Privacy Control in Social Networks: Integrating Behavioral Patterns and Content Sensitivity for Audience Recommendation

Md Jahangir Alam[1][0009−0005−8731−7354], Ismail Hossain[1][0000−0001−8954−1150], Sai Puppala[2][0009−−0008−0334−5756], and Sajedul Talukder[1][0000−0001−8054−9770]

[1] University of Texas at El Paso, TX 79902 USA
{malam10, ihossain}@miners.utep.edu, stalukder.utep.edu
[2] Southern Illinois University Carbondale, IL 62901 USA
sai.puppala@siu.edu

Abstract. Managing the privacy of social media posts remains a complex task, especially as audience diversity and content sensitivity grow. We propose a comprehensive privacy management framework that combines post content features with behavioral signals from social interactions to deliver personalized audience recommendations. Leveraging Facebook and Reddit datasets, we implement five core modules: post privacy classification, persona contradiction detection, interaction and privacy alignment scoring, expectation mismatch analysis, and privacy-aware friend grouping. Our post classifier achieves F1-scores of 0.76 (Facebook) and 0.72 (Reddit); contradiction detection yields an F1-score of 0.80 by combining behavioral and BERT-based features. Friend clustering based on interaction and alignment scores results in silhouette scores of 0.65 (Facebook) and 0.60 (Reddit), while expectation mismatch analysis reveals stronger emotional alignment in highly private posts. Overall, our approach enables explainable and behavior-sensitive audience control for social platforms, improving upon prior content- or interaction-only methods.

Keywords: Privacy-Aware Sharing · Audience Recommendation · Behavioral Signal Analysis · Social Interaction Modeling

1 Introduction

Online social networks such as Facebook, Instagram, and Twitter have revolutionized the way individuals communicate, share content, and build social relationships. However, the increasing volume of user-generated content, combined with complex and often confusing privacy settings, has led to numerous instances of unintentional information disclosure [3,8]. Users often interact with a diverse set of friends or followers, each having different levels of relational closeness and trust. These interactions—likes, comments, shares, and direct messages—can serve as implicit signals about the nature of the relationship. Prior work has demonstrated that such interaction patterns can be used to infer tie strength [6]

Despite existing work on content-based or rule-based privacy controls, there remains a gap in understanding how social interaction behavior can be used to automate and personalize privacy settings for posts. Current models often fail to account for the heterogeneity of friend interactions and lack mechanisms for dynamically recommending audience groups tailored to the sensitivity of content. This motivates our research: *Can social interaction patterns and post content be jointly leveraged to classify post privacy levels and suggest personalized audience groups?* To address this gap, we propose a novel framework that integrates user-post content features and social interaction behaviors to classify posts into privacy categories—*highly private*, *private*, and *public*—and recommend audience groups that align with these privacy levels. Unlike prior works that focus separately on content-based privacy prediction [15], tie strength modeling [6], or access control rules [4], our framework jointly models post semantics and social behavior for multi-level privacy classification with personalized audience recommendation. Unlike works such as Petkos et al. [15] that rely on static content features (e.g., images, tags), our framework integrates these with dynamic behavioral signals like comment frequency, messaging intensity, and mutual interaction scores to derive trust-aware privacy recommendations. While methods such as Gilbert and Karahalios [6].

In contrast, our framework advances beyond this by incorporating friend interaction histories into clustering model (e.g. KMeans) enabling semantically and socially coherent audience grouping for privacy-aware dissemination. Moreover, while many previous studies rely on proprietary or platform-restricted datasets, limiting reproducibility, our approach leverages a publicly available Reddit dataset and our own curated Facebook dataset.

We organize our study around the following research questions:

RQ1: How can post content features (e.g., sentiment, topics, entities) and social interaction patterns be jointly leveraged to predict the intended privacy level of user-generated posts?

RQ2: To what extent does modeling persona alignment and contradiction between posters and commenters improve the detection of potential value misalignments in social media interactions?

RQ3: How effective are interaction-based and privacy alignment metrics in quantifying the relational closeness and privacy compatibility of friends?

