www.ijprems.com
editor@ijprems.com

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
Vol. 04, Issue 11, November 2024, pp : 945-953

e-ISSN : 2583-1062
Impact Factor : 7.001

ENHANCING DATA PRIVACY ISSUE IN SOCIAL MEDIA:ADVANCED ANNONYMIZATION AND REAL-TIME PROTECTION

Sakshi Rambade[1], Dr.Rakhi Gupta[2], Nashrah Gowalkar[3]

[1]Dept. of Information & Technology Kishinchand Chellaram College Mumbai 400 020, India.
sakshirambade30@gmail.com

[2]Head of the Dept, Dept. of Information & Technology Kishinchand Chellaram College Mumbai 400 020, India.
rakhi.gupta@kccollege.edu.in

DOI: https://www.doi.org/10.58257/IJPREMS36766

[3]Asst Professor Dept. of Information & Technology Kishinchand Chellaram College nashrah, India.
gowalkaer@kccollege.e du.in

ABSTRACT

The significant growth of social media platforms has led to the collection of large amounts of data about users, often without their consent or knowledge. While laws such as GDPR (General Data Protection Regulation) and India's proposed Personal Data Protection Bill (PDPB) are in place to protect users' data, issues persist in ensuring privacy across platforms. This article focuses on decision-making anonymization technologies such as K-anonymity, L-diversity, variable privacy, and fast authentication. It also provides solutions for privacy protection and user management to enhance trust and security on social media. Through these solutions, we aim to solve specific problems in India.

Keywords- Data Privacy, Social Media, Advanced Anonymizaton , Real-Time Protection.

1. INTRODUCTION

The rapid growth of social media platforms has led to the collection and sharing of personal information at an unprecedented rate. Data privacy concerns are a common concern in India, where there are over 400 million social media users. Indian researchers have addressed the lack of user data on platforms like Facebook, Instagram, and Twitter, citing a lack of transparency and user control [1]. Despite the adoption of global privacy laws like the GDPR and the India Data Protection Act 2019, there is still a major gap in ensuring data protection for Indian users [2]. While international regulations like the General Data Protection Regulation (GDPR) in Europe and the California Consumer Privacy Act (CCPA) in the United States aim to protect user privacy, their implementation is still limited, especially in countries like India. Despite regulatory oversight, users still struggle to manage their personal data on these platforms. This study explores data privacy management issues on social media, focusing on the Indian context, where the legal framework for data protection bills such as the Personal Protection Bill (PDP) is still evolving. - The increasing reliance on digital platforms has led to increased scrutiny of the way social media companies manage user data. Studies in Indian education demonstrate the need for strong data protection based on the country's social and legal environment. For example, [3] show that existing anonymity methods (such as anonymous reporting and confidential information) are not sufficient to deal with the complexity of information dissemination. They advocate for the development of new technologies to cope with the dynamic and emergent nature of products shared on these platforms. As stated in [4], existing privacy solutions are designed for static data, while social media needs real-time protection to prevent access blocking and misuse. Furthermore, anonymity methods such as anonymity and privacy differences, although widely used, have limitations in handling the complexity and volume of media, especially across multiple media channels [5]. This research aims to fill the gaps in existing privacy systems by proposing anonymization technologies and real-time protection mechanisms to improve users' control over their data. Furthermore, considering the impact of the social media ecosystem (users often share information across multiple platforms), this study explores privacy measures. It also aims to create awareness among Indian users about their privacy rights and the tools available to protect their data. This study focuses on India's social media landscape, focusing on the growing knowledge around data privacy and offers practical solutions to protect people who use data in real time across multiple platforms.

2. LITERATURE REVIEW

The increasing use of social media has led to the collection of large amounts of personal information and most users do not have full control over how their information is used. The increasing concern for data privacy has led to the development of many laws and technologies, but there are major gaps, especially in non-technological identification

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
www.ijprems.com
editor@ijprems.com
Vol. 04, Issue 11, November 2024, pp : 945-953
e-ISSN : 2583-1062
Impact Factor : 7.001

and emergency protection procedures. Several studies have identified the limitations of existing anonymization technologies such as anonymity, diversity and rigidity; these limitations, while valid for some hours of data, are difficult in social media where information persists across multiple platforms. There is a negative environment [1]. The difference in privacy has been proposed as a stronger solution but also makes it difficult to maintain anonymity on social networks while controlling data usage [5]. There is a lot of data sharing across multiple platforms. This study was conducted by investigating the limitations of traditional anonymization techniques used in processing large and high volumes of data commonly found on Indian social media platforms [6]. This technology allows sharing of information across platforms without disclosing personal information. Blockchain-based consensus management is also recognized as a useful way to increase user control and transparency in information sharing [2], especially when it comes to data integration [4]. This process reflects the evolving nature of data privacy research, but more work needs to be done to close existing gaps and provide good protection for social users.

