NLP Privacy Risk Identification in Social Media (NLP-PRISM): A Survey

Dhiman Goswami, Jai Kruthunz Naveen Kumar, Sanchari Das
George Mason University
Fairfax, VA, USA
{dgoswam,jnaveenk,sdas35}@gmu.edu

Abstract

Natural Language Processing (NLP) is integral to social media analytics but often processes content containing Personally Identifiable Information (PII), behavioral cues, and metadata raising privacy risks such as surveillance, profiling, and targeted advertising. To systematically assess these risks, we review 203 peer-reviewed papers and propose the *NLP Privacy Risk Identification in Social Media (NLP-PRISM)* framework, which evaluates vulnerabilities across six dimensions: data collection, preprocessing, visibility, fairness, computational risk, and regulatory compliance. Our analysis shows that transformer models achieve F1-scores ranging from 0.58–0.84, but incur a $1\% - 23\%$ drop under privacy-preserving fine-tuning. Using NLP-PRISM, we examine privacy coverage in six NLP tasks: sentiment analysis (16), emotion detection (14), offensive language identification (19), code-mixed processing (39), native language identification (29), and dialect detection (24) revealing substantial gaps in privacy research. We further found a ($\downarrow$ $2\% - 9\%$) trade-off in model utility, MIA AUC (membership inference attacks) 0.81, AIA accuracy 0.75 (attribute inference attacks). Finally, we advocate for stronger anonymization, privacy-aware learning, and fairness-driven training to enable ethical NLP in social media contexts.

1 Introduction

Social media platforms such as X (formerly Twitter), Facebook, and Reddit generate vast volumes of user-generated content (Luca, 2015; Naveen Kumar et al., 2026; Gupta et al., 2024; Noman et al., 2019; Das et al., 2021, 2020), providing valuable resources for natural language processing (NLP) (Adhikari et al., 2023, 2025b, 2022). However, this content's informal, multilingual, and noisy nature characterized by slang, emojis, abbreviations, and code-mixing poses challenges for model robustness and generalization (Gharehchopogh and Khalifelu, 2011; Singh and Manoj, 2017; Bioglio and Pensa, 2022; Cho et al., 2020). Core NLP tasks commonly applied to social media include sentiment analysis (Muhammad et al., 2023; Barnes et al., 2022; Patwa et al., 2020; Yue et al., 2019; Xu et al., 2025), emotion detection (Giorgi et al., 2024; Mohammad and Bravo-Marquez, 2017; Andalibi and Buss, 2020; Ortiz-Clavijo et al., 2023), offensive language identification (Zampieri et al., 2019b, 2020; Ataei et al., 2022), code-mixed processing (Aguilar et al., 2018; Sravani et al., 2021; Yong et al., 2023; Das and Gambäck, 2013; Roy and Kumar, 2025), native language identification (Tetreault et al., 2013; Malmasi et al., 2017), and dialect detection (Malmasi et al., 2016a; Zampieri et al., 2019c; Gaman et al., 2020; Chifu et al., 2024).

These tasks enable applications such as opinion mining (Sharma and Jain, 2020), emotion inference (Canales and Martínez-Barco, 2014; Gill et al., 2008), hate speech moderation (Davidson et al., 2017; Hounsel et al., 2018; He et al., 2024; Vishwamitra et al., 2024; Xu, 2024), multilingual communication (Doğruöz et al., 2021), second language acquisition (Malmasi et al., 2016b; Stokes et al., 2023; Reitmaier et al., 2022; Buschek et al., 2021; Strengers et al., 2020), and sociolinguistic analysis (Faisal et al., 2024; Eleta, 2012). While large language models have advanced these tasks, they also raise concerns about interpretability and user privacy (Meier, 2024; Silva et al., 2022; Adhikari et al., 2025a).

Privacy risks in social media NLP arise from adversaries such as platform operators, scrapers, or state actors capable of accessing raw data or models to perform profiling, de-anonymization, or inference attacks (Beigi and Liu, 2020; Zhang et al., 2018). Given the prevalence of personally identifiable information (PII), geolocation, and behavioral cues in social media text (Lucas and Borisov, 2008), such data is highly vulnerable to surveillance. Moreover, large models have been shown

Findings of the Association for Computational Linguistics: EACL 2026, pages 1519–1541

to memorize sensitive information during training (Das et al., 2024; Pan et al., 2020), prompting privacy-preserving efforts such as differential privacy (DP), federated learning (FL), and data anonymization (Bonneau et al., 2009; Adu-Oppong et al., 2008; Mondal et al., 2014; Moore et al., 2024). DP injects training noise to protect user identity (Wang and Sinnott, 2017), while FL decentralizes learning to reduce data leakage (Mistry et al., 2024; Khalil et al., 2024; Li et al., 2020). However, these techniques remain sparsely adopted in real-world NLP systems. Although prior works (e.g., (Mahendran et al., 2021)) discuss general privacy in NLP, task-specific analyses for social media contexts are limited.

To address this gap, we introduce **NLP Privacy Risk Identification in Social Media (NLP-PRISM)**, a framework for systematically characterizing privacy risks and mitigation strategies across social media NLP tasks. Using NLP-PRISM, we conducted a systematic review of 3,982 papers and selected 203 peer-reviewed works from major NLP, privacy, and HCI venues, guided by research question: *What specific privacy risks emerge when key NLP tasks such as sentiment analysis, emotion detection, offensive language identification, code-mixed text processing, native language identification, and dialect detection are applied to social media data, where models may inadvertently reveal or infer sensitive personal attributes, user identities, or demographic information?*

Key Contributions:

(1) We present the first comprehensive systematization of privacy vulnerabilities in social media NLP, analyzing 203 peer-reviewed studies across six core tasks. Our analysis identifies previously underexplored threats, including latent user re-identification, attribute inference, and representation-level demographic leakage.

(2) We propose *NLP-PRISM*, a six-dimensional framework that captures privacy risks across data collection and usage, preprocessing and anonymization, visibility and profiling, computational vulnerabilities, bias, fairness and discrimination, and regulatory and ethical AI considerations.

(3) We conduct a quantitative assessment of the privacy utility trade-off using transformers (XLM-R, GPT-2, and FLAN-T5) combined with privacy-preserving interventions such as named-entity masking, text perturbation, and noise addition. Adversarial evaluations using membership inference (MIA) and attribute inference (AIA) attacks demonstrate significant information leakage (MIA AUC up to 0.81 and AIA accuracy up to 0.75) and measurable performance degradation (F1 score decreases ranging from 1% to 23%), with identity-centric tasks showing the highest vulnerability.

2 NLP-PRISM: Privacy Risk Framework

The *NLP Privacy Risk Identification in Social Media (NLP-PRISM)* framework offers a comprehensive, multi-dimensional view of privacy-aware NLP by analyzing vulnerabilities, computational risks, and mitigation strategies across the NLP lifecycle. Built upon six interrelated dimensions, it adapts and extends the *Cross-Cultural Privacy Framework* (Ur and Wang, 2013), *OVERRIDE* (Raghavan et al., 2012), and the *Privacy Risk Assessment Framework (PRAF)* (Saka and Das, 2024). The Cross-Cultural Privacy Framework introduces key factors such as cultural norms, user expectations, and legal contexts shaping privacy behaviors like pseudonymity and disclosure. The OVERRIDE model adds context-aware controls adjusting privacy preferences dynamically, a concept vital for social media environments. Finally, PRAF provides a foundation for assessing compliance, usability, and third-party data sharing. Figure 1 summarizes the NLP-PRISM architecture.

2.1 Data Collection and Usage

This initial step identifies how privacy risks vary across NLP tasks that depend on user-generated content, often collected without explicit consent. Such data can expose identities, behavioral traits, or demographics (Mondal et al., 2014). For instance, sentiment and emotion analysis infer users' moods and mental states, while offensive language identification misclassifies minority speech as toxic, reinforcing social biases. Code-mixed data reveal bilingual or cultural traits, and native language or dialect identification can expose ethnicity or regional origin (Acharya et al., 2025). Thus, NLP-PRISM underscores that collection processes are inherently tied to identity leakage, emphasizing the need for privacy-aware sourcing practices.