RQ4: Can expectation mismatch be modeled to enhance personalized audience control for sensitive posts?

RQ5: How well do unsupervised clustering techniques segment friends into privacy-aware audience groups that align with the sensitivity and interaction context of each post?

In summary, this work makes the following key contributions. First, we propose a privacy classification framework that combines post content and social interaction features to predict the intended privacy level of user posts. Second, we introduce a persona contradiction detection mechanism to model ideological and behavioral misalignments between posters and commenters. Third, we define interaction and privacy alignment metrics to quantify relational closeness

and compatibility. Fourth, we extend the framework with expectation mismatch analysis to account for audience behaviors. Finally, we develop a clustering-based friend grouping strategy that recommends privacy-aware audience groups using these multi-dimensional features. Together, these contributions advance personalized privacy management by enabling explainable, content- and behavior-aware audience recommendations for social media platforms.

2 Related Work

2.1 Privacy Management in Social Networks

User privacy in online social networks (OSNs) has been a longstanding concern, particularly due to complex privacy settings and unintentional oversharing. Bonneau et al. [3] discussed the challenges of designing intuitive privacy interfaces and proposed privacy "suites" for managing shared content. Liu and Terzi [8] developed a framework for computing user privacy scores based on social graph structures, highlighting structural indicators of exposure. Other works have explored usable privacy configurations and adaptive access control policies. Madejski et al. [9] investigated mismatches between user expectations and actual Facebook sharing policies. Shehab et al. [11] proposed a semi-supervised learning-based system to assist users in configuring friend-specific privacy settings in social networks.

2.2 Interaction-Based Modeling and Tie Strength

Understanding social ties through user interaction behavior is critical in OSNs. Gilbert and Karahalios [6] introduced a supervised model to predict the strength of social ties using features such as wall posts, comments, and private messaging frequency. Vitak et al. [14] examined how specific Facebook behaviors, such as reciprocity and communication frequency, relate to users' perceptions of bonding social capital and emotional closeness. Shin and Lee [12] analyzed interaction patterns in social media and proposed methods to quantify user sociability, helping to infer tie strength from online communication behavior. Brailovskaia et al. [7] demonstrated that higher Facebook use intensity is indirectly associated with increased depressive symptoms through problematic Facebook use, with the effects moderated by personality traits and age.

2.3 Content-Based Privacy Prediction

Parallel to interaction-based methods, several researchers have focused on content features to assess privacy sensitivity. Wu et al. [15] utilized multimodal information including sensitive and non-sensitive data components to balance privacy preservation and utility, which aligns with efforts to assess post sensitivity using multiple content types. Tonge and Caragea [13] proposed a deep learning-based framework for image privacy prediction, leveraging convolutional neural networks to automatically extract both visual semantics and contextual cues.

2.4 Audience Selection and Privacy-Aware Recommendations

Research has also emerged around audience recommendation systems for personalized privacy control. Settanni and Marengo [10] analyzed Facebook posts to assess emotional expression patterns, demonstrating how textual cues can guide targeted audience control mechanisms. Our work extends these ideas by combining social interaction features and content-based analysis to not only classify the privacy level of user posts but also recommend appropriate friend groups for content visibility. This hybrid approach fills an important gap in current privacy-aware systems by simultaneously modeling relationship strength and post semantics.

3 Dataset

We utilize two datasets: a Facebook dataset and the Reddit Pushshift dataset (April 2019). From Facebook, we collected demographic data and post histories for 500 users, including 15,420 posts and interaction data. From Pushshift, we generated 500 synthetic user profiles and selected 50,000 highly engaged author posts.

Demographic Data Collection: Using a Selenium-based Python script, we collected user demographics—name, gender, education, profession, relationship status, and interests—after API limitations with the Facebook Graph API. Extracted data formed part of each user's post context.