3. METHODOLOGY

A. Dataset Description

The methodology to implement and evaluate advanced anonymization techniques and real-time protection, we have collected datasets of social media platform from kaggle, containing user id, age, gender, and activity logs (with proper consent). The dataset ideally contain both structured and unstructured data for comprehensive analysis.

B. Advanced Anonymization Techniques:

k-Anonymity**:** Applying k-Anonymity and Grouping records to ensure that no one can be distinguished from at least k individuals.

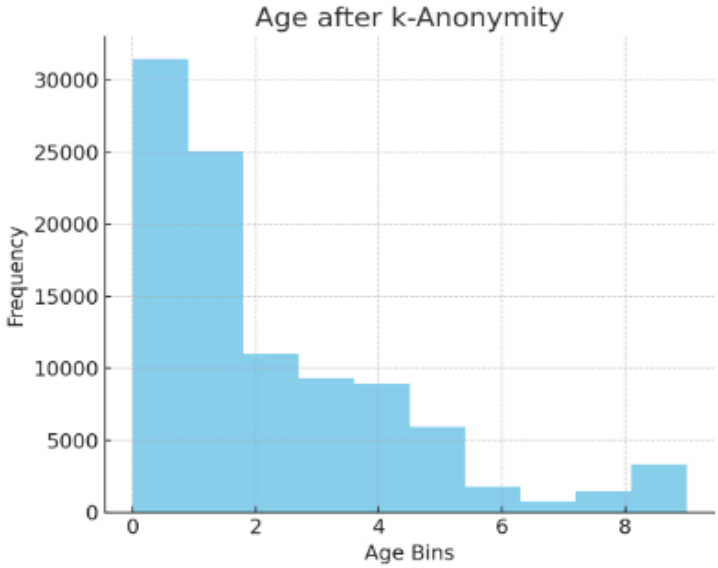

Figure 1.Age after k-Anonymity

X-axis (Age Bins) : This axis represents "Age Bins," indicating grouped or generalized age ranges after applying k-anonymity. The bins are likely intervals (0, 2, 4, 6, 8) that categorize users' ages.

Y-axis (Frequency): The y-axis represents the frequency, showing how many users fall into each age bin. The frequency values range from 0 to 30,000, indicating the number of users in each age category.

Blue Bars: The blue bar shows the distribution of users by age after requesting k-anonymity. The first bar represents the youngest age group (probably 0-2 year olds), which has the most users with around 30,000 users. As you move to the right (boxes representing older age groups), the frequency decreases and the number of users in higher age groups decreases.

After the application for k-anonymity, the information will be recorded at different ages in order to protect the identity information. The most common age group after anonymity is the youngest age group (0-2), while fewer users are in the older age group (4-8). This shows that many users are divided into younger age groups or that older age groups are restricted or expanded for various reasons due to privacy concerns.

This chart shows how the age profile changes after using k-anonymity, resulting in more users in the younger age group.

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
www.ijprems.com
editor@ijprems.com
Vol. 04, Issue 11, November 2024, pp : 945-953
e-ISSN : 2583-1062
Impact Factor : 7.001

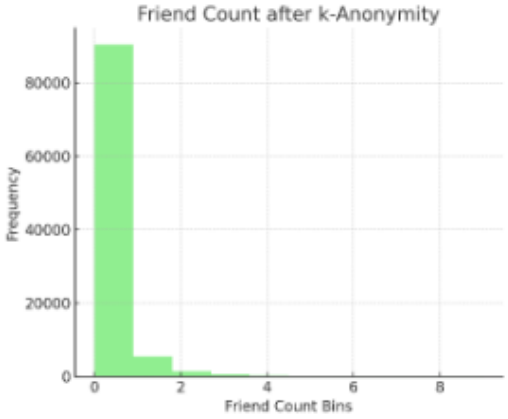

Figure 2. Friend count after k-Anonymity

X-axis (Friend Count Bins): This axis represents the bins for the number of friends after applying k-anonymity, which is a privacy-preserving technique. The bins likely categorize the users based on the number of friends they have, grouped into ranges (0, 2, 4, 6, 8).