2.2 Data Preprocessing and Anonymization

Traditional anonymization through entity masking or token replacement often fails to protect against linguistic leakage. Subtle cues like stylistic markers, slang, or phonetic spellings persist, enabling re-identification, particularly in emotion detection and toxicity tasks (Danescu-Niculescu-Mizil et al.,

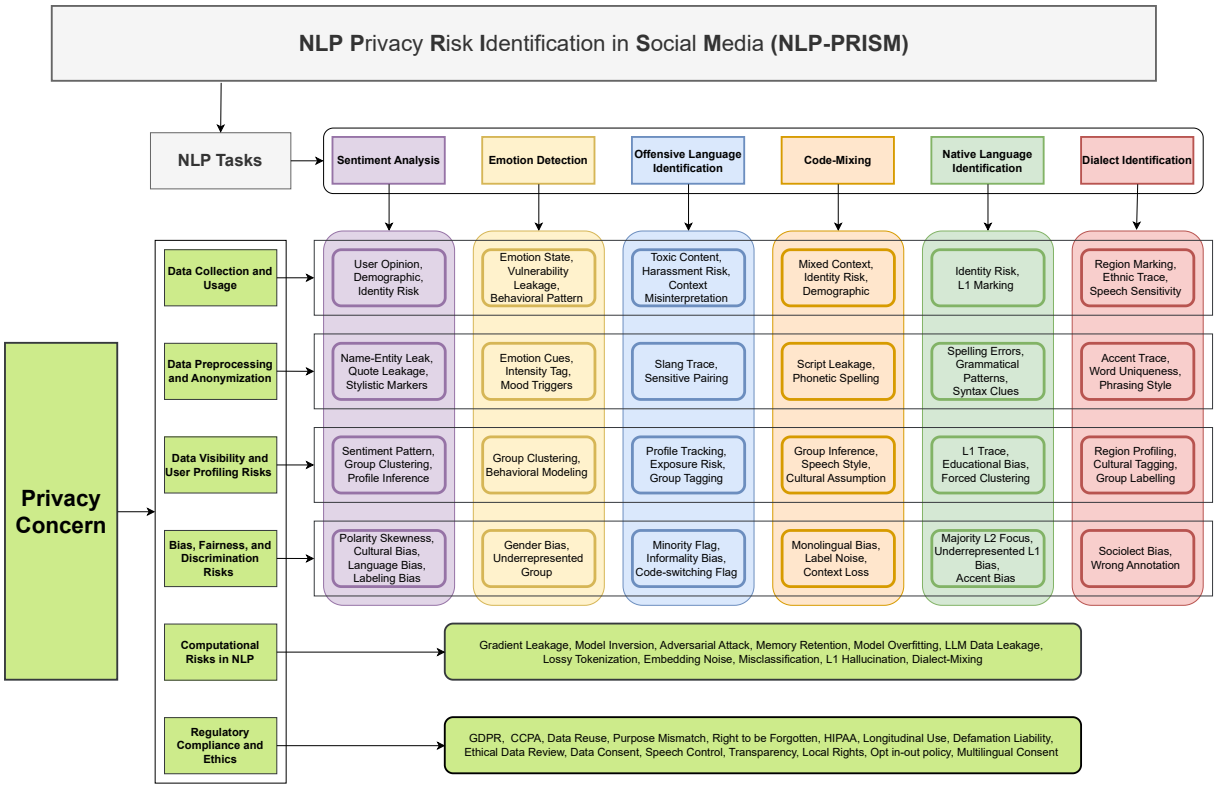

Figure 1: NLP-PRISM Framework Highlighting Privacy Risks in NLP-based Social Media Applications

2011). Dialect and native language models retain accent and syntax traces, while code-mixed data preserve transliteration artifacts and language switches. These residual signals can be exploited by deep learning models for inference attacks (Mei et al., 2017). NLP-PRISM thus maps these limitations across tasks, showing how linguistic artifacts remain potent carriers of implicit identity cues.

2.3 Data Visibility and User Profiling Risks

As preprocessed data enter models, visibility and profiling risks intensify (Yanushkevich et al., 2018). Sentiment and emotion models may infer psychological or political attributes; offensive language models can marginalize minority dialects by labeling them as toxic. Code-mixed data expose bilingual behavior, and dialect identification reveals demographic traits. When these linguistic features are cross-linked with external datasets, they enable user clustering, behavioral profiling, and discrimination (Zafarani and Liu, 2013). NLP-PRISM highlights this transition as a pivotal point where model visibility amplifies privacy exposure.

2.4 Bias, Fairness, and Discrimination Risks

Bias-driven inequities compound privacy harms by reinforcing marginalization (Mai, 2016). Offensive language models often flag dialectal or culturally marked expressions as toxic, while sentiment and emotion models misinterpret cultural affect (Liu, 2023). Underrepresented code-mixed data face monolingual bias and label noise, marginalizing minority voices (Raza Ur Rehman et al., 2025). Native language and dialect identification can misclusters speakers by ethnic or regional traits. In NLP-PRISM, these discriminatory tendencies are viewed as privacy violations and expose users' social or cultural affiliations without consent and exacerbating reputational risks.

2.5 Computational Privacy Risks in NLP

Beyond task-specific issues, NLP architectures face systemic vulnerabilities. Large models fine-tuned on social media text are susceptible to membership inference (Shokri et al., 2017) and model inversion (Fredrikson et al., 2015) attacks that reconstruct sensitive samples. Features like lossy tokenization, embedding leakage, and gradient tracing may expose training data even in federated setups. Moreover, hallucinations in generative models can reproduce verbatim text, heightening risk. Hence, NLP-PRISM classifies computational risks as persistent threats requiring privacy-preserving training and inference controls.

2.6 Regulatory Compliance and Ethics

This dimension examines how legal and ethical standards constrain NLP use. Systems must comply with GDPR (Voigt and Von dem Bussche, 2017), CCPA (Goldman, 2020), and HIPAA, yet enforcement is hindered by opaque data reuse and the inference of unregulated attributes. Sentiment and language identification models may reveal psychological or demographic traits beyond current protections. Ethical lapses arise when model outputs misclassify or stigmatize users, especially in culturally sensitive contexts. Without transparent, opt-in frameworks, NLP deployments risk violating both legal and moral boundaries.

3 Privacy Risk with Large Language Models (LLMs)

While the NLP-PRISM framework characterizes privacy risks across core NLP tasks, the adoption of LLMs introduces additional model-centric privacy threats that cut across the six tasks. Unlike task-specific models, LLMs operate as general-purpose systems trained on massive, heterogeneous corpora (Goswami et al., 2025), often collected with limited transparency or control. Due to their scale, contextual learning, and training dynamics, LLMs are vulnerable to risks such as training data extraction, attribute inference, activation inversion, prompt-based leakage, and user-level inference (Das et al., 2025), which align with and also extend the dimensions captured by NLP-PRISM.

In LLM-based sentiment analysis, privacy risks extend beyond output-level predictions to the exposure of sensitive training content and subjective user expressions. Since models often process highly personal opinions related to health, politics, or lived experiences, leakage at the model level can directly compromise individual privacy. Dai et al. (2025) show that adversaries can recover parts of training data through activation inversion attacks, revealing sentiment-bearing personal text even when raw data is not accessible. In addition, privacy auditing studies by Meng et al. (2025) demonstrate that LLMs may reproduce sentiment-specific expressions from training data under targeted prompting, raising concerns about inadvertent disclosure of sensitive opinions or experiences.

Emotion detection by LLMs is particularly vulnerable to attribute inference attacks, where sensitive emotional or psychological attributes can be inferred from model outputs or latent representations. Because emotional states are closely tied to mental health and personal well-being, even indirect inference poses significant privacy harm. Research on recollection and ranking behavior in LLMs by Meng et al. (2025) shows that emotionally salient content is more likely to be memorized and resurfaced during inference, increasing exposure risks under probing prompts. Furthermore, studies on named-entity and attribute inference by Sutton et al. (2025) demonstrate that even anonymized emotional text can leak private information when processed by LLMs during training or output generation.

For offensive language identification, LLMs amplify privacy risks through their ability to memorize and regenerate toxic, abusive, or sensitive examples encountered during training. Such data often originates from real user interactions and may contain identifiable information or contextual cues. Privacy evaluations of LLMs by (Ye and Luo, 2025) reveal that memorized offensive content indicates a risk of unintended reproduction of harmful or identifying text during downstream use. Additionally, inference-based attacks show that adversaries may deduce whether specific offensive instances or user-generated content were used during training, posing privacy concerns for moderation systems and content reporting pipelines (Mattern et al., 2023).

