysis

Next, the correlation between the ζ_υ are computed using PCC to avoid suboptimal generalization and improve the effectiveness of customer behavior analysis.

10.48047/jocaaa.2023.31.04.51

$$C_r = \frac{\sum(\zeta_1 - \bar{\zeta}_1)(\zeta_2 - \bar{\zeta}_2)}{\sqrt{\sum(\zeta_1 - \bar{\zeta}_1)^2 \sum(\zeta_2 - \bar{\zeta}_2)^2}} \quad (25)$$

Where, C_r indicates the correlation outcomes.

3.9.6 Customer Behavior Analysis

Based on ζ_υ and C_r, the U.S. bank customer behavior is analyzed by using GLSNSTM. Generally, Long Short Term Memory (LSTM) effectively handles varying length input sequences, making them flexible for real-world applications. Also, LSTMs employ cell states and gates for controlling gradient flow, which allows them to learn over long sequences without significant loss of information. However, LSTM had computational complexity due to the improper activation function. Likewise, it had overfitting issues due to the poor initialization of weights. To address these issues, the Schatten-p Norm regularization technique and Gating Adaptive Blending Unit (GABU) technique are employed in LSTM. The structural layout of the proposed GLSNSTM classifier is shown in Figure 2.

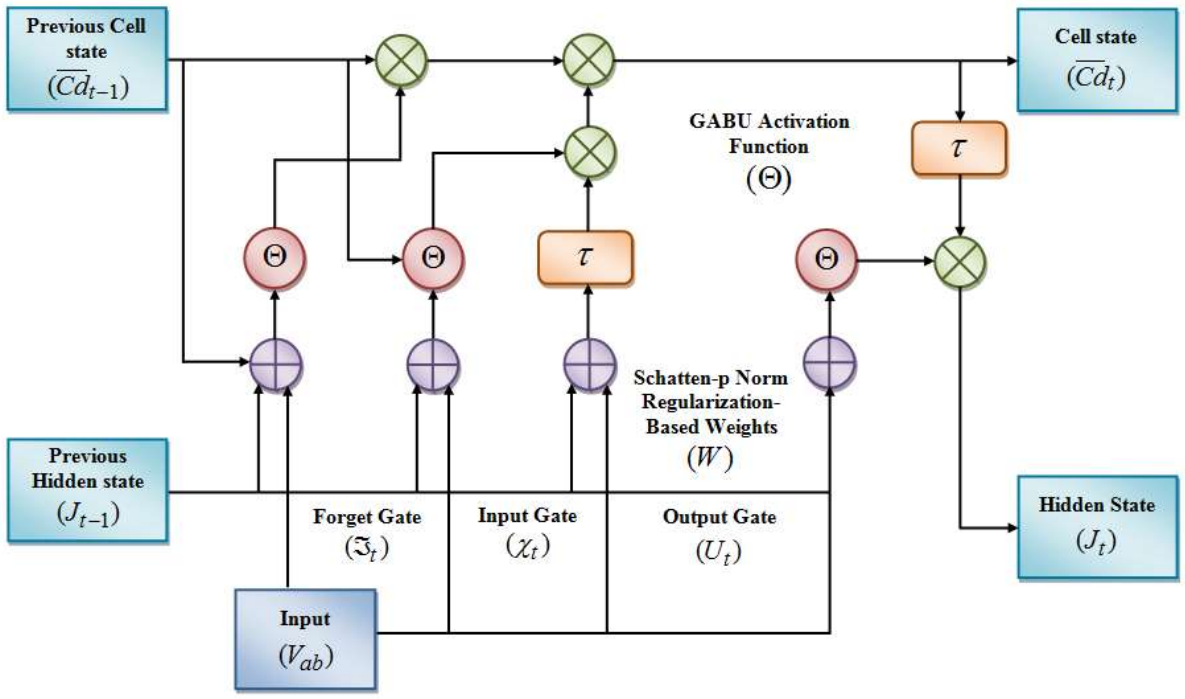


Figure 2: Structural layout of the proposed GLSNSTM classifier

Here, the GLSNSTM consists of a forget gate, input gate, candidate cell state, cell state, and output gate. The inputs ζ_υ and C_r are assumed as V_{ab} , which is expressed as,

$$\zeta_\upsilon, C_r \rightarrow V_{ab} \quad (26)$$

The step-by-step mathematical expression of the GLSNSTM classifier is explained below,

Step 1: Forget Gate

10.48047/jocaaa.2023.31.04.51

Forget gate is the first step of the GLSNSTM classifier. The forget gate removes the unwanted information from the gates and is mathematically defined as,

$$\Im_t = \Theta \times \left\langle \left\{ W \times \left| V_{ab} \cup J_{t-1} \right| \right\} + A \right\rangle \tag{27}$$

$$\Theta = \sum l\sigma(y) wg(V_{ab}) \tag{28}$$

$$W = \lambda \left(\sum (V_{ab})^j \right)^{\frac{1}{j}} \tag{29}$$

Where, $\Im_t$ specifies the forget gate operation, A denotes the bias term, J_{t-1} exemplifies the previous hidden state, Θ defines the GABU activation, which diminishes the computational complexity, W illustrates the Schatten-p Norm regularization-based weights, which solves the overfitting issue, $l\sigma$ signifies the logistic sigmoid function, y denotes the trainable parameter that controls the weight of activation function (wg), λ is the regularization parameter, and j implies the fractional value.