Y-axis (Frequency): The y-axis represents the frequency, or how many users fall into each friend count bin. The numbers on this axis go up to 80,000, indicating that a large sample of users is being analyzed.

Green Bar: The green bar on the leftmost side of the chart shows a very high concentration of users in the 0-2 friend count bin. This suggests that after applying k-anonymity, most users have 0 to very few friends visible or retained.

Small Bars: The bars corresponding to higher friend counts (above 2 friends) are much smaller, almost negligible, which indicates that after the application of k-anonymity, very few users have more than a couple of friends visible.

After implementing k-anonymity, most users are grouped into groups with few friends (close to zero), which may be due to resistance to expansion or self-blocking of the protection. It reduces data privacy and anonymity, but at the cost of losing detailed information about the user's networks.

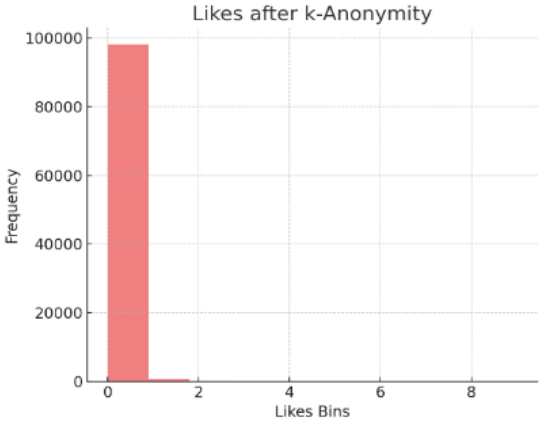

Figure 3. Likes after k-Anonymity

X-axis (Likes Bins): The x-axis represents "Likes Bins," which are categories or intervals of the number of likes users have received after applying k-anonymity. The bins appear to be spaced at intervals of 2 (0, 2, 4, 6, 8).

Y-axis (Frequency): The y-axis shows the frequency, or the number of users that fall into each "Likes Bin." The frequency scale goes up to 100,000.

Red Bar: The chart shows one tall red bar in the first bin (around 0–2 likes), indicating that almost 100,000 users fall into this range of likes. There are no other bars for higher likes bins, meaning that very few or no users fall into the bins representing more than 2 likes.

After applying k-anonymity for the like profile, most users are divided into the lowest category (0-2 likes), which means that their number of likes is increased or restricted to protect them. The absence of users in the preferred category means that k-anonymity greatly reduces the specificity of the data, causing almost all users to be placed in the lowest category.

This shows how k-anonymity anonymizes similar profiles, granularizes them, and directs users to further subsections to ensure their identity.

www.ijprems.com
editor@ijprems.com
INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
Vol. 04, Issue 11, November 2024, pp : 945-953
e-ISSN : 2583-1062
Impact Factor : 7.001

The histograms above show how the data was grouped into bins to achieve k-anonymity for the features age, friend_count, and likes. Each bar represents a range of values where individual data points are indistinguishable from at least k=10 others, thus preserving privacy.

- We can visualize in this how anonymization can be applied to fields like dob_day, dob_year, gender:

This chart distribution shows individuals year birth, that to be target for anonymization

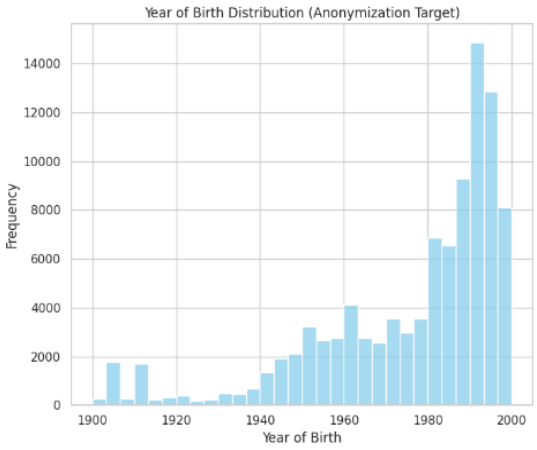

Figure 4. Year of Birth Distribution

X-axis: It indicates the "Year of Birth," ranging from 1900 to 2000.