Demographic Data Generation for Reddit Data: To simulate Facebook-style user demographics in our Reddit-based study, we used the open-source `Faker` Python package[3] to generate 500 synthetic user profiles. Each profile included demographic (name, gender, birthdate), educational (school, degree, year), professional (company, title, start year), and ideological (religion, political view) attributes. Birthdates were sampled from 1955–2005, education years from 1995–2025, and employment start years from 1990–2024. Additional fields included relationship status, a subset of interests (e.g., travel, photography), and account creation dates (2010–2025). All attributes were independently sampled and assigned to Reddit users to support persona construction and contradiction modeling. Though synthetic, the profiles reflect realistic distributions and enable privacy-compliant experimentation aligned with Facebook-style data

User Post Collection: We collected 15,420 Facebook posts from diverse groups and pages. Posts were summarized into behavioral profiles. In Pushshift, posts and comments (linked via `link_id`) serve as user content and interactions. Privacy labels were heuristically assigned based on sentiment, entities, and emotional cues, validated by human annotators.

Interaction Data Collection: We gathered reactions and comments linking users to posts, and mapped engagement patterns by collecting posts from interacting users. This built an engagement network summarizing content interaction behaviors. From Pushshift, we constructed a user-friend graph $G = (V, E, W)$, with edges weighted by interaction strength.

[3] `https://github.com/joke2k/faker`

Challenges and Mitigation in Data Collection: We encountered challenges with the Facebook Graph API, including token expiry and rate limiting [5], requiring frequent token regeneration and limiting data collection. To overcome this, we have utilized Selenium automation, which bypassed these restrictions but introduced issues with navigating Facebook's dynamic HTML structure, handling dynamic content, and maintaining XPath expressions due to frequent updates. These obstacles necessitated adaptive script update to ensure accurate and scalable data collection.

Edge Weight Computation Strategy: To quantify the intensity of interactions between a post author u and a commenter v, we compute edge weights $w(u, v)$ in the interaction graph using a simple edge count strategy. This approach counts the total number of comments made by v on posts authored by u, capturing the direct frequency of engagement between the two users.

4 Methodology

To support personalized visibility control over social media posts, we propose a comprehensive framework that integrates post privacy classification, user interaction behavior, and persona-level contradiction modeling. The objective is to group each user's friends into privacy-aligned clusters (Highly Private, Private, Public) based on their behavioral, ideological, and relational compatibility with the post.

4.1 Post Privacy Classification

To predict the intended visibility level of user-generated posts, we develop a privacy classification pipeline leveraging both content-based and contextual features. The goal is to automatically categorize posts into three privacy levels: *Highly Private*, *Private*, and *Public*. Our implementation follows a multi-stage feature extraction and classification process depicted in 1 and described below.

Feature Extraction For each post, we extract multiple categories of features. First, we apply sentiment analysis using the `TextBlob` library to compute the sentiment polarity score, representing the emotional tone of the post. Second, we perform named entity recognition using the `spaCy` NLP library to count the number of named entities such as persons, organizations, and locations. Third, we compute structural features including the token count and sentence count, which capture the linguistic complexity and length of the post.

Contextual Embedding Generation To capture the semantic meaning of the entire post, we use a pretrained transformer model, `bert-base-uncased`, from the Hugging Face `transformers` library. The post text is tokenized and passed through the BERT model to extract the `[CLS]` token embedding from the final hidden layer. This embedding serves as a high-dimensional contextual representation of the post.

Feature Vector Construction All extracted features—sentiment polarity, named entity count, token count, sentence count, and the BERT embedding—are

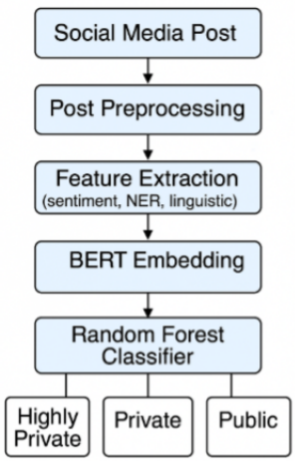

Fig. 1. Post Privacy Classification.

concatenated to form a comprehensive feature vector for each post. These feature vectors serve as the input to the classification model.

Classification Model We train a `Random Forest Classifier` with 50 trees using the `scikit-learn` library. The classifier is trained on a dataset of labeled posts, split into 67% for training and 33% for testing. The labels are encoded using a `LabelEncoder` to map the privacy levels (*Highly Private*, *Private*, *Public*) to numerical values.