Step 2: Input Gate

In the following step, the input gate operation is done, which discovers the essential information that is required to pass via the gates. The input gate operation is written as,

$$\chi_t = \left\langle \left[\left\{ W \times \left| V_{ab} \cup J_{t-1} \right| \right\} + A \right] * \Theta \right\rangle \tag{30}$$

Here, χ_t exemplifies the input gate operation.

Step 3: Candidate Cell State

Subsequently, the candidate cell state (Cd_t) is computed as,

$$Cd_t = \left\langle \left[\left\{ W \times \left| V_{ab} \cup J_{t-1} \right| \right\} + A \right] * \tau \right\rangle \tag{31}$$

Where, τ denotes the hyperbolic tangent function.

Step 4: Cell State

After that, the cell state operation $\left(\overline{Cd}_t\right)$ is performed. Here, the cell state concatenates the forget gate and input gate. It is equated as,

$$\overline{Cd}_t = \left(\Im_t \bullet \overline{Cd}_{t-1} \right) + \left(\chi_t \bullet Cd_t \right) \tag{32}$$

Where, $\overline{Cd}_{t-1}$ illustrates the previous cell state and $\bullet$ demonstrates the element-wise multiplication process.

Step 5: Output Gate

Then, the decisions about the bank customer behavior analysis are made by the output gate (U_t). The output gate operation is provided as,

$$U_t = \left\langle \left[\left\{ W \times \left| V_{ab} \cup J_{t-1} \right| \right\} + A \right] * \Theta \right\rangle \tag{33}$$

10.48047/jocaaa.2023.31.04.51

$$J_t = U_t \bullet \Theta\left(\overline{Cd}_t\right) \tag{34}$$

Where, J_t specifies the present hidden state. The analyzed bank customer behavior (BA) is written as,

$$BA \to \langle N, \lambda \rangle \tag{35}$$

Where, N signifies the normal behavior and λ depicts the fraud behavior. The pseudo-code for GLSNSTM is depicted as,

Pseudocode for GLSNSTM

Input: Extracted Features (ζ_υ) and Correlation Outcomes (C_r)

Output: Analysed Bank Customer Behavior (BA)

Begin

Initialize (ζ_υ), (C_r), cell state $\left(\overline{Cd}\right)$, hidden state (J), and bias (A)

Assume (ζ_υ) and (C_r) as V_{ab}

For each V_{ab}

Estimate forget gate

$$\Im_t = \Theta \times \langle \{W \times |V_{ab} \cup J_{t-1}|\} + A \rangle$$

Discover $\Theta = \sum l\sigma(y) wg(V_{ab})$

Find Schatten-p Norm regularization-based weights (W)

Implement input gate operation

$$\chi_t$$

Compute candidate cell state

$$Cd_t = \langle [\{W \times |V_{ab} \cup J_{t-1}|\} + A] * \tau \rangle$$

Find cell state

$$\overline{Cd}_t$$

Perform output gate operation $U_t = \langle [\{W \times |V_{ab} \cup J_{t-1}|\} + A] * \Theta \rangle$

Estimate present hidden state

$$J_t = U_t \bullet \Theta\left(\overline{Cd}_t\right)$$

End For

Obtain $BA \to \langle N, \lambda \rangle$

End

Thus, the proposed GLSNSTM excellently performed U.S. bank customer behavior analysis. If the behavior is normal, the preferences are provided for the customer; otherwise, access is denied.

3.9.7 DeepXplainer

Then, the explanations about the analyzed U.S. bank customer behavior (BA) are provided by using Za-SHAP. Generally, SHapley Additive exPlanation (SHAP) provides insights into how input features influence prediction. However, the kernel function that is used for assigning the weight might not capture the underlying structure of the data, leading to

10.48047/jocaaa.2023.31.04.51

suboptimal results. To address this problem, Zak's function is employed in SHAP to assign the weight value.

In Za-SHAP, each feature is treated as a player in a game in which the goal is to detect the output. Regarding Shapley's values from cooperative game theory, the SHAP values (Shp) are generated. It is defined as,

$$Shp = \sum_{\beta s \subseteq \Im} \frac{|\beta s|!\left(\left\|\Im\right\| - |\beta s| - 1\right)!}{\left\|\Im\right\|!}\left[\left\|\Im\right|\left(\beta s \cup \{c\}\right) - \left(BA\right)\right] \tag{36}$$

Where, βs denotes the subset of features, $\Im$ depicts the full set of features, and c is the number of features. Here, Zak's function is used for assigning the SHAP values to features. The Zak's function excellently captures the underlying structure of the data. It is expressed as,

$$Shp^{*} = \left(\left(BA\right)^{2} + 10^{6}\, BA^{2}\right) \tag{37}$$

Lastly, the transformed SHAP values $\left(Shp^{*}\right)$ are visualized to indicate the essential global features. Thus, the proposed methodology excellently detected the U.S. bank customer preferences and analyzed the bank customer behavior.

4. RESULT AND DISCUSSION

Here, the performance analysis and comparative validation of the proposed and existing techniques are done to prove the effectiveness of the proposed model. Also, the proposed methodology is implemented in the working platform of PYTHON.