LLMs trained on multilingual and code-mixed data introduce unique privacy risks due to cross-lingual representation leakage. Code-mixed content often reflects individual identity, community membership, or migration history, making leakage particularly sensitive. Activation inversion attacks demonstrate that sensitive mixed-language training samples can be reconstructed from intermediate representations in large models, even in decentralized or collaborative training settings (Song et al., 2025). Moreover, privacy auditing work indicates that multilingual LLMs may inadvertently encode user-specific language-mixing patterns, increasing the risk of re-identification or demographic inference (Kim et al., 2024).

Native language identification using LLM embeddings raises concerns related to user-level inference, where an attacker can infer whether a particular user, or data from that user, contributed to model training. Since native language often correlates with nationality, ethnicity, or migration status, such inference can have broader societal implications. Recent studies by (Choi et al., 2025) show that LLMs encode subtle linguistic signals that per-

sist even after mitigation attempts, enabling adversaries to infer sensitive language-related attributes beyond the intended task.

Dialect identification tasks further compound privacy risks, as dialectal features often correlate with regional, cultural, or socio-demographic attributes at both individual and community levels. When dialect detection is embedded within LLM pipelines, latent representations may capture fine-grained linguistic markers that users did not explicitly consent to share. Research on LLM privacy leakage by (Ye and Luo, 2025) shows that latent representations can expose such linguistic patterns, enabling attribute inference or reconstruction attacks even when outputs appear benign.

4 Methodology

For this survey, we followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines (McInnes et al., 2018; Moher et al., 2010) to ensure a transparent, systematic, and evidence-based process for identifying, screening, and selecting relevant studies. We also adopted the methodological design established in prior works by the last author of this paper (Huang et al., 2025; Shrestha et al., 2022; Düzgün et al., 2022; Tazi et al., 2022; Das et al., 2022; Zezulak et al., 2023; Tazi et al., 2023, 2022; Grover and Das, 2025; Tazi et al., 2024; Saka and Das, 2025; Podapati et al., 2025; Majumdar and Das, 2021; Agarwal et al., 2025; Kishnani et al., 2023; Jones et al., 2021; Noah and Das, 2021; Das et al., 2019; Shrestha and Das, 2022). In addition, our analysis was guided by the NLP-PRISM framework (detailed in Section 2). Figure 2 provides an overview of the PRISMA-based methodology employed in this study.

4.1 Paper Identification

We queried major academic sources that publish high-impact work on NLP, privacy & security, and HCI, namely *DBLP, ACM Digital Library, IEEE Xplore, Springer,* and *Elsevier*. In addition, we included top-ranked conferences and journals (A*/A venues per CORE ranking[1]) and used Google Scholar to capture influential papers not indexed in those databases. The distribution of retrieved items by venue is shown in Appendix Table 5.

[1] https://portal.core.edu.au/conf-ranks/

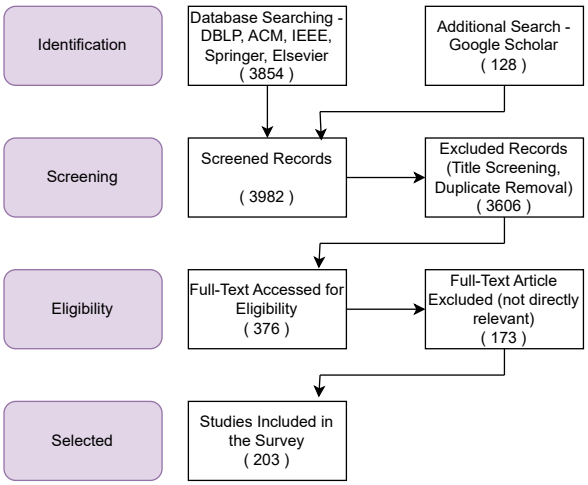

Figure 2: Overview of Data Collection : PRISMA.

4.2 Paper Screening

We constructed structured search queries combining general privacy phrases (e.g., *social media privacy, privacy concerns in social media, social media privacy risk*) with task-specific terms (e.g., *privacy risks in sentiment analysis, emotion detection, offensive language identification, code-mixed text processing, native language identification, dialect identification*). Consequently, our search targeted six core NLP tasks that are especially sensitive in social media contexts: sentiment analysis, emotion detection, offensive language identification, code-mixing, native language identification, and dialect identification. Briefly, these tasks can enable large-scale emotional profiling, deeper mood inferences (including mental-health signals), surveillance or biased moderation, cross-lingual identity leakage, geolocation or demographic inference, and socioeconomic or cultural profiling respectively.

We also included privacy-centric keywords such as *privacy-aware NLP, differential privacy in NLP, federated learning for NLP*, and *anonymization in text processing*. To remain current, the search covered publications up to *December 2024*. Appendix Table 6 reports the task-wise paper counts, including privacy-related studies.

4.3 Paper Exclusion and Final Selection

Our initial retrieval returned **3,982** records. We applied a multi-stage filtering pipeline: title/metadata screening to remove duplicates and clearly irrelevant items, abstract-level screening to assess topical relevance to NLP and privacy, and a full-text eligibility review to judge methodological rigor and contributions to privacy-enhancing techniques (e.g., DP, FL, anonymization, adversarial robust-

ness). Two researchers independently conducted the title/abstract and full-text screenings; disagreements were escalated to a third researcher when necessary. After abstract screening, **3,606** papers were excluded, leaving **376** for full-text review; a further **173** were removed on full-text inspection, producing a final set of **203** studies.

Using the NLP-PRISM framework, we coded each paper across six predefined dimensions: *Data Collection and Usage*, *Data Preprocessing and Anonymization*, *Data Visibility and User Profiling Risks*, *Computational Privacy Risks in NLP*, *Bias, Fairness, and Discrimination Risks*, and *Regulatory Compliance and Ethics*. Two annotators independently labeled a random 25% subset to assess reliability, yielding an inter-annotator agreement of 87.8% and Cohen's Kappa of 0.756, which indicates strong reliability as per McHugh (2012).

Finally, to complement our qualitative synthesis, we ran empirical evaluations that quantify privacy leakage under adversarial settings aligned with NLP-PRISM. Specifically, we measured membership inference and attribute inference attacks on fine-tuned transformer models across the six tasks, and we tested privacy interventions such as named-entity masking and noise addition. The task-wise effects on model utility and attack success are summarized in Table 2.

5 Results

Here we provide an overview of the datasets, benchmark competitions, and computational methodologies. Key trends across these components are summarized in Appendix Table 7, with emphasis on their relevance to privacy challenges.

5.1 Privacy Risks (NLP-PRISM Analysis)

Table 1 summarizes the analyzed papers and the identified privacy risks. A detailed analysis of these risks is provided in Appendix 10.

Data Collection and Usage: Privacy compliance in data collection and usage remains a critical challenge, yet 19 of the 203 papers in our study highlight this. For instance, Xiao et al. (2018) and Sharma and Jain (2020) could improve transparency by clarifying data collection practices in sentiment analysis. Lohar et al. (2021) note privacy issues in public health applications. Likewise, Raihan et al. (2024) and Teodorescu et al. (2023) could strengthen data policies to prevent inadvertent disclosures. Efforts in offensive language identification, such as Zampieri et al. (2024) and Arango et al. (2024), would benefit from stronger control in dataset collection. Similarly, Sigurbergsson and Derczynski (2020), Rosenthal et al. (2021), and Deng et al. (2022) could reinforce privacy protections in large-scale curation. Furthermore, while Hidayatullah et al. (2022) and Bierner (2001) contribute valuable code-mixed resources, clearer data usage measures would improve ethical standards. In NLI, Goldin et al. (2018) and Huang (2015) could refine metadata handling to protect user identities. Finally, Fleisig et al. (2024) emphasize compliance when using proprietary models for dialect analysis. Collectively, these studies underline the necessity of robust frameworks for privacy-conscious data collection and usage.

Data Preprocessing and Anonymization: Effective anonymization is vital for privacy protection, yet only 12 papers offer explicit methods with room for improvement. For example, Lohar et al. (2021) and Xiao et al. (2018) could strengthen anonymization by detailing specific steps, while Teodorescu et al. (2023) and Leonardelli et al. (2021) could integrate clearer anonymization protocols. Research on offensive language, such as Jeong et al. (2022) and Sigurbergsson and Derczynski (2020), would benefit from improved anonymization during annotation. Similarly, Hidayatullah et al. (2022) and Malmasi et al. (2015) could enhance documentation of anonymization practices, ensuring compliance. Moreover, Ferragne and Pellegrino (2007) and Mahmud et al. (2023) provide insights into dialect and offensive language detection, where integrating anonymization would further enhance privacy. Overall, these works highlight the growing awareness of user data protection and the need for transparent anonymization standards.