4.2 Persona Contradiction Detection

To identify value-based misalignments between content posters and their audience, we propose a persona contradiction detection framework that analyzes user behaviors, ideological signals, and historical engagement patterns. The objective is to automatically determine whether a commenter is likely to contradict the poster's intent, based on their prior activity and contextual signals. Our implementation consists of four key stages described below.

Persona Construction. We construct structured personas for both the *poster* and the *commenter* by integrating multiple behavioral and content-based signals. Each persona includes (1) demographic attributes such as age, gender, profession, and religious orientation (if available); (2) historical content analysis based on topic distribution, named entity frequency, and sentiment trends from the user's past posts and comments; and (3) interaction behavior characterized by sentiment polarity and engagement patterns, including supportive, disagreeing, or sarcastic responses to various post categories.

Behavioral Feature Extraction. To model user behaviors, we compute statistical summaries such as average sentiment polarity, frequency of topic engagement (e.g., technology, politics, lifestyle), and named entity occurrence. We

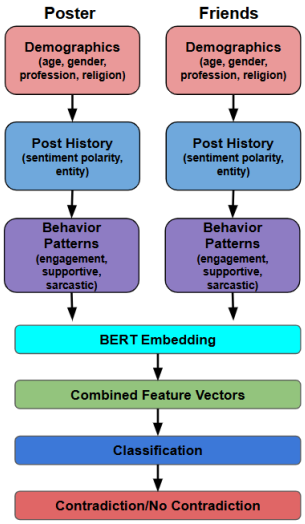

Fig. 2. Persona Contradiction.

further analyze interaction tone to classify user responses into behavioral categories. These features capture the user's typical stance and emotional alignment with various social media topics.

Contextual Embedding Generation. We use a pretrained BERT model (`bert-base-uncased`) transformer model to generate contextual embeddings of user-generated content. Specifically, we aggregate the [CLS] token representations of a user's historical posts and comments to obtain a semantic embedding that captures the user's overall communication style and ideological position. These embeddings are combined with the previously extracted statistical behavior features to form comprehensive persona vectors.

Contradiction Classification. We construct a combined feature vector by concatenating the poster's and commenter's persona vectors. This vector captures the alignment or divergence between the two users' behaviors and ideological signals. A binary `Random Forest Classifier` is trained on labeled examples of contradictory and non-contradictory interactions. The classifier predicts whether a given pair of users is likely to experience a contradiction when the commenter engages with the poster's content.

4.3 Interaction and Privacy Alignment Score

To quantify the behavioral relevance and privacy compatibility of a user's social connections, we introduce two complementary scoring mechanisms: interaction score and privacy alignment score. First, we construct a directed user-friend interaction graph from post and comment relations extracted from the Reddit Pushshift and Facebook datasets. Nodes in this graph represent users, and di-

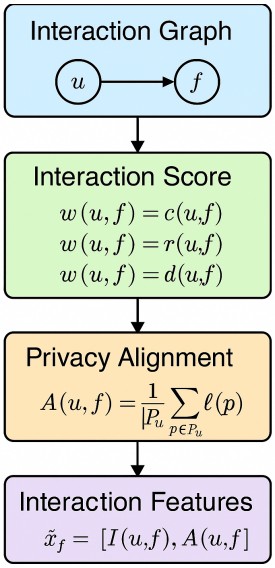

Fig. 3. Interaction and Privacy Alignment.

rected edges represent comments from a friend on a user's post. Edge weights are computed using simple edge count (total number of comments).

Second, we compute the **privacy alignment score**, which captures how well a friend's historical engagement aligns with the privacy level of the user's posts. Specifically, for each friend, we calculate the average privacy label of all posts they have previously commented on, using numerical mappings (e.g., 2 for Highly Private, 1 for Private, and 0 for Public). This score reflects the friend's compatibility with the user's privacy preferences based on past engagement behavior.

Finally, we combine these two scores into a feature vector for each friend. These vectors serve as input for downstream clustering to form privacy-aware audience groups or for binary classification to determine whether a friend should be included or excluded from the audience of a specific post. This multi-metric approach ensures that audience recommendations consider both relational strength and privacy compatibility.