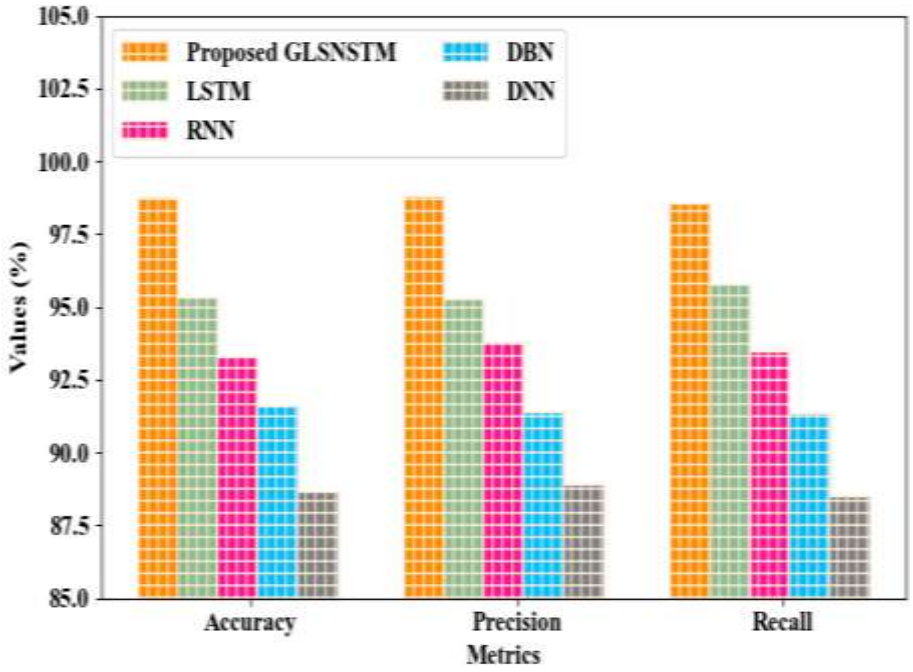
4.1 Dataset Description

The proposed model uses the "Chat logs/ Email" dataset, which is collected from publicly available sources for bank customer preference detection. Also, the proposed model employs the "Bank Account Fraud Dataset Suite" for bank customer behavior analysis. This dataset is collected from publicly available sources, and the dataset link is provided under the reference section. Here, the "Bank Account Fraud Dataset Suite" consists of 1000000 numbers of data. Among the whole data, 80% of the data is employed for training, and the remaining 20% of the data is used for testing purposes.

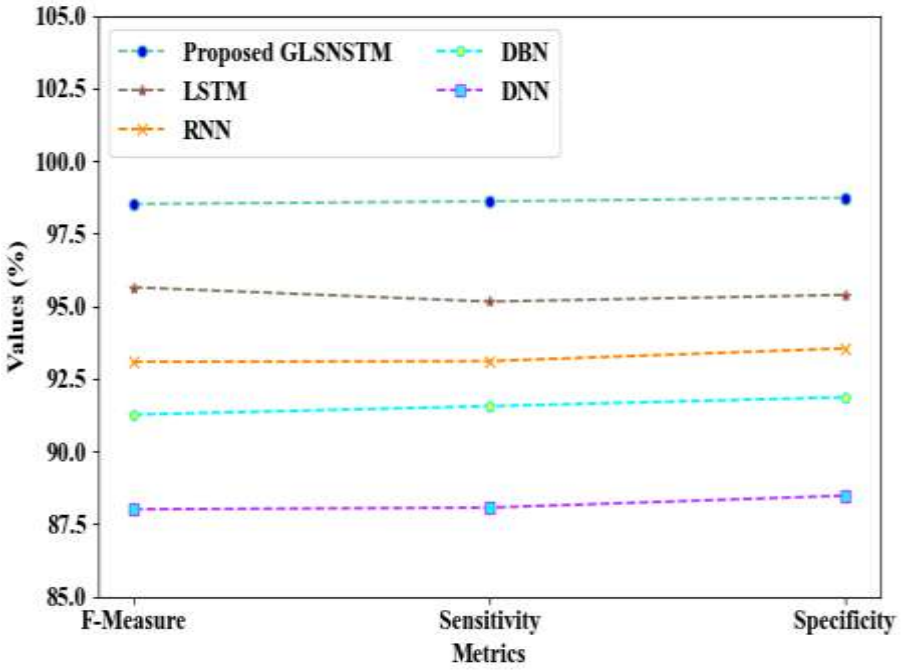
4.2 Performance Assessment

Here, the performance of the proposed model is compared with existing techniques.

10.48047/jocaaa.2023.31.04.51



(a)



(b)

Figure 3: Pictorial diagram regarding (a) Accuracy, Precision, and Recall And (B) F-Measure, Sensitivity, and Specificity

Figures 3 (a) and (b) depict the pictorial analysis of the proposed GLSNSTM and prevailing techniques based on performance metrics. Here, the proposed GLSNSTM achieved high accuracy, precision, recall, F-Measure, sensitivity, and specificity of 98.76%, 98.81%, 98.57%, 98.52%, 98.61%, and 98.73%, respectively. However, the existing methods, such as LSTM, Recurrent Neural Network (RNN), Deep Belief Network (DBN), and Deep Neural