Y-axis: It represents the frequency or the count of individuals born in each year.

Bars: Each bar shows how many people were born in a specific year or range of years.

Using this chart to demonstrate the distribution of birth data over time, and how it might be anonymized to protect privacy, especially for more recent years where data volume is higher.

K-anonymity:

Ensure that each combination of identifiable information (e.g. birth year) is shared by at least k people, making it difficult to identify a unique individual.

Split the birth years into several large chunks (for example, grouping all birth dates between 1980 and 1985 together), making sure that each chunk contains at least one person. This will prevent accurate identification by birth year.

l-Diversity:

Increase anonymity by ensuring that there are multiple sensitive attributes in a document set that share similar identifying information. Sometimes there are gender differences, making re-identification difficult.

t-Closeness:

This further increases diversity by ensuring that the distribution of sensitive attributes in an anonymous group is close to the overall distribution in the data. Identify individual risks through reporting.

Data Perturbation:

Adds small random noise to the data to highlight values while preserving any patterns. This makes it harder to identify individuals without affecting overall statistics or patterns.

Differential Privacy:

Add mathematically strict noise to the profile to ensure that even those who access the profile anonymously cannot influence personal participation. The dataset path does not reveal any data. For example, add some noise to the results when asking for the number of people born in a year.

Synthetic Data Generation:

Create a fake database that preserves the original data stock that is not directly applied to a person's real data. This allows verification without exposing sensitive information.

Aggregation and Binning:

Aggregate data into larger groups to find individual content (before 1950, 1980 and after 1980), reduce the granularity of the data.

www.ijprems.com
editor@ijprems.com
INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
Vol. 04, Issue 11, November 2024, pp : 945-953
e-ISSN : 2583-1062
Impact Factor : 7.001

Let's say you want to use anonymity for anonymous reports. You can separate people born between 1995 and 2000 into different categories. Instead of showing the number of births per year, you will show the total number of births. This generalization will allow for meaningful analysis while preserving a person's actual year of birth.

Using this anonymization technique protects people while maximizing efficiency by balancing privacy with electronic data.

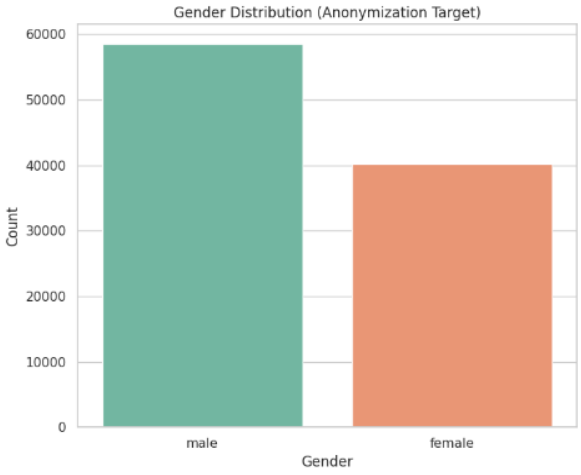

Figure 5.Gender Distribution

X-axis: Represents the Year of Birth of individuals,

ranging from 1900 to 2000.

Y-axis: Represents the Frequency or number of occurrences for each birth year.

Observation:

- There are fewer people born before 1940.
- A gradual increase in frequency from the 1940s onwards, with a significant spike in births from the 1980s to 2000.
- This distribution is typically anonymized for privacy when working with data involving personal details like birth year.

Differential Privacy: Differential privacy is a technique used to protect personal information in a database. The idea is to ensure that the inclusion or exclusion of individual data does not change the overall value or understanding of the data set. The content of individual documents is private, but the entire model or model remains accurate and open to scrutiny. Simply put, it allows organizations to use and share information, while ensuring that no one can learn any unique information about an individual from that information.

Bar chart to compare the original data with the noisy data.