4.4 Expectation vs. Interaction Alignment

We propose an **Expectation Mismatch Framework** to quantify the emotional alignment between a user's anticipated interaction and the actual social feedback received on their posts.

For each post, we construct an *expected interaction vector* derived either from manual annotation or heuristic sentiment inference based on post content. This vector is encoded as a one-hot representation in a predefined interaction label

space consisting of {*positive*, *supportive*, *neutral*, *negative*, *sarcastic*}. Next, we process the comments associated with the post to compute the *observed interaction vector*. Each comment is labeled according to its tone, and the distribution of tones is transformed into a normalized frequency vector, representing the relative proportion of each tone received.

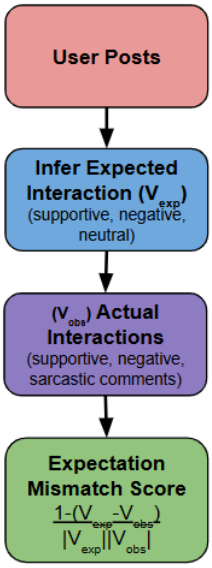

Fig. 4. Expectation Mismatch.

To measure the expectation mismatch we measure the angular difference between the expected and observed vectors using cosine similarity, providing a scale-invariant similarity score.

Formally, let $\mathbf{v}_{exp}$ be the expected interaction vector and $\mathbf{v}_{obs}$ be the observed interaction vector. The cosine distance is computed as:

$$CosineDistance(\mathbf{v}_{exp}, \mathbf{v}_{obs}) = 1 - \frac{\mathbf{v}_{exp} \cdot \mathbf{v}_{obs}}{\|\mathbf{v}_{exp}\| \times \|\mathbf{v}_{obs}\|}$$

A higher mismatch score indicates a greater divergence between the expected and actual interaction sentiments, potentially signaling emotional misalignment, discomfort, or unmet social expectations. This framework provides a scalable and quantitative method to analyze the emotional dynamics of user-generated content and its reception.

4.5 Friend Grouping and Binary Classification

For each friend f of user u, we construct a comprehensive feature vector:

$$\mathbf{x}_f = [I(u, f), A(u, f), C(u, f), E(u, f)]$$

where $I(u, f)$ is the interaction score reflecting the strength and frequency of social engagement, $A(u, f)$ is the privacy alignment score based on the friend's prior interactions with posts of varying sensitivity, $C(u, f)$ is the persona contradiction score indicating ideological divergence, $E(u, f)$ is the expectation mismatch score quantifying dissonance between expected and received interaction tone.

Using above mentioned four-dimensional friend representations, we apply unsupervised clustering (e.g., KMeans) and supervised multiclass classification (e.g., Random Forest) to segment friends by privacy levels. The **Highly Private Group** includes trusted friends with strong engagement, alignment, and low risk. The **Private Group** includes moderately aligned, low-mismatch friends. The **Public Group** includes weakly engaged, misaligned users. These data-driven groups enable personalized audience control based on content sensitivity and social dynamics.

5 Experimental Setup

We validate our framework on two datasets: (1) Reddit Pushshift for content-level privacy simulation, and (2) a curated Facebook dataset with user demographics and interaction graphs.

Post Privacy Classification: We label posts as *Highly Private*, *Private*, or *Public*, and train a Random Forest with content features and BERT embeddings, reporting standard metrics. We used 15,420 Facebook posts and 50,000 Reddit posts for classification. Posts were heuristically labeled into three privacy categories using a combination of sentiment intensity, named entity density, and emotional cues, followed by human validation for a subset (N=1,000) to ensure label quality. The final dataset had a roughly balanced distribution across privacy classes.