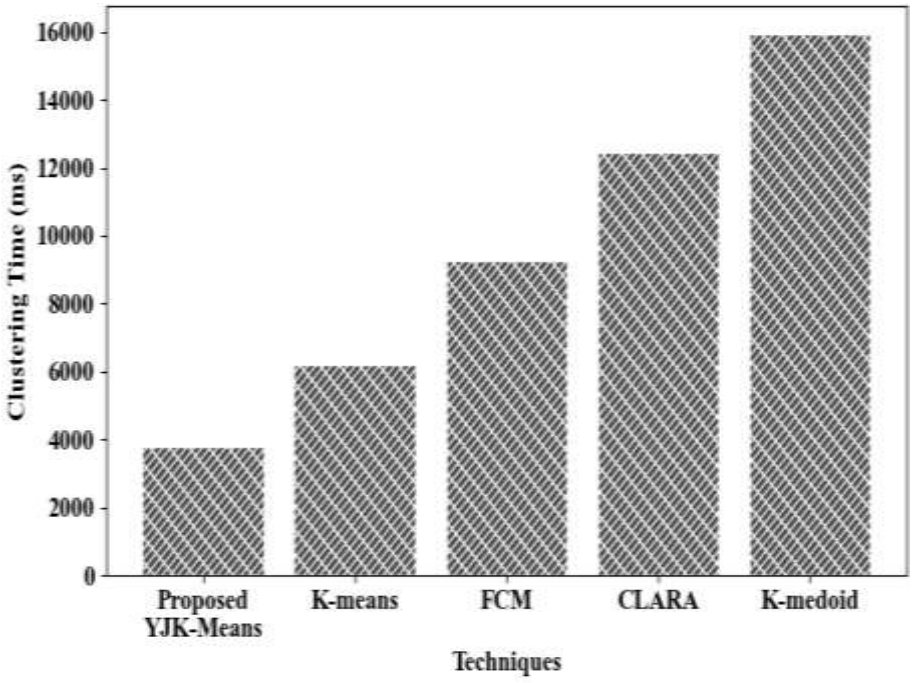
10.48047/jocaaa.2023.31.04.51

Network (DNN) attained a low average accuracy, precision, recall, F-Measure, sensitivity, and specificity of 92.21%, 92.34%, 92.27%, 92%, 91.97%, and 92.32%, respectively. Here, the proposed GLSNSTM employed the GABU activation function and Schatten-p Norm regularization technique to avoid overfitting issues and computational complexity. Thus, the trustworthiness of the proposed model was proved.

Table 1: Cohen's Kappa analysis

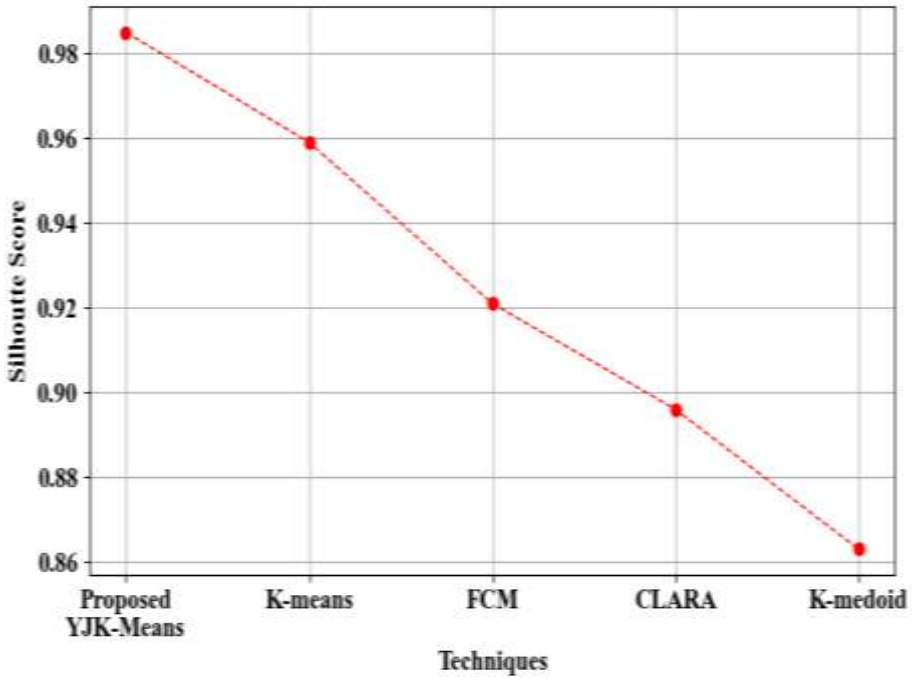
Methods	Cohen's Kappa
Proposed GLSNSTM	0.95
LSTM	0.89
RNN	0.81
DBN	0.74
DNN	0.67

Cohen's kappa analysis of the proposed GLSNSTM and prevailing techniques is depicted in Table 1. Here, the proposed GLSNSTM superiorly analyzed the bank customer behavior due to the usage of GABU activation and the Schatten-p Norm regularization technique. The proposed GLSNSTM achieved a high Cohen's kappa of 0.95, whereas the existing LSTM, RNN, DBN, and DNN attained a low Cohen's kappa of 0.89, 0.81, 0.74, and 0.67, respectively. Thus, the effectiveness of the proposed model was demonstrated.