Data Visibility and User Profiling Risks: Managing data visibility and minimizing profiling risks are crucial for privacy, yet only 4% of the reviewed papers address these issues. For instance, Sharma and Jain (2020) raise concerns about AI-driven sentiment monitoring, underscoring the need for informed consent. Similarly, Lohar et al. (2021) and Leonardelli et al. (2021) discuss risks in sentiment tracking and offensive language classification, calling for clearer data protection mechanisms. In addition, Lukas et al. (2023) highlight data leakage risks in large-scale models, which demand stricter access control. Hidayatullah et al. (2022) and Goldin et al. (2018) illustrate how dataset ex-

Tasks	Privacy Risk
Sentiment Analysis	Data Collection and Usage (Xiao et al., 2018; Sharma and Jain, 2020; Lohar et al., 2021), Data Preprocessing and Anonymization (Lohar et al., 2021; Xiao et al., 2018), Data Visibility and User Profiling Risks (Lohar et al., 2021; Sharma and Jain, 2020), Bias, Fairness, and Discrimination Risks (Konate and Du, 2018; Barbieri et al., 2016), Regulatory Compliance and Ethics (Xiao et al., 2018; Raihan et al., 2023a)
Emotion Detection	Data Collection and Usage (Raihan et al., 2024; Teodorescu et al., 2023; Arango et al., 2024; Zampieri et al., 2024), Data Preprocessing and Anonymization (Teodorescu et al., 2023), Computational Privacy Risks in NLP (Arango et al., 2024), Regulatory Compliance and Ethics (Mohamed et al., 2022; Raihan et al., 2024; Teodorescu et al., 2023)
Offensive Language Identification	Data Collection and Usage (Zampieri et al., 2024; Sigurbergsson and Derczynski, 2020; Rosenthal et al., 2021; Arango et al., 2024; Deng et al., 2022), Data Preprocessing and Anonymization (Leonardelli et al., 2021; Jeong et al., 2022; Sigurbergsson and Derczynski, 2020; Rosenthal et al., 2021), Data Visibility and User Profiling Risks (Leonardelli et al., 2021; Rosenthal et al., 2021), Computational Privacy Risks in NLP (Arango et al., 2024; Xiao et al., 2024; Morabito et al., 2024), Regulatory Compliance and Ethics (Zampieri et al., 2024; Jeong et al., 2022)
Code-Mixing	Data Collection and Usage (Hidayatullah et al., 2022), Data Preprocessing and Anonymization (Hidayatullah et al., 2022), Data Visibility and User Profiling Risks (Hidayatullah et al., 2022; Zhang et al., 2023), Regulatory Compliance and Ethics (Hidayatullah et al., 2022; Garg et al., 2018)
Native Language Identification	Data Collection and Usage (Staicu et al., 2023; Bierner, 2001; Goldin et al., 2018; Jauhiainen et al., 2019), Data Preprocessing and Anonymization (Berzak et al., 2017), Data Visibility and User Profiling Risks (Jauhiainen et al., 2019; Goldin et al., 2018; Aoyama and Schneider, 2024), Regulatory Compliance and Ethics (Jauhiainen et al., 2019; Nguyen et al., 2021)
Dialect Identification	Data Collection and Usage (Jørgensen et al., 2015; Huang, 2015; Fleisig et al., 2024; Barot et al., 2024), Data Preprocessing and Anonymization (Malmasi et al., 2015; Ferragne and Pellegrino, 2007; Mahmud et al., 2023), Computational Privacy Risks in NLP (Mahmud et al., 2023), Bias, Fairness, and Discrimination Risks (Fleisig et al., 2024), Regulatory Compliance and Ethics (Huang, 2015; Fleisig et al., 2024; Faisal et al., 2024; Barot et al., 2024)

Table 1: Survey of Privacy Risks in Social Media NLP Tasks

posure can lead to unintended profiling, advocating privacy-aware dataset structuring. Finally, Zhang et al. (2023), Jauhiainen et al. (2019), and Aoyama and Schneider (2024) emphasize privacy concerns in code-mixed processing and NLI, stressing the need for stronger institutional privacy policies.

Computational Privacy Risks in NLP: Computational privacy concerns arise when data handling, model architectures, or annotation procedures do not fully align with privacy regulations. *We have found this aspect to be addressed in only about* 2% *papers of our study*. Arango et al. (2024) emphasize the importance of ensuring privacy compliance in cross-domain adaptation. Similarly, Xiao et al. (2024) and Morabito et al. (2024) explore the detection of offensive language in large models, and further evaluation of computational privacy risks could refine their approaches. Mahmud et al. (2023) and Rizwan et al. (2020) contribute to low-resource language datasets, where clearer annotation and masking strategies would mitigate privacy concerns.

Bias, Fairness, and Discrimination Risks: Ensuring fairness in NLP remains challenging, as large language models (LLMs) inherit biases from training data, resulting in unequal outcomes across gender, race, dialect, and region. For instance, Fleisig et al. (2024) highlight dialect identification challenges, showing how opaque proprietary model training can lead to performance inequities across linguistic groups. Similarly, Tang et al. (2024) expose gender disparities in language model outputs, illustrating systemic biases from imbalanced datasets. In parallel, Beytía et al. (2022) reveal how gender biases embedded in large-scale resources like Wikipedia propagate into NLP systems, influencing their behavior in consequential ways. These three studies collectively underscore the urgent need for fairness-aware training and critical evaluation of model outputs.

Regulatory Compliance and Ethics: Regulatory Compliance and Ethics are essential for responsible NLP research, yet only 8% of the reviewed papers effectively address these issues. For instance, Xiao et al. (2018) and Teodorescu et al. (2023) could enhance compliance through clearer anonymization, while Raihan et al. (2023a, 2024) stress standardized procedures for code-mixing data collection. Likewise, Huang (2015) emphasize documentation in dialect processing, and Fleisig et al. (2024) with Hidayatullah et al. (2022) highlight ethical risks in language discrimination and

Tasks	XLM-R				GPT-2				FLAN-T5			
	F1	F1(P)	MIA AUC	AIA Acc.	F1	F1(P)	MIA AUC	AIA Acc.	F1	F1(P)	MIA AUC	AIA Acc.
Sentiment Analysis	0.84	0.84	0.62	0.57	0.86	0.85	0.66	0.61	0.33	0.33	0.55	0.50
Emotion Detection	0.62	0.58	0.70	0.63	0.59	0.56	0.68	0.60	0.52	0.31	0.72	0.65
Offensive Language Identification	0.84	0.82	0.74	0.69	0.86	0.83	0.76	0.72	0.79	0.72	0.71	0.65
Code-Mixing	0.63	0.57	0.69	0.66	0.59	0.54	0.68	0.64	0.63	0.60	0.67	0.61
Native Language Identification	0.58	0.35	0.81	0.75	0.49	0.27	0.79	0.73	0.19	0.12	0.76	0.70
Dialect Identification	0.64	0.60	0.73	0.68	0.55	0.42	0.77	0.72	0.41	0.33	0.75	0.70

Table 2: Finetuning (FT) Results. F1: F1 score of direct finetuning; F1(P): F1 score of privacy-preserved finetuning.

code-mixing. Furthermore, Jeong et al. (2022) and Zampieri et al. (2024) call for transparency in offensive language identification, while Jauhiainen et al. (2019) and Mohamed et al. (2022) advocate ethical representation in multilingual datasets. Studies on code-switching, such as Garg et al. (2018) and Nguyen et al. (2021), expose annotation and profiling risks, and Faisal et al. (2024) present a large-scale dataset requiring regulatory scrutiny.

5.2 Experimental Results

Task	XLM-R	GPT-2	FLAN-T5
Sentiment Analysis	Minor	Minor	Minor
Emotion Detection	Moderate	Moderate	High
Offensive Language Identification	Minor	Minor	Moderate
Code-Mixing	Moderate	Moderate	Minor
Native Language Identification	High	High	High
Dialect Identification	Moderate	High	High

Table 3: Performance Trade-off with Transformer Models Across NLP Tasks. • Minor = $\leq 5\%$ • Moderate = $5\% < \text{F1} \leq 10\%$ • High $\geq 10\%$

In line with the privacy evaluation outlined in NLP-PRISM, we conducted experiments to assess the impact of privacy-aware methodologies across six NLP tasks (Table 2). We used the binary-labeled dataset Sentiment140 for Sentiment Analysis (Go et al., 2009), MELD (Poria et al., 2019) with seven emotion classes for Emotion detection, and OLID (Zampieri et al., 2019a) for Offensive Language Identification. Code-Mixing experiments employed the Malayalam-English dataset from Dravidian CodeMix at FIRE 2020 (Chakravarthi et al., 2020a), while TOEFL11 (Blanchard, 2013) with eleven L1s was used for Native Language Identification. Dialect Identification experiment combined English and Spanish dialect corpora from the VarDial 2024 shared task (Chifu et al., 2024), yielding nine dialect labels. Additional experimental results are summarized in Appendix Table 8.