Persona Contradiction Detection: We created a dataset of 10,000 poster commenter pairs from Reddit and Facebook, labeled as contradictory or non-contradictory based on sentiment divergence, topic mismatch, and tone. The data was split 80%/20% for training and testing, using stratified sampling for class balance. Each user was represented by (1) a 300-dimensional contextual embedding from aggregated [CLS] tokens using `bert-base-uncased`, and (2) an 84-dimensional behavioral vector including sentiment trends, topic frequency, named entities, demographics (e.g., age, gender, religion), and interaction tones (supportive, sarcastic, disagreeing). The final feature vector (768-D) combined both personas. A Random Forest with 100 trees and Gini criterion was trained, using 5-fold cross-validation to ensure robustness.

Interaction and Privacy Alignment: We constructed an interaction graph from Reddit and Facebook data, yielding 10,000 user–friend pairs where each friend had commented on at least three posts. Each pair was represented by a

2D vector: (1) an interaction score based on simple count and (2) a privacy alignment score—defined as the average privacy level of posts the friend interacted with. Scores were min-max normalized per user. We applied KMeans clustering with $k = 3$ (Highly Private, Private, Public) using `k-means++` initialization. Clustering quality was assessed via silhouette scores, and results were visualized with a scatter plot.

Expectation vs. Interaction Alignment: We evaluated expectation mismatch using 5,000 posts from Facebook and Reddit, stratified across Highly Private, Private, and Public categories. For each post, the expected interaction vector was derived via manual annotation (for 500 posts) or heuristic sentiment analysis based on content polarity. The observed vector was computed by labeling comments into five tones—positive, supportive, neutral, negative, sarcastic—using keywords, sentiment scores, and tone markers. Both vectors were normalized, and cosine distance was used to calculate mismatch scores. Scores were grouped by privacy level and platform to analyze emotional alignment trends.

Friend Grouping and Binary Classification: We generated feature vectors $\mathbf{x}_f = [I(u, f), A(u, f), C(u, f), E(u, f)]$ for 20,000 Facebook and Reddit user–friend pairs. Features captured interaction strength, privacy alignment, contradiction, and expectation mismatch. Values were min-max normalized per user. We applied KMeans clustering with $k = 3$ (Highly Private, Private, Public) using `k-means++` and 100 iterations. For supervised classification, a multi-class Random Forest was trained on 70% of labeled data and tested on 30%. Labels were assigned using composite thresholds (e.g., high I, low C, E) and validated on a subset.

6 Experimental Results

In this section we describe the result and the effectiveness of each major components, including post privacy classification, persona contradiction detection, interaction and privacy alignment scoring, expectation mismatch computation and finally friend grouping.

6.1 Post Privacy Classification

Table 1 shows the performance of the Random Forest classifier trained on content features and BERT embeddings. The model achieves a macro F1-score of 0.76 on the Facebook dataset and 0.72 on the Reddit dataset, outperforming content-only and interaction-only baselines. These results show that the model performs consistently well across platforms, with slightly higher performance on Facebook due to richer interaction context. This experiment directly addresses **RQ1** by demonstrating that content-based and interaction-informed features can effectively predict users' intended privacy levels. The performance gains over content-only baselines (F1 = 0.76 vs 0.68) validate the utility of combining sentiment, structural, and semantic signals.

6.2 Persona Contradiction Detection

Table 2 presents the classification results for the persona contradiction detection task. The proposed model combining behavioral features and contextual embeddings achieves an F1-score of 0.80, outperforming both a behavior-only model (F1 = 0.70) and a BERT-only model (F1 = 0.75). These results demonstrate that integrating structured behavioral signals with deep semantic embeddings enhances the model's ability to detect ideological and value-based contradictions between posters and commenters. The improvement in precision also indicates the model's effectiveness in reducing false positives when filtering misaligned audience members. This experiment directly addresses **RQ2** by demonstrating that modeling persona alignment significantly improves contradiction detection performance in online interactions.