(a)

10.48047/jocaaa.2023.31.04.51



(b)

Figure 4: Performance validation in terms of (a) Clustering time and (b) Silhouette score

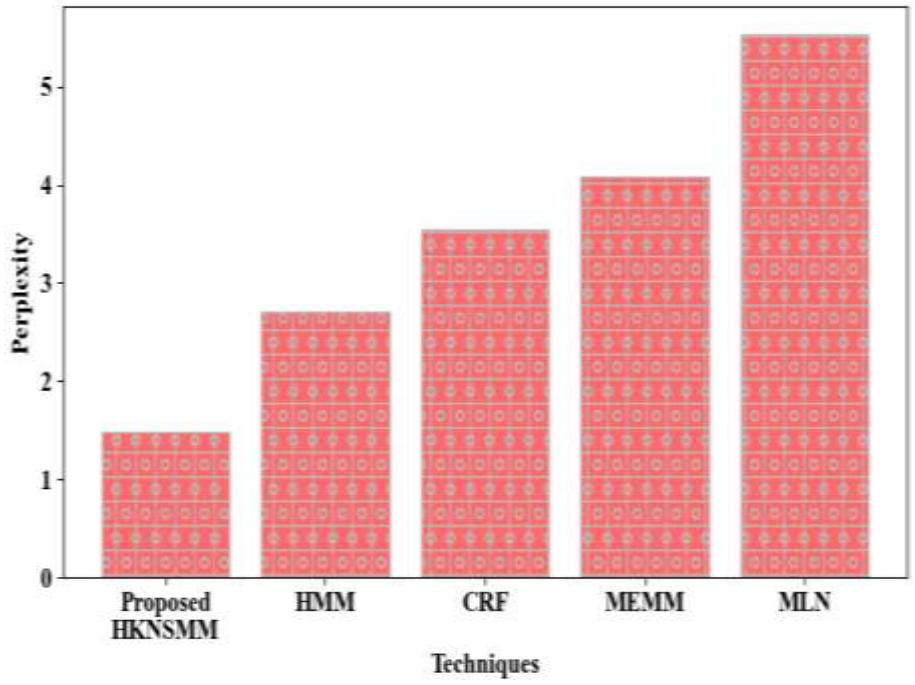
Figures 4 (a) and (b) display performance validation of the proposed YJK-Means and conventional methods in terms of clustering time and silhouette score. Here, the proposed YJK-Means achieved a high silhouette score (0.985) and low clustering time (3782ms). But, the prevailing techniques, such as K-Means and Fuzzy C-Means (FCM), attained a low silhouette score of 0.959 and 0.921, respectively. Also, the prevailing Clustering Large Applications (CLARA) and K-medoid obtained high clustering times of 12456ms and 15965ms, correspondingly. Thus, the experimental outcomes proved that the proposed model had less complexity due to the inclusion of Jeffreys Entropy and Yule distance.

Table 2: Clustering accuracy evaluation

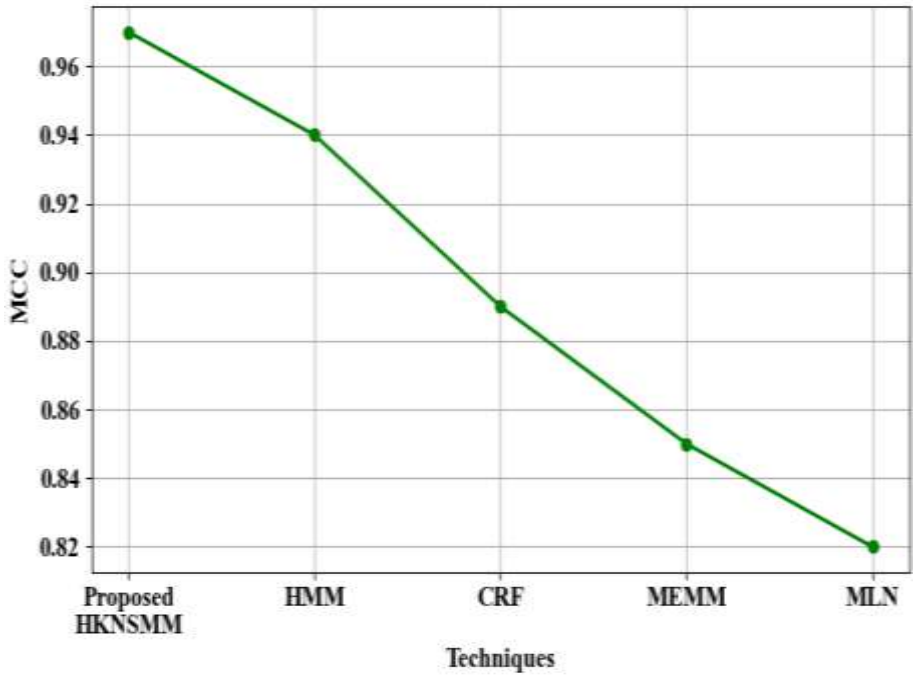
Techniques	Clustering accuracy (%)
Proposed YJK-Means	98.56
K-means	94.28
FCM	91.04
CLARA	88.76
K-medoid	85.29

Clustering accuracy evaluation of the proposed YJK-Means and conventional methods is depicted in Table 2. Here, the proposed YJK-Means detected the bank customer preferences with high accuracy due to the usage of Jeffreys Entropy and Yule distance. The proposed YJK-Means achieved a high clustering accuracy of 98.56%, whereas the conventional K-means, FCM, CLARA, and K-medoid attained low accuracies of 94.28%, 91.04%, 88.76%, and 85.29%, correspondingly. Thus, the reliability of the proposed model was proved.