We fine-tuned XLM-R (encoder), GPT-2 (decoder), and FLAN-T5 (encoder–decoder) on task-specific datasets, integrating anonymization and noise addition. Following NER (Sharma et al., 2022), entities such as Person (PER), Organization (ORG), Location (LOC), and Geo-Political Entities (GPE) were masked to reduce re-identification risks. Noise was introduced through 5% text perturbations (character swaps, insertions, deletions, vowel-to-random-consonant substitutions) to mitigate inference attacks. These methods reduced data visibility and profiling risks: PII masking limited exposure, while noise reduced re-identification from model weights, addressing computational privacy. The anonymized, noise-resilient data further supported fairness and ethical AI compliance. Optimal hyperparameters were: batch size 8, learning rate 1e-5, epochs 3, and noise 0.05, with experiments run on a 40 GB institutional GPU cluster.

The privacy-utility trade-off varied across tasks and architectures (Table 3). Sentiment Analysis remained stable with minor degradation in GPT-2, while Emotion Detection showed greater drops, especially for FLAN-T5. Offensive Language Identification was least affected due to strong lexical cues. Code-Mixing tasks showed small declines, with FLAN-T5 displaying higher resilience. Native Language Identification suffered the greatest loss, particularly in GPT-2 and FLAN-T5, whereas XLM-R's multilingual pretraining ensured robustness; it also performed best in Dialect Identification. Overall, encoder-only models like XLM-R demonstrated superior stability under privacy constraints. Finally, membership inference attacks (MIA) (Table 2) were performed using shadow models to test adversarial distinguishability, while attribute inference attacks (AIA) (Table 2) were evaluated via probing classifiers predicting protected attributes (e.g., dialect, gender) from model representations.

6 Discussion

Social media NLP presents multifaceted risks arising from the intersection of language, identity, and algorithmic inference. Our survey addresses the core research question by identifying privacy vulnerabilities across six NLP tasks. However, we found that *only 2%–9% of the papers across differ-*

NLP Task	Data Collection	Anonymization	Visibility & Profiling	Bias & Fairness	Computational Privacy	Compliance & Ethics
Sentiment Analysis	▲	✓	✗	▲	▲	✗
Emotion Detection	▲	✓	✓	▲	✗	✗
Offensive Language	✓	✓	✓	▲	✗	✗
Code-Mixed Processing	✓	▲	✓	▲	✗	✗
Native Language Identification	✓	✓	▲	▲	▲	✗
Dialect Identification	✓	✓	✓	✓	▲	✗

✓= Well-covered in literature, ▲= Partially addressed, ✗= Largely missing.

Table 4: Literature Gaps Across NLP Tasks Based on NLP-PRISM Dimensions

ent tasks explicitly discuss privacy preservation in our comprehensive analysis conducted within the NLP-PRISM framework (Table 4).

Sentiment analysis and emotion detection use behavioral data that infer personal attributes like mental health or political stance. These risks heighten in sentiment (Yelp, Sentiment140, SentMix-3L) and emotion datasets (MELD, EmoMix-3L). *About 63% of sentiment analysis and 62% of emotion detection papers haven't addressed privacy issues such as data collection, anonymization, or profiling.* Offensive language identification faces fairness and privacy risks. Datasets like OLID and SOLID yield high-performing models but embed biases harming minority dialects. Transformers frequently misclassify dialect-rich text, showing representativeness gaps; *73% of studies exhibited such issues leading to user profiling and computational privacy risks.* Code-mixed text processing introduces distinct privacy threats due to multilingual blending and transliteration, which can expose identity across languages. Datasets such as Hinglish, Tamlish, and Tenglish lack standardized anonymization, causing inconsistent protection.

We identified *data anonymization, visibility, and compliance risks in 80% of code-mixed studies*. Native language and dialect identification tasks exhibited a high re-identification potential through linguistic fingerprinting. *In* 55% *of cases, even anonymized text was traceable via stylistic and syntactic cues*, with datasets such as Reddit-L2, AfriDial, and TwitterAAE found to be particularly vulnerable. Dialect datasets further posed geo-linguistic profiling risks, potentially revealing users' locations or social status; such issues were observed in 69% *of dialect identification papers*. Moreover, 40% *of the analyzed papers* cautioned that unanonymized competition datasets and LLMs increase the likelihood of identity disclosure.

While analyzing privacy risks, we also examined mitigation strategies (Appendix Table 15) and found only 19 papers addressing them. Key approaches include identity protection (e.g., named-entity masking, topic filtering) (Teodorescu et al., 2023; Leonardelli et al., 2021; Hidayatullah et al., 2022), model explainability tools (xAI, SHAP, LIME) (Xiao et al., 2018; Mohamed et al., 2022), and secured computation (SMPC, HE) (Arango et al., 2024; Mehta and Passi, 2022). Data refinement methods such as phonetic normalization and aggregation (Jeong et al., 2022; Sigurbergsson and Derczynski, 2020; Berzak et al., 2017; Ferragne and Pellegrino, 2007) further reduce re-identification risks. Beyond the six NLP tasks, NLP-PRISM can also extend to other social and domain-specific applications with privacy risks, enabling systematic analysis of vulnerabilities in areas like misinformation detection, stance classification, user profiling, and privacy-sensitive fields such as healthcare, education, and finance.

7 Conclusion

We conducted a comprehensive analysis of privacy risks in social media NLP, synthesizing evidence from 203 studies across six core tasks. Using the *NLP-PRISM* framework, we identified systemic vulnerabilities across data handling, feature exposure, computational processes, fairness, and regulatory compliance. The findings show that fewer than 10% of studies address privacy in deployment contexts, and only a small subset incorporate formal threat modeling. Empirical evaluations with XLM-R, GPT-2, and FLAN-T5 indicate substantial information leakage (MIA AUC up to 0.81, AIA accuracy up to 0.75) and significant performance degradation (F1 reductions of up to 23%) under privacy-preserving fine-tuning. These results underscore the need for standardized privacy evaluation protocols, reproducible benchmarking, and regulation-aligned auditing to foster secure, fair, and socially responsible NLP systems.

Limitations

In this work, we analyze peer-reviewed research on privacy risks in social media NLP. While peer-reviewed studies provide a reliable foundation for systematic analysis, we acknowledge that important research often appears on pre-print platforms and may not be covered within the scope of this work. Moreover, empirical validation in real-world applications and integration across NLP models remain ongoing challenges. Additionally, emerging generative AI models introduce new risks, including adversarial attacks and memorization vulnerabilities, which require further exploration. Moreover, it is found that, while applying privacy preserving techniques, native language and dialect identifications are the most susceptible in terms of performance due to the compromise of linguistic cues and it is difficult to balance this trade-off. Moving forward, we will refine privacy-aware techniques by analyzing on multiple datasets on downstream NLP tasks, improve fairness-aware learning, and develop adaptive methods that balance privacy with model utility. We will also implement explainable privacy indicators and real-time user controls to enhance transparency and user trust, ensuring more robust and privacy-aware NLP solutions.

Acknowledgment

We would like to acknowledge the Data Agency and Security (DAS) Lab at George Mason University. The opinions expressed in this work are solely those of the authors.

References

Muhammad Abdul Mageed, AbdelRahim Elmadany, Chiyu Zhang, ElMoatez Billah Nagoudi, Houda Bouamor, and Nizar Habash. 2023. NADI 2023: The Fourth Nuanced Arabic Dialect Identification Shared Task. In *Proceedings of ArabicNLP*.

Muhammad Abdul Mageed, Amr Keleg, AbdelRahim Elmadany, Chiyu Zhang, Injy Hamed, Walid Magdy, Houda Bouamor, and Nizar Habash. 2024. NADI 2024: The Fifth Nuanced Arabic Dialect Identification Shared Task. In *Proceedings of ArabicNLP*.

Muhammad Abdul Mageed, Chiyu Zhang, Houda Bouamor, and Nizar Habash. 2020. NADI 2020: The First Nuanced Arabic Dialect Identification Shared Task. In *Proceedings of WANLP*.