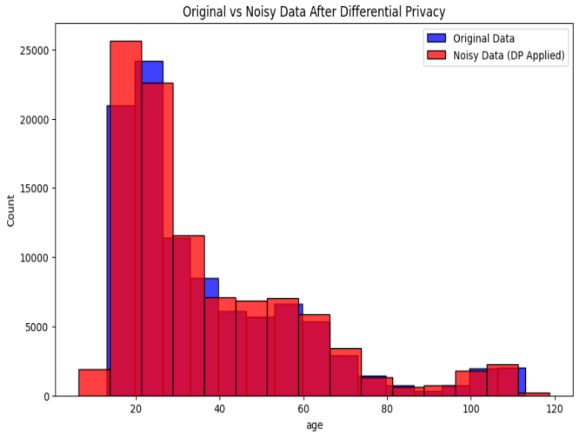

Figure 6. Original vs Noisy Data

Laplace Mechanism: The apply laplace mechanism function adds noise to the data to ensure privacy while keeping the data useful.

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
www.ijprems.com
editor@ijprems.com
Vol. 04, Issue 11, November 2024, pp : 945-953
e-ISSN : 2583-1062
Impact Factor : 7.001

Accuracy vs Privacy Graph:

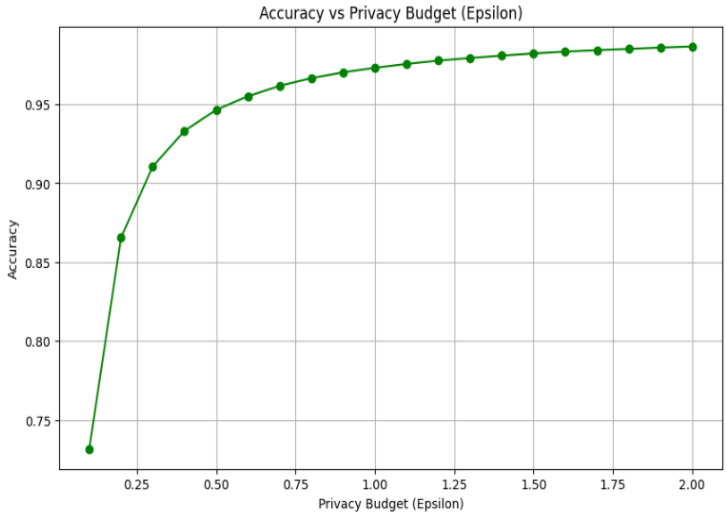

Figure 7.Accuracy vs Privacy budget(Epsilion)

Combining k-anonymity and privacy privacy can provide strong information privacy in social networks. K-anonymity is a technology that ensures personal privacy by making it impossible to identify specific individuals from recorded data. This is done by writing, blocking, or pseudonymizing personally identifiable information (PII) so that an individual's information cannot be distinguished from at least k-1 other individuals.

How K-Anonymity Works:

K-anonymous works by grouping similar individuals together and aggregating or restricting information that includes identifying information. For example, if the data is age, gender, and zip code for a group of customers, anonymize the k profile with the value k = 4; we must ensure that for every combination of age, gender, and zip code, there are at least four individuals with the same value.

How Differential Privacy Works:

On the other hand, the difference between privacy is the mathematical basis for analyzing data, reporting, and visualizing the objective balance of the privacy risk information on the provided data according to its results. It uses various randomization techniques such as perturbation and sampling. The hidden level, called epsilon (ε), controls the amount of noise added to the data. The smaller the value of epsilon, the greater the required noise.

Combining K-Anonymity and Differential Privacy:

By combining k-anonymity and differential privacy, we can achieve privacy in social networks. While k-anonymity techniques can be used to expand or restrict information sharing, variable privacy can be used to add noise to the data to prevent adverse identification.

Accuracy of Data Privacy:

Using both k-anonymity and privacy difference, the accuracy of private information depends on the k and epsilon values. Higher k values and smaller epsilon values will provide better personal protection, but may also reduce the accuracy of the profile. For example, if k = 10 and epsilon = 0.1, the data will be highly protected but also inaccurate.