10.48047/jocaaa.2023.31.04.51



(a)



(b)

Figure 5: Graphical representation with respect to (a) Perplexity and (b) MCC

Figures 5 (a) and (b) display the graphical representation of the proposed HKNSMM and prevailing techniques with respect to perplexity and Matthew's Correlation Coefficient (MCC). Here, the proposed HKNSMM achieved a low perplexity (1.478) and high MCC (0.97). However, the prevailing methods like HMM, Conditional Random Fields (CRF), Maximum Entropy Markov Model (MEMM), and Markov Logic Networks (MLN) obtained a high average perplexity of 3.972 and low average MCC of 0.87. Here, the proposed

10.48047/jocaaa.2023.31.04.51

HKNSMM excellently tagged the words due to the usage of the Kneser–Ney Smoothing technique, which effectively dealt with unseen data.

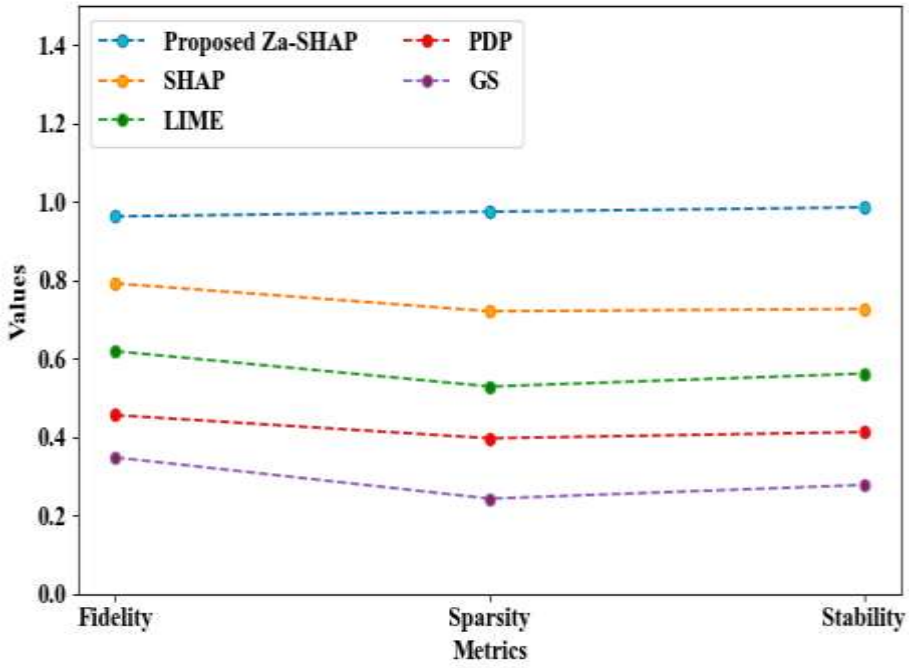


Figure 6: Performance assessment regarding Fidelity, Sparsity, and Stability

Performance assessment of the proposed Za-SHAP and existing methods in terms of fidelity, sparsity, and stability is shown in Figure 6. The proposed Za-SHAP proficiently provided an explanation of the bank customer behavior outcomes with the help of Zak's function. Here, the proposed Za-SHAP achieved a high fidelity, sparsity, and stability of 0.963, 0.975, and 0.986, respectively. However, the existing SHAP and Local-agnostic Model for Explanations (LIME) obtained low fidelities of 0.792 and 0.619, respectively. Also, the conventional Partial Dependence Plot (PDP) and Global Surrogate (GS) attained low stabilities of 0.413 and 0.278, respectively. Likewise, the conventional methods obtained low sparsity values. Thus, the proposed model was better than the existing methods.

4.3 Comparative Validation

Here, the comparative validation is carried out for the proposed and related works to prove the proposed model's trustworthiness.

Table 3: Comparative Evaluation

Authors' name	Objectives	Methods	Advantages	Limitations
Proposed model	Bank customer behavior analysis and preference detection	GLSNSTM	The proposed methodology achieved high reliability and efficacy.	Yet, it didn't analyze the user behavior and preferences evolving over time.
(Hosseini et al., 2022)	Customer behavior analysis in	RFM and K-Means	By using this model, banks could provide	However, the model had lack of contextual

10.48047/jocaaa.2023.31.04.51

	traditional and electronic banking		the right service to the right person.	understanding.
(Cui et al., 2021)	Multicontextual Behavior Profiling for Online Banking	RME	It analyzed the behavior patterns under different contexts.	However, this research offered limited interpretability, leading to lower trust.
(Abedin et al., 2023)	Bank customer behavior modeling	RF	The model guided the customers to take the right steps.	Nevertheless, in this work, noisy, redundant, and missing features could degrade performance.
(Mytnyk et al., 2023)	Fraudulent behavior recognition	LR	The experimental outcomes showed the effectiveness of the model.	Yet, the model had overfitting issues.
(Kalinin et al., 2020)	Customer financial behavior analysis	GE	It attained a very low median absolute percentage error.	But, it struggled to extract reliable financial behavior indicators from the unstructured data.

Table 3 depicts the comparative validation of the proposed and related works. Here, the proposed GLSNSTM achieved high reliability and efficacy in bank customer behavior analysis and preference detection. However, the existing Recency, Frequency, and Monetary (RFM) and K-Means algorithm had a lack of contextual understanding. Also, the conventional Ranking Metric Embedding (RME) offered limited interpretability. Similarly, the existing Graph Embedding (GE) struggled to extract reliable financial behavior indicators from the unstructured data. Likewise, the prevailing Random Forest (RF) and Logistic Regression (LR) obtained poor performance. Thus, the analysis proved that the proposed model had high effectiveness.