Muhammad Abdul-Mageed, Chiyu Zhang, AbdelRahim Elmadany, Houda Bouamor, and Nizar Habash. 2021. NADI 2021: The second nuanced arabic dialect identification shared task. In *Proceedings of WANLP*.

Muhammad Abdul Mageed, Chiyu Zhang, AbdelRahim Elmadany, Houda Bouamor, and Nizar Habash. 2022. NADI 2022: The third nuanced Arabic dialect identification shared task. In *Proceedings of WANLP*.

Poorvi Acharya, J Elizabeth Liebl, Dhiman Goswami, Kai North, Marcos Zampieri, and Antonios Anastasopoulos. 2025. Tracing l1 interference in english learner writing: A longitudinal corpus with error annotations. In *Proceedings of EMNLP*.

Andrick Adhikari, Sanchari Das, and Rinku Dewri. 2022. Privacy policy analysis with sentence classification. In *2022 19th Annual International Conference on Privacy, Security & Trust (PST)*, pages 1–10. IEEE.

Andrick Adhikari, Sanchari Das, and Rinku Dewri. 2023. Evolution of composition, readability, and structure of privacy policies over two decades. *Proceedings on Privacy Enhancing Technologies*, 3:138ś153.

Andrick Adhikari, Sanchari Das, and Rinku Dewri. 2025a. Natural language processing of privacy policies: A survey. *arXiv preprint arXiv:2501.10319*.

Andrick Adhikari, Sanchari Das, and Rinku Dewri. 2025b. Policypulse: Precision semantic role extraction for enhanced privacy policy comprehension. *Proceedings of NDSS*.

Fabeah Adu-Oppong, Casey K Gardiner, Apu Kapadia, and Patrick P Tsang. 2008. Social circles: Tackling privacy in social networks. In *Proceedings of SOUPS*.

Noëmi Aepli, Antonios Anastasopoulos, Adrian Chifu, William Domingues, Fahim Faisal, Mihaela Găman, Radu Tudor Ionescu, and Yves Scherrer. 2022. Findings of the VarDial evaluation campaign 2022. In *Proceedings of ICCL (VarDial)*.

Neha Agarwal, Ethan Mackin, Faiza Tazi, Mayank Grover, Rutuja More, and Sanchari Das. 2025. Systematic literature review of vulnerabilities and defenses in vpns, tor, and web browsers. In *International Conference on Information Systems Security*, pages 357–375. Springer.

Gustavo Aguilar, Fahad AlGhamdi, Victor Soto, Mona Diab, Julia Hirschberg, and Thamar Solorio. 2018. Overview of the CALCS 2018 Shared Task: Named Entity Recognition on Code-switched Data. In *Proceedings of CALCS*.

Mohamed Alloghani, Mohammed M Alani, Dhiya Al-Jumeily, Thar Baker, Jamila Mustafina, Abir Hussain, and Ahmed J Aljaaf. 2019. A systematic review on the status and progress of homomorphic encryption technologies. *Journal of Information Security and Applications*, 48.

M Anand Kumar, HB Barathi Ganesh, Shivkaran Singh, KP Soman, and Paolo Rosso. 2017. Overview of the inli pan at fire-2017 track on indian native language identification. In *CEUR workshop proceedings*.

Nazanin Andalibi and Justin Buss. 2020. The human in emotion recognition on social media: Attitudes, outcomes, risks. In *Proceedings of CHI*.

Tatsuya Aoyama and Nathan Schneider. 2024. Modeling nonnative sentence processing with l2 language models. In *Proceedings of EMNLP*.

Aymé Arango, Parisa Kaghazgaran, Sheikh Muhammad Sarwar, Vanessa Murdock, and Cj Lee. 2024. Multifold: Multi-source domain adaption for offensive language detection. In *Proceedings of ICWSM*, volume 18, pages 86–99.

Taha Shangipour Ataei, Kamyar Darvishi, Soroush Javdan, Amin Pourdabiri, Behrouz Minaei-Bidgoli, and Mohammad Taher Pilehvar. 2022. Pars-off: a benchmark for offensive language detection on farsi social media. *IEEE Transactions on Affective Computing*, 14:2787–2795.

Premjth B, Bharathi Raja Chakravarthi, Prasanna Kumar Kumaresan, Saranya Rajiakodi, Sai Prashanth Karnati, Sai Rishith Reddy Mangamuru, and Janakiram Chandu. 2024. Findings of the shared task on hate and offensive language detection in telugu codemixed text (hold-telugu). In *Proceedings of DravidianLangTech*.

F. Balouchzahi, S. Butt, A. Hegde, N. Ashraf, H.l. Shashirekha, Grigori Sidorov, and Alexander Gelbukh. 2022. Overview of CoLI-kanglish: Word level language identification in code-mixed Kannada-English texts at ICON 2022. In *Proceedings of ICON*.

Francesco Barbieri, Valerio Basile, Danilo Croce, Malvina Nissim, Nicole Novielli, Viviana Patti, and 1 others. 2016. Overview of the evalita 2016 sentiment polarity classification task. In *CEUR Workshop Proceedings*.

Jeremy Barnes, Laura Oberlaender, Enrica Troiano, Andrey Kutuzov, Jan Buchmann, Rodrigo Agerri, Lilja Øvrelid, and Erik Velldal. 2022. SemEval 2022 task 10: Structured sentiment analysis. In *Proceedings of SemEval*.

Jay Barot, Ali Allami, Ming Yin, and Dan Lin. 2024. Tonecheck: Unveiling the impact of dialects in privacy policy. In *Proceedings of SACMAT*.

Ghazaleh Beigi and Huan Liu. 2020. A survey on privacy in social media: Identification, mitigation, and applications. *ACM Transactions on Data Science*, 1:1–38.

Mohamed Berrimi, Abdelouahab Moussaoui, Mourad Oussalah, and Mohamed Saidi. 2020. Arabic dialects identification: North african dialects case study. In *Proceedings of IAM*.

Yevgeni Berzak, Chie Nakamura, Suzanne Flynn, and Boris Katz. 2017. Predicting native language from gaze. In *Proceedings of ACL*.

Pablo Beytía, Pushkal Agarwal, Miriam Redi, and Vivek K Singh. 2022. Visual gender biases in wikipedia: A systematic evaluation across the ten most spoken languages. In *Proceedings of ICWSM*.

Gann Bierner. 2001. Alternative phrases and natural language information retrieval. In *Proceedings of ACL*.

Livio Bioglio and Ruggero G Pensa. 2022. Analysis and classification of privacy-sensitive content in social media posts. *EPJ Data Science*, 11:12.

D Blanchard. 2013. Toefl11: A corpus of non-native english. *Educational Testing Service*.

Su Lin Blodgett, Lisa Green, and Brendan O'Connor. 2016. Demographic dialectal variation in social media: A case study of african-american english. In *Proceedings of EMNLP*.

Joseph Bonneau, Jonathan Anderson, and Luke Church. 2009. Privacy suites: shared privacy for social networks. In *Proceedings of SOUPS*.

Daniel Buschek, Martin Zürn, and Malin Eiband. 2021. The impact of multiple parallel phrase suggestions on email input and composition behaviour of native and non-native english writers. In *Proceedings of CHI*.

Lea Canales and Patricio Martínez-Barco. 2014. Emotion detection from text: A survey. In *Proceedings of JISIC*, pages 37–43.

Bharathi Raja Chakravarthi, Navya Jose, Shardul Suryawanshi, Elizabeth Sherly, and John Philip McCrae. 2020a. A sentiment analysis dataset for code-mixed malayalam-english. In *Joint Proceedings of SLTU and CCURL*.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Navya Jose, Anand Kumar M, Thomas Mandl, Prasanna Kumar Kumaresan, Rahul Ponnusamy, Hariharan R L, John P. McCrae, and Elizabeth Sherly. 2021. Findings of the shared task on offensive language identification in Tamil, Malayalam, and Kannada. In *Proceedings of DravidianLangTech*.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Vigneshwaran Muralidaran, Shardul Suryawanshi, Navya Jose, Elizabeth Sherly, and John P McCrae. 2020b. Overview of the track on sentiment analysis for dravidian languages in code-mixed text. In *Proceedings of FIRE*.

Bharathi Raja Chakravarthi, N Sripriya, B Bharathi, K Nandhini, S Chinnaudayar Navaneethakrishnan, Thenmozhi Durairaj, Rahul Ponnusamy, Prasanna Kumar Kumaresan, Kishore Kumar Ponnusamy, and Charmathi Rajkumar. 2023. Overview of the shared task on sarcasm identification of dravidian languages (malayalam and tamil) in dravidiancodemix. In *Proceedings of FIRE*.