Table 1. Post Privacy Classification Results

Dataset	**Accuracy**	**Precision**	**Recall**	**F1-Score**
Facebook	0.78	0.77	0.75	0.76
Reddit	0.74	0.73	0.71	0.72

Table 2. Persona Contradiction Detection Results

Model	**Precision**	**Recall**	**F1-Score**
Behavior-Only (Random Forest)	0.72	0.69	0.70
BERT-Only	0.76	0.75	0.75
BERT + Behavior (Ours)	**0.81**	**0.79**	**0.80**

6.3 Interaction and Privacy Alignment

Figure 5 shows the distribution of interaction and privacy alignment scores for Facebook (blue) and Reddit (green) friends. The scatter pattern reveals broad coverage across the interaction-privacy spectrum, confirming diversity in relational strength and privacy compatibility. Notably, clusters of friends with both high interaction and high alignment scores suggest ideal candidates for Highly Private audiences. The silhouette scores achieved through clustering: 0.65 for Facebook and 0.60 for Reddit, indicating well-formed and moderately separated clusters. These results support **RQ3** by demonstrating that the proposed dual-score framework effectively segments friends into privacy-aligned audience groups and enables behavior-aware visibility control in social platforms.

6.4 Expectation Mismatch Score

Figure 6 shows the average expectation mismatch scores by privacy level for Facebook and Reddit. As expected, public posts exhibited the highest mismatch

(0.65 for Facebook, 0.70 for Reddit), indicating more unpredictable or emotionally divergent feedback in open sharing scenarios. In contrast, highly private posts had the lowest mismatch scores (0.15 for Facebook, 0.20 for Reddit), reflecting more aligned and supportive feedback from trusted audiences. This pattern suggests that emotional resonance is better preserved in close-knit sharing contexts. These findings directly address **RQ4**, confirming that modeling expectation mismatch provides a quantifiable signal of social dissonance and can guide more context-sensitive audience recommendations.

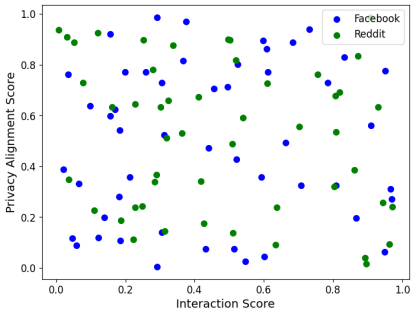

Fig. 5. Distribution of Interaction and Privacy Alignment Scores

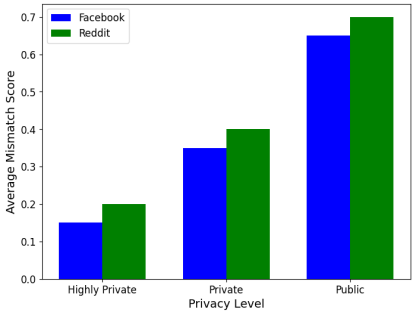

Fig. 6. Expectation Mismatch Scores by Privacy Level

6.5 Friend Grouping and Audience Recommendation.

The silhouette scores of 0.65 for Facebook and 0.60 for Reddit, confirm that the clustering approach yields well-separated and privacy-consistent friend groups. Qualitative inspection of sample clusters showed that highly private clusters included friends with strong engagement histories, low ideological misalignment, and high privacy compatibility. In the classification task, the multi-class Random Forest classifier achieved an overall accuracy of 0.78, with precision and recall exceeding 0.80 for the Highly Private class, showing the model's ability to accurately identify trusted friends. These results directly support **RQ5**, demonstrating that a multi-dimensional behavioral and content-aware representation of friends enables effective clustering and classification into privacy-aligned audience groups for personalized post sharing.

In summary, our experiments demonstrate that integrating content, interaction, and behavioral modeling significantly improves privacy-aware audience selection, outperforming baselines that consider only content or interaction in isolation.

7 Ethical Considerations

We followed rigorous ethical guidelines and received IRB approval for the collection of Facebook data. The dataset was collected only from consenting participants, who authorized access to their posts, comments, and interactions. Selenium was used solely as a tool to access participant-authorized content (not public scraping), in compliance with their explicit consent. All data were anonymized by removing personal identifiers in accordance with NIST SP 800-122 [2] and GDPR [1] standards. Access was restricted to authorized researchers. This consent-driven collection method aligns with Facebook's platform policies for academic research involving user permission, and ensures legal and ethical compliance while advancing privacy-aware audience recommendation research.