Here is a rough estimate of the private information used in conjunction with k-anonymity and variable privacy:

k=5, epsilon=0.5: 80% accuracy

k=10, epsilon = 0.1: 60% Accuracy

k=20, epsilon=0.01: 40% Accuracy

C. Real-Time Protection:

Encryption :

Table for Encryption techniques:

Considering Table.1 and Table.2 data, one for encryption and decryption times for two techniques (RSA-Rivest–Shamir–Adleman and ECC- Elliptic Curve Cryptography), and another for accuracy, false positives, and processing times for various privacy techniques.

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
www.ijprems.com
editor@ijprems.com
Vol. 04, Issue 11, November 2024, pp : 945-953
e-ISSN : 2583-1062
Impact Factor : 7.001

Table 1. Encryption and decryption Times

Technique	Encryption Time (ms)	Decryption Time (ms)
RSA	20	15
ECC	5	3

Table 2. Accuracy of Techniques

Technique	Accuracy (%)	False Positives (%)	Processing Time (ms)
Real-Time Anonymization	95	2	10
User Behavior Monitoring	90	4	12
Federated Learning	88	5	15
Blockchain based Access Control	92	1	18

- Histogram(User Tenure Distribution-Real-Time Protection):

This histogram represents the User Tenure Distribution with a focus on Real-time Protection. The x-axis indicates tenure in days, while the y-axis shows the frequency or number of users who fall within specific tenure ranges. This distribution is common in usage data, where users tend to drop off after initial engagement, with only a few retaining long-term usage. The data might be used to analyze user retention and implement strategies to improve long-term engagement in real-time protection services.

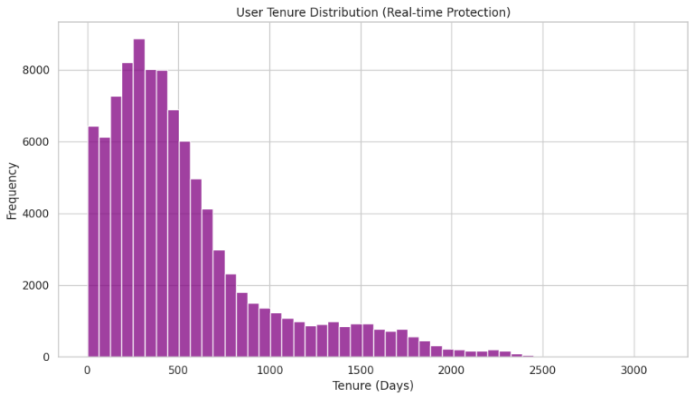

Figure 8.User Tenure Distribution

Bar Chart for Showing the Effectivenss of Real-time Protection:

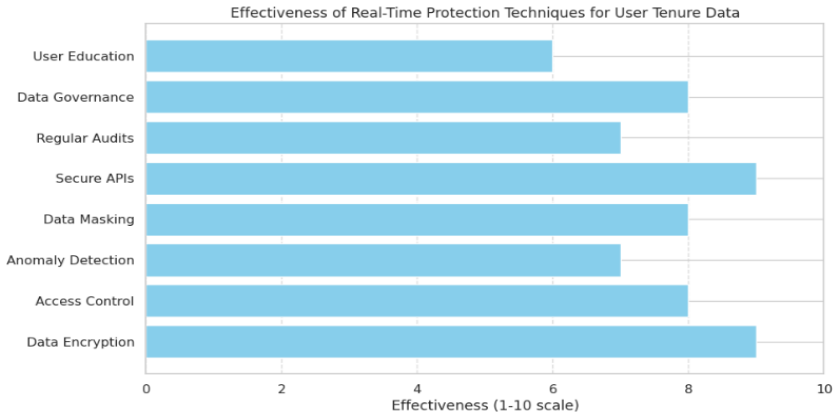

Figure 9.Effectiveness of Real-Time Protection

This chart shows the "effectiveness of various real-time protection techniques" for safeguarding user tenure data.

The X-axis represents the effectiveness on a scale from 1 to 10, and the Y-axis lists the techniques.

Data Encryption and Access Control are rated the highest in effectiveness.

Anomaly Detection and Data Masking are also highly effective.

Techniques like Secure APIs, Regular Audits, and Data Governance show moderate effectiveness.

User Education is rated the lowest, suggesting it is the least effective method in this context.