Khyathi Raghavi Chandu, Ekaterina Loginova, Vishal Gupta, Josef van Genabith, G ünter Neuman, Manoj Chinnakotla, Eric Nyberg, and Alan Black. 2018. Code-mixed question answering challenge: Crowdsourcing data and techniques. *ACL 2018*, page 29.

Adrian Gabriel Chifu, Goran Glavaš, Radu Tudor Ionescu, Nikola Ljubešić, Aleksandra Miletić, Filip Miletić, Yves Scherrer, and Ivan Vulić. 2024. VarDial evaluation campaign 2024: Commonsense reasoning in dialects and multi-label similar language identification. In *Proceedings of VarDial*.

Hichang Cho, Pengxiang Li, and Zhang Hao Goh. 2020. Privacy risks, emotions, and social media: A coping model of online privacy. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 27:1–28.

Yujin Choi, Youngjoo Park, Junyoung Byun, Jaewook Lee, and Jinseong Park. 2025. Safeguarding privacy of retrieval data against membership inference attacks: Is this query too close to home? In *Findings of EMNLP*.

Chenxi Dai, Lin Lu, and Pan Zhou. 2025. Stealing training data from large language models in decentralized training through activation inversion attack. In *Proceedings of ACL*.

Cristian Danescu-Niculescu-Mizil, Michael Gamon, and Susan Dumais. 2011. Mark my words! linguistic style accommodation in social media. In *Proceedings of WWW*, pages 745–754.

Amitava Das and Björn Gambäck. 2013. Code-mixing in social media text. *Traitement Automatique des Langues*, 54:41–64.

Badhan Chandra Das, M Hadi Amini, and Yanzhao Wu. 2024. Security and privacy challenges of large language models: A survey. *ACM Computing Surveys*.

Badhan Chandra Das, M Hadi Amini, and Yanzhao Wu. 2025. Security and privacy challenges of large language models: A survey. *ACM Computing Surveys*, 57:1–39.

Sanchari Das, Tousif Ahmed, Apu Kapadia, and Sameer Patil. 2021. Does this photo make me look good? how posters, outsiders, and friends evaluate social media photo posts. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–32.

Sanchari Das, Andrew Kim, and Sayar Karmakar. 2020. Change-point analysis of cyberbullying-related twitter discussions during covid-19. In *Proceedings of the 16th Annual Social Informatics Research Symposium ("Sociotechnical Change Agents: ICTs, Sustainability, and Global Challenges") in Conjunction with the 83rd Association for Information Science and Technology (ASIS&T)*.

Sanchari Das, Andrew Kim, Zachary Tingle, and Christena Nippert-Eng. 2019. All about phishing exploring user research through a systematic literature review. In *Proceedings of the Thirteenth International Symposium on Human Aspects of Information Security & Assurance (HAISA 2019)*.

Sanchari Das and 1 others. 2022. Sok: a proposal for incorporating accessible gamified cybersecurity awareness training informed by a systematic literature review. In *Proceedings of the workshop on usable security and privacy (USEC)*.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of ICWSM*, volume 11.

Jiawen Deng, Jingyan Zhou, Hao Sun, Chujie Zheng, Fei Mi, Helen Meng, and Minlie Huang. 2022. COLD: A benchmark for Chinese offensive language detection. In *Proceedings of EMNLP*.

Amirita Dewani, Mohsin Ali Memon, Sania Bhatti, Adel Sulaiman, Mohammed Hamdi, Hani Alshahrani, Abdullah Alghamdi, and Asadullah Shaikh. 2023. Detection of cyberbullying patterns in low resource colloquial roman urdu microtext using natural language processing, machine learning, and ensemble techniques. *Applied Sciences*, 13.

Elisa Di Nuovo, Cristina Bosco, Elisa Corino, and 1 others. 2020. How good are humans at native language identification? a case study on italian l2 writings. In *Proceedings of CLiC-it*.

A Seza Doğruöz, Sunayana Sitaram, Barbara Bullock, and Almeida Jacqueline Toribio. 2021. A survey of code-switching: Linguistic and social perspectives for language technologies. In *Proceedings of ACL - IJCNLP*.

Reyhan Düzgün, Naheem Noah, Peter Mayer, Sanchari Das, and Melanie Volkamer. 2022. Sok: A systematic literature review of knowledge-based authentication on augmented reality head-mounted displays. In *Proceedings of the 17th International Conference on Availability, Reliability and Security*, pages 1–12.

Irene Eleta. 2012. Multilingual use of twitter: social networks and language choice. In *Proceedings of CSCW*.

Fahim Faisal, Orevaoghene Ahia, Aarohi Srivastava, Kabir Ahuja, David Chiang, Yulia Tsvetkov, and Antonios Anastasopoulos. 2024. DIALECTBENCH: An NLP benchmark for dialects, varieties, and closely-related languages. In *Proceedings of ACL*.

Emmanuel Ferragne and François Pellegrino. 2007. Automatic dialect identification: A study of british english. *Speaker Classification II: Selected Projects*.

Eve Fleisig, Genevieve Smith, Madeline Bossi, Ishita Rustagi, Xavier Yin, and Dan Klein. 2024. Linguistic bias in ChatGPT: Language models reinforce dialect discrimination. In *Proceedings of EMNLP*.

Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. 2015. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of CCS*.

Mihaela Gaman, Dirk Hovy, Radu Tudor Ionescu, Heidi Jauhiainen, Tommi Jauhiainen, Krister Lindén, Nikola Ljubešić, Niko Partanen, Christoph Purschke, Yves Scherrer, and Marcos Zampieri. 2020. A report on the VarDial evaluation campaign 2020. In *Proceedings of VarDial*.

Rupali Gangarde, Amit Sharma, Ambika Pawar, Rahul Joshi, and Sudhanshu Gonge. 2021. Privacy preservation in online social networks using multiple-graph-properties-based clustering to ensure k-anonymity, l-diversity, and t-closeness. *Electronics*, 10.

Saurabh Garg, Tanmay Parekh, and Preethi Jyothi. 2018. Code-switched language models using dual rnns and same-source pretraining. In *Proceedings of EMNLP*.

Binyam Gebrekidan Gebre, Marcos Zampieri, Peter Wittenburg, and Tom Heskes. 2013. Improving native language identification with tf-idf weighting. In *Proceedings of BEA*.

Farhad Soleimanian Gharehchopogh and Zeinab Abbasi Khalifelu. 2011. Analysis and evaluation of unstructured data: text mining versus natural language processing. In *Proceedings of AICT*. IEEE.

Koyel Ghosh, Apurbalal Senapati, and Aditya Shankar Pal. 2023. Annihilate hates (task 4 hasoc 2023): Hate speech detection in assamese bengali and bodo languages. In *FIRE (Working Notes)*.

Alastair J Gill, Robert M French, Darren Gergle, and Jon Oberlander. 2008. The language of emotion in short blog texts. In *Proceedings of CSCW*.

Salvatore Giorgi, João Sedoc, Valentin Barriere, and Shabnam Tafreshi. 2024. Findings of WASSA 2024 shared task on empathy and personality detection in interactions. In *Proceedings of WASSA*.

Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision. *CS224N project report, Stanford*, 1:2009.

Gili Goldin, Ella Rabinovich, and Shuly Wintner. 2018. Native language identification with user generated content. In *Proceedings of EMNLP*.

Eric Goldman. 2020. An introduction to the california consumer privacy act (ccpa). *Santa Clara Univ. Legal Studies Research Paper*.

Dhiman Goswami, Md Nishat Raihan, Antara Mahmud, Antonios Anastasopoulos, and Marcos Zampieri. 2023. Offmix-3l: A novel code-mixed test dataset in bangla-english-hindi for offensive language identification. In *Proceedings of SocialNLP*.

Dhiman Goswami, Marcos Zampieri, Kai North, Shervin Malmasi, and Antonios Anastasopoulos. 2025. Multilingual native language identification with large language models. In *Proceedings of NAACL (SRW)*, pages 193–199.

Mayank Grover and Sanchari Das. 2025. Sok: a systematic review of privacy and security in healthcare robotics. In *International Conference on Social Robotics*, pages 212–234. Springer.

Naman Gupta, Kate Walsh, Sanchari Das, and Rahul Chatterjee. 2024. " i really just leaned on my community for support": Barriers, challenges, and coping mechanisms used by survivors of {Technology-Facilitated} abuse to seek social support. In *33rd USENIX Security Symposium (USENIX Security 24)*, pages 4981–4998.