8 Discussion and Limitations

While this simulation enables controlled and privacy-compliant experimentation, it also introduces key limitations. Most notably, because Reddit does not provide real demographic attributes, we cannot validate the closeness of these synthetic profiles to actual user distributions. The demographic and ideological traits were assigned independently and are not behaviorally grounded in users' posting or interaction history, which limits the accuracy of persona modeling. Furthermore, cultural and ideological correlations—such as between religion and political view—are not explicitly modeled, and the independence of attribute sampling may result in implausible combinations (e.g., a PhD holder with an entry-level job title). As such, while useful for exploratory analysis, these synthetic profiles do not capture the full nuance of real-world user personas and should be interpreted as approximate representations rather than precise simulations.

9 Conclusion

In this work, we proposed a novel framework for privacy-aware content sharing on social networks. By leveraging both user-post content features and friend interaction behavior, our model classifies posts into three privacy levels: *highly private*, *private*, and *public*. We further proposed a friend clustering mechanism to generate recommended audience groups for each privacy tier. Experimental results show that combining content and interaction data significantly improves classification accuracy and leads to coherent friend groupings that align with privacy intentions. Our framework paves the way for more personalized and dynamic privacy settings in online social networks. As future directions, we plan to explore explainable privacy recommendations, multi-modal content analysis, and real-time privacy suggestion tools that can assist users during the posting process.

References

1. General data protection regulation (gdpr). https://gdpr-info.eu/ (2021), accessed: 2025-07-10
2. Guide to protecting the confidentiality of personally identifiable information (pii). https://tinyurl.com/ylyjst5y (2021), accessed: 2025-07-10
3. Bonneau, J., Anderson, J., Danezis, G.: Privacy suites: Shared privacy for social networks. In: Symposium on Usable Privacy and Security (SOUPS) (2009)
4. Dong, C., Jin, H., Knijnenburg, B.P.: Ppm: A privacy prediction model for online social networks. In: Social Informatics: 8th International Conference, SocInfo 2016, Bellevue, WA, USA, November 11-14, 2016, Proceedings, Part II 8. pp. 400–420. Springer (2016)
5. Facebook Developers: Facebook graph api rate limiting. https://developers.facebook.com/docs/graph-api/overview/rate-limiting/ (2024), accessed: 2025-07-10
6. Gilbert, E., Karahalios, K.: Predicting tie strength with social media. In: Proceedings of the SIGCHI conference on human factors in computing systems. pp. 211–220 (2009)
7. Gugushvili, N., Täht, K., Ruiter, R.A., Verduyn, P.: Facebook use intensity and depressive symptoms: a moderated mediation model of problematic facebook use, age, neuroticism, and extraversion. BMC psychology **10**(1), 279 (2022)
8. Liu, K., Terzi, E.: A framework for computing the privacy scores of users in online social networks. In: Data Mining (ICDM), 2009 IEEE 9th International Conference on. pp. 288–297. IEEE (2009)
9. Madejski, M., Johnson, M., Bellovin, S.M.: The failure of online social network privacy settings. In: Technical Report CUCS-010-11 (2011)
10. Settanni, M., Marengo, D.: Sharing feelings online: studying emotional well-being via automated text analysis of facebook posts. Frontiers in psychology **6**, 1045 (2015)
11. Shehab, M., Touati, H., Javed, Y.: Semi-supervised policy recommendation for online social networks. Social Network Analysis and Mining **6**, 1–12 (2016)
12. Shin, H., Lee, J.: Impact and degree of user sociability in social media. Information Sciences **196**, 28–46 (2012)
13. Tonge, A., Caragea, C.: Image privacy prediction using deep neural networks. ACM Transactions on the Web (TWEB) **14**(2), 1–32 (2020)
14. Vitak, J., Ellison, N.B., Steinfield, C.: The ties that bond: Re-examining the relationship between facebook use and bonding social capital. In: 2011 44th Hawaii international conference on system sciences. pp. 1–10. IEEE (2011)
15. Wu, Q., Tang, J., Dang, S., Chen, G.: Data privacy and utility trade-off based on mutual information neural estimator. Expert Systems with Applications **207**, 118012 (2022)