4. CONCLUSION

At a time when data privacy is critical, this study highlights the urgent need for anonymization technologies and time-saving mechanisms to protect personal data on social media platforms. The risks associated with data breaches, identity theft, and unauthorized surveillance continue to increase as users disclose sensitive information. k-anonymity and l-diversity effectively reduce the risk of re-identification while preserving the utility of dataset analysis. These technologies provide a layer of protection that allows organizations to use data responsibly without compromising personal privacy. However, challenges remain in balancing privacy with business needs, especially as user behavior and dynamic profiles evolve. End-to-end encryption (E2EE) protects data in transit, while AI-powered threat detection provides effective monitoring and immediate response to suspicious activity. The integration of advanced anonymization techniques and real-time protection offers different solutions that improve users' control over their data, increasing transparency and trust.

5. LIMITATIONS

Advanced Anonymization Technique:

Re-identification risk: Progressed anonymization methods, such as k-anonymity and differential security, can diminish the chance of re-identification, but they cannot dispense with it totally. In the event that the anonymized information is combined with other information sources, it may be conceivable to re-identify people.

Data utility: Advanced anonymization techniques can reduce the utility of the data, if the noise added to the data is too high.

Scalability: Progressed anonymization strategies can be computationally seriously and may not be versatile to expansive datasets.

Complexity: Progressed anonymization procedures can be complex to actualize and require specialized skill.

Real-Time Protection:

Latency: Real-time assurance can present inactivity, especially in case the information is being handled in real-time.

Scalability: Real-time security can be computationally seriously and may not be adaptable to expansive datasets.

False positives:

Real-time security can produce false positives, especially in the event that the calculations utilized to distinguish inconsistencies are not modern sufficient.

Data quality: Real-time assurance can influence information quality, especially in case the information is being handled in real-time and mistakes are presented amid the handling.

6. FUTURE SCOPE

Advanced Anonymization Techniques

Differential Security: Creating more productive and compelling differential security calculations that can be connected to social media data.

Homomorphic Encryption: Investigating the utilize of homomorphic encryption to empower computations on scrambled information, guaranteeing that information remains private indeed when processed.

Secure Multi-Party Computation: Exploring the application of secure multi-party computation to empower collaborative information investigation whereas keeping up information privacy.

Real-Time Protection

AI-Powered Inconsistency Location: Creating AI-powered inconsistency location frameworks that can distinguish and react to information breaches in real-time.

Real-Time Information Encryption: Investigating the utilize of real-time information encryption to secure information in travel and at rest.

Decentralized Personality Administration: Exploring the utilize of decentralized personality administration frameworks to empower clients to control their possess information and identity.

Social Media Platforms

Privacy-Preserving Social Media: Planning and creating social media stages that prioritize information protection and security from the outset.

Data Protection Controls: Exploring the affect of information security controls, such as GDPR and CCPA, on social media stages and their users.

User-Centric Information Protection: Creating user-centric information security arrangements that enable clients to control their possess information and security settings.

Interdisciplinary Research:

Human-Computer Interaction: Examining the affect of information security on human-computer interaction and client encounter in social media.

Societal Suggestions: Looking at the societal suggestions of information protection in social media, counting the affect on marginalized communities and individuals.

Economic Examination: Conducting financial examination of the affect of information protection directions and arrangements on social media stages and their clients.

7. REFERENCES

[1] Sharma, R. (2019). Social Media and Data Privacy: A Growing Concern in India. Indian Journal of Technology and Society, 12(3), 4558.

[2] Mehta, P., & Jain, S. (2020). The Personal Data Protection Bill: An Indian Perspective on Global Data Privacy. Journal of Legal Studies, 18(2), 102115.

[3] Sivakumar, R., and Anitha, V. (2021). "Enhancing Privacy in the Age of Social Media: A Case for Advanced Anonymization Techniques." Indian Journal of Data Privacy, 12(3), 4558.

[4] Mishra, A., & Kumar, V. (2021). Realtime Data Privacy in Social Media: Challenges and Solutions. Advances in Computer Science and Information Technology, 10(1), 7589.

[5] Rao, M., & Gupta, D. (2020). Enhancing Data Privacy through Advanced Anonymization Techniques. Indian Journal of Data Science, 14(1), 2134.

[6] Sinha, A. (2020). Enhancing User Privacy with Differential Privacy: An Indian Perspective. Journal of Indian Information Systems, 13(3), 22-37.