5. CONCLUSION

This paper presented an effective framework named inter and intra-sentence relationship analysis-aware NLP-based U.S bank customer behavior analysis and preference detection system using GLSNSTM. Here, significant processes, such as code-mixed language handling, inter and intra-sentence relationship analysis, syntactic structure analysis, keywords extraction, bank customer preference detection, and bank customer behavior analysis were performed. The proposed GLSNSTM achieved high accuracy, precision, and recall of 98.76%, 98.81%, and 98.57%, respectively for behavior analysis of bank customers. Likewise, the proposed YJK-Means took a minimum clustering time of 3782ms for bank customer preference detection. Also, the proposed HKNSMM achieved a low perplexity of 1.478 for PoS tagging. Overall, the proposed methodology achieved high reliability and

10.48047/jocaaa.2023.31.04.51

efficacy. Although the proposed model analyzed inter and intra-sentence relationships between the chat logs/email data, it didn't analyze the user behavior and preferences evolving over time.

Future Scope

In the future, enhanced temporal NLP techniques will be integrated for tracking customer behavior and preferences more accurately.

REFERENCES

Dataset link: https://www.kaggle.com/datasets/sgpjesus/bank-account-fraud-dataset-neurips-2022?select=Base.csv

Abbasimehr, H., & Shabani, M. (2019). A new methodology for customer behavior analysis using time series clustering: A case study on a bank's customers. *Kybernetes*, *50*(2), 221–242. https://doi.org/10.1108/K-09-2018-0506

Abedin, M. Z., Hajek, P., Sharif, T., Satu, M. S., & Khan, M. I. (2023). Modelling bank customer behaviour using feature engineering and classification techniques. *Research in International Business and Finance*, *65*, 1–16. https://doi.org/10.1016/j.ribaf.2023.101913

Andrian, B., Simanungkalit, T., Budi, I., & Wicaksono, A. F. (2022). Sentiment Analysis on Customer Satisfaction of Digital Banking in Indonesia. *International Journal of Advanced Computer Science and Applications*, *13*(3), 466–473. https://doi.org/10.14569/IJACSA.2022.0130356

Chao, M. H., Trappey, A. J. C., & Wu, C. T. (2021). Emerging Technologies of Natural Language-Enabled Chatbots: A Review and Trend Forecast Using Intelligent Ontology Extraction and Patent Analytics. *Complexity*, *2021*, 1–26. https://doi.org/10.1155/2021/5511866

Cui, J., Yan, C., & Wang, C. (2021). ReMEMBeR: Ranking Metric Embedding-Based Multicontextual Behavior Profiling for Online Banking Fraud Detection. *IEEE Transactions on Computational Social Systems*, *8*(3), 1–12. https://doi.org/10.1109/TCSS.2021.3052950

Domingos, E., Ojeme, B., & Daramola, O. (2021). Experimental analysis of hyperparameters for deep learning-based churn prediction in the banking sector. *Computation*, *9*(3), 1–19. https://doi.org/10.3390/computation9030034

Egbuhuzor, N. S., Ajayi, A. J., Akhigbe, E. E., Oluwadamilola, O., Ewim, C. P.-M., & Ajiga, D. I. (2021). Cloud-based CRM systems : Revolutionizing customer engagement in the financial sector with artificial intelligence Cloud-based CRM systems : Revolutionizing customer engagement in the financial sector with artificial intelligence. *International Journal of Science and Research Archive*, *3*(1), 215–234. https://doi.org/10.30574/ijsra.2021.3.1.0111

Elrefai, A. T., Elgazzar, M. H., & Khodeir, A. N. (2021). Using Artificial Intelligence in Enhancing Banking Services. *2021 IEEE 11th Annual Computing and Communication Workshop and Conference, CCWC 2021*, 980–986.

10.48047/jocaaa.2023.31.04.51

https://doi.org/10.1109/CCWC51732.2021.9375993

Gavval, R., & Ravi, V. (2020). Clustering Bank Customer Complaints on Social Media for Analytical CRM via Multi-objective Particle Swarm Optimization. *Nature Inspired Computing for Data Science*, 213–239. https://doi.org/10.1007/978-3-030-33820-6_12

Hosseini, M., Abdolvand, N., & Harandi, S. R. (2022). Two-dimensional analysis of customer behavior in traditional and electronic banking. *Digital Business*, *2*(2), 1–10. https://doi.org/10.1016/j.digbus.2022.100030

Ibitoye, A. O. J., & Onifade, O. F. W. (2022). Improved customer churn prediction model using word order contextualized semantics on customers' social opinion. *International Journal of Advances in Applied Sciences*, *11*(2), 107–112. https://doi.org/10.11591/ijaas.v11.i2.pp107-112

Kalinin, A., Vaganov, D., & Bochenina, K. (2020). Discovering patterns of customer financial behavior using social media data. *Social Network Analysis and Mining*, *10*(1), 1–14. https://doi.org/10.1007/s13278-020-00690-3

Katsafados, A. G., Androutsopoulos, I., Chalkidis, I., Fergadiotis, E., Leledakis, G. N., & Pyrgiotakis, E. G. (2021). Using textual analysis to identify merger participants: Evidence from the U.S. banking industry. *Finance Research Letters*, *42*, 1–15. https://doi.org/10.1016/j.frl.2021.101949