Xinlei He, Savvas Zannettou, Yun Shen, and Yang Zhang. 2024. You only prompt once: On the capabilities of prompt learning on large language models to tackle toxic content. In *Proceedings of IEEE S&P*.

Ahmad Fathan Hidayatullah, Atika Qazi, Daphne Teck Ching Lai, and Rosyzie Anna Apong. 2022. A systematic review on language identification of code-mixed text: techniques, data availability, challenges, and framework development. *IEEE access*, 10.

Austin Hounsel, Prateek Mittal, and Nick Feamster. 2018. Automatically generating a large,{Culture-Specific} blocklist for china. In *Proceedings of FOCI*.

Fei Huang. 2015. Improved arabic dialect classification with social media data. In *Proceedings of EMNLP*.

Yue Huang, Marthie Grobler, Lauren S Ferro, Georgia Psaroulis, Sanchari Das, Jing Wei, and Helge Janicke. 2025. Systemization of knowledge (sok): Goals, coverage, and evaluation in cybersecurity and privacy games. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–27.

Scott Jarvis, Yves Bestgen, and Steve Pepper. 2013. Maximizing classification accuracy in native language identification. In *Proceedings of BEA*.

Tommi Jauhiainen, Marco Lui, Marcos Zampieri, Timothy Baldwin, and Krister Lindén. 2019. Automatic language identification in texts: A survey. *Journal of Artificial Intelligence Research*, 65.

Younghoon Jeong, Juhyun Oh, Jongwon Lee, Jaimeen Ahn, Jihyung Moon, Sungjoon Park, and Alice Oh. 2022. KOLD: Korean offensive language dataset. In *Proceedings of EMNLP*.

John M Jones, Reyhan Duezguen, Peter Mayer, Melanie Volkamer, and Sanchari Das. 2021. A literature review on virtual reality authentication. In *International Symposium on Human Aspects of Information Security and Assurance*, pages 189–198. Springer.

Anna Jørgensen, Dirk Hovy, and Anders Søgaard. 2015. Challenges of studying and processing dialects in social media. In *Proceedings of WNUT*.

Samar Samir Khalil, Noha S Tawfik, and Marco Spruit. 2024. Federated learning for privacy-preserving depression detection with multilingual language models in social media posts. *Patterns*, 5.

Hwichan Kim, Jun Suzuki, Tosho Hirasawa, and Mamoru Komachi. 2024. Pruning multilingual large language models for multilingual inference. In *Findings of EMNLP*.

Urvashi Kishnani, Srinidhi Madabhushi, and Sanchari Das. 2023. Blockchain in oil and gas supply chain: a literature review from user security and privacy perspective. In *International Symposium on Human Aspects of Information Security and Assurance*, pages 296–309. Springer.

Arouna Konate and Ruiying Du. 2018. Sentiment analysis of code-mixed bambara-french social media text using deep learning techniques. *Wuhan University Journal of Natural Sciences*, 23.

Moshe Koppel, Jonathan Schler, and Kfir Zigdon. 2005. Determining an author's native language by mining a text for errors. In *Proceedings of ACM SIGKDD*.

Ritesh Kumar, Atul Ojha, Shervin Malmasi, and Marcos Zampieri. 2020. Evaluating aggression identification in social media. In *Proceedings of TRAC*.

Ritesh Kumar, Atul Kr. Ojha, Shervin Malmasi, and Marcos Zampieri. 2018. Benchmarking aggression identification in social media. In *Proceedings of TRAC*.

Shivani Kumar, Md. Shad Akhtar, Erik Cambria, and Tanmoy Chakraborty. 2024. SemEval 2024 - task 10: Emotion discovery and reasoning its flip in conversation (EDiReF). In *Proceedings of SemEval*.

BS Sowmya Lakshmi and BR Shambhavi. 2017. An automatic language identification system for code-mixed english-kannada social media text. In *Proceedings of CSITSS*.

Priyadarshini Lamabam and Kunal Chakma. 2016. A language identification system for code-mixed english-manipuri social media text. In *Proceedings of ICETECH*.

Lung-Hao Lee, Liang-Chih Yu, Suge Wang, and Jian Liao. 2024. Overview of the sighan 2024 shared task for chinese dimensional aspect-based sentiment analysis. In *Proceedings of SIGHAN*.

Elisa Leonardelli, Stefano Menini, Alessio Palmero Aprosio, Marco Guerini, and Sara Tonelli. 2021. Agreeing to disagree: Annotating offensive language datasets with annotators' disagreement. In *Proceedings of EMNLP*.

Li Li, Yuxi Fan, Mike Tse, and Kuo-Yi Lin. 2020. A review of applications in federated learning. *Computers & Industrial Engineering*, 149.

Zhaoming Liu. 2023. Cultural bias in large language models: A comprehensive analysis and mitigation strategies. *Journal of Transcultural Communication*, 3(2):224–244.

Pintu Lohar, Guodong Xie, Malika Bendechache, Rob Brennan, Edoardo Celeste, Ramona Trestian, and Irina Tal. 2021. Irish attitudes toward covid tracker app & privacy: sentiment analysis on twitter and survey data. In *Proceedings of ARES*.

Michael Luca. 2015. User-generated content and social media. In *Handbook of media Economics*, volume 1, pages 563–592. Elsevier.

Matthew M Lucas and Nikita Borisov. 2008. Flybynight: mitigating the privacy risks of social networking. In *Proceedings of WPES*.

Nils Lukas, Ahmed Salem, Robert Sim, Shruti Tople, Lukas Wutschitz, and Santiago Zanella-Béguelin. 2023. Analyzing leakage of personally identifiable information in language models. In *Proceedings of S&P*. IEEE.

Darshini Mahendran, Changqing Luo, and Bridget T Mcinnes. 2021. Privacy-preservation in the context of natural language processing. *IEEE Access*, 9:147600–147612.

Tanjim Mahmud, Michal Ptaszynski, Juuso Eronen, and Fumito Masui. 2023. Cyberbullying detection for low-resource languages and dialects: Review of the state of the art. *Information Processing & Management*, 60.

Jens-Erik Mai. 2016. Marginalization and exclusion: Unraveling systemic bias in classification. *Knowledge Organization*, 43(5).

Ritajit Majumdar and Sanchari Das. 2021. Sok: An evaluation of quantum authentication through systematic literature review. In *Proceedings of the Workshop on Usable Security and Privacy (USEC)*.

Shervin Malmasi, Keelan Evanini, Aoife Cahill, Joel Tetreault, Robert Pugh, Christopher Hamill, Diane Napolitano, and Yao Qian. 2017. A report on the 2017 native language identification shared task. In *Proceedings of BEA*.

Shervin Malmasi, Eshrag Refaee, and Mark Dras. 2015. Arabic dialect identification using a parallel multidialectal corpus. In *Proceedings of PACLING*.

Shervin Malmasi, Marcos Zampieri, Nikola Ljubešić, Preslav Nakov, Ahmed Ali, and Jörg Tiedemann. 2016a. Discriminating between similar languages and Arabic dialect identification: A report on the third DSL shared task. In *Proceedings of VarDial*.

Shervin Malmasi and 1 others. 2016b. *Native language identification: explorations and applications*. Ph.D. thesis, Macquarie University, Faculty of Science and Engineering, Department of ....

Thomas Mandl, Sandip Modha, Anand Kumar M, and Bharathi Raja Chakravarthi. 2020. Overview of the hasoc track at fire 2020: Hate speech and offensive language identification in tamil, malayalam, hindi, english and german. In *Proceedings of FIRE*.

Thomas Mandl, Sandip Modha, Prasenjit Majumder, Daksh Patel, Mohana Dave, Chintak Mandlia, and Aditya Patel. 2019. Overview of the hasoc track at fire 2019: Hate speech and offensive content identification in indo-european languages. In *Proceedings of FIRE*.

Justus Mattern, Fatemehsadat Mireshghallah, Zhijing Jin, Bernhard Schölkopf, Mrinmaya Sachan, and Taylor Berg-Kirkpatrick. 2023. Membership inference attacks against language models via neighbourhood comparison. In *Findings of ACL*.

Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica*, 22.

Matthew DF McInnes, David Moher, Brett D Thombs, Trevor A McGrath, Patrick M Bossuyt, Tammy Clifford, Jérémie F Cohen, Jonathan J Deeks, Constantine Gatsonis, Lotty Hooft, and 1 others. 2018. Preferred reporting items for a systematic review and meta-analysis of diagnostic test accuracy studies: the prisma-dta statement