Kaur, R., Sandhu, R. S., Gera, A., Kaur, T., & Gera, P. (2020). Intelligent Voice Bots for Digital Banking. *Smart Innovation, Systems and Technologies*, *141*, 401–408. https://doi.org/10.1007/978-981-13-8406-6_38

Kumar, S., Ahmed, R., Bharany, S., Shuaib, M., Ahmad, T., Tag Eldin, E., Rehman, A. U., & Shafiq, M. (2022). Exploitation of Machine Learning Algorithms for Detecting Financial Crimes Based on Customers' Behavior. *Sustainability (Switzerland)*, *14*(21), 1–24. https://doi.org/10.3390/su142113875

Maheswari, B., Aswini, J., & Anita, M. (2021). Hybrid feature selection approach for naive bayes to improve consumer behavior analysis. *Proceedings of the 3rd International Conference on Intelligent Communication Technologies and Virtual Mobile Networks, ICICV 2021*, 1200–1204. https://doi.org/10.1109/ICICV50876.2021.9388439

Mashaabi, M., Alotaibi, A., Qudaih, H., Alnashwan, R., & Al-khalifa, H. (2022). Natural Language Processing in Customer Service: A Systematic Review. *ArXiv*, 1–27. https://arxiv.org/pdf/2212.09523

Mishev, K., Gjorgjevikj, A., Vodenska, I., Chitkushev, L. T., & Trajanov, D. (2020). Evaluation of Sentiment Analysis in Finance: From Lexicons to Transformers. *IEEE Access*, *8*, 131662–131682. https://doi.org/10.1109/ACCESS.2020.3009626

Mytnyk, B., Tkachyk, O., Shakhovska, N., Fedushko, S., & Syerov, Y. (2023). Application of Artificial Intelligence for Fraudulent Banking Operations Recognition. *Big Data and Cognitive Computing*, *7*(2), 1–19. https://doi.org/10.3390/bdcc7020093

Ogunleye, B., Maswera, T., Hirsch, L., Gaudoin, J., & Brunsdon, T. (2023). Comparison of Topic Modelling Approaches in the Banking Context. *Applied Sciences (Switzerland)*,

10.48047/jocaaa.2023.31.04.51

13(2), 1–14. https://doi.org/10.3390/app13020797

Olujimi, P. A., & Ade-Ibijola, A. (2023). NLP techniques for automating responses to customer queries: a systematic review. *Discover Artificial Intelligence*, *3*(1), 1–19. https://doi.org/10.1007/s44163-023-00065-5

Örpek, Z., Tural, B., & Özmen, S. (2023). Review on the Use of NLP in the Banking Sector. *International Conference on Engineering, Science and Technology*, *1*, 154–160. https://www.researchgate.net/profile/Istes-Publication/publication/378706403_Proceedings_of_International_Conference_on_Engineering_Science_and_Technology_2023/links/65e5fc14adf2362b6377cd4f/Proceedings-of-International-Conference-on-Engineering-Science-and-Technology-2023.pdf#page=161

Papadia, G., Pacella, M., & Giliberti, V. (2022). Topic Modeling for Automatic Analysis of Natural Language: A Case Study in an Italian Customer Support Center. *Algorithms*, *15*(6), 1–17. https://doi.org/10.3390/a15060204

Rustamov, S., Bayramova, A., & Alasgarov, E. (2021). Development of dialogue management system for banking services. *Applied Sciences (Switzerland)*, *11*(22), 1–18. https://doi.org/10.3390/app112210995

Venkateswararao, M., Vellela, S. S., Venkateswara Reddy, B., Vullam, N., Sk, K. B., & Roja, D. (2023). Credit Investigation and Comprehensive Risk Management System based Big Data Analytics in Commercial Banking. *2023 9th International Conference on Advanced Computing and Communication Systems, ICACCS 2023*, 2387–2391. https://doi.org/10.1109/ICACCS57279.2023.10113084

Vo, N. N. Y., Liu, S., Li, X., & Xu, G. (2021). Leveraging unstructured call log data for customer churn prediction. *Knowledge-Based Systems*, *212*, 1–21. https://doi.org/10.1016/j.knosys.2020.106586

Wang, Y., Li, Y., & Wu, T. (2021). Research on Compliance Supervision Data Analysis Model Based on Mass Chat Records in the Inter-Bank Market. *2021 IEEE 2nd International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering, ICBAIE 2021*, 368–380. https://doi.org/10.1109/ICBAIE52039.2021.9389994

Xiong, T., Ma, Z., Li, Z., & Dai, J. (2021). The analysis of influence mechanism for internet financial fraud identification and user behavior based on machine learning approaches. *International Journal of System Assurance Engineering and Management*, *13*, 996–1007. https://doi.org/10.1007/s13198-021-01181-0

Zhang, X., Agarwal, S., Choy, R., Wong, K. J., Lim, L., Lee, Y. Y., & Lu, J. J. (2020). Personalized Digital Customer Services for Consumer Banking Call Centre using Neural Networks. *Proceedings of the International Joint Conference on Neural Networks*, 1–7. https://doi.org/10.1109/IJCNN48605.2020.9206709

Zouari, G., & Abdelhedi, M. (2021). Customer satisfaction in the digital era: evidence from Islamic banking. *Journal of Innovation and Entrepreneurship*, *10*(1), 1–18. https://doi.org/10.1186/s13731-021-00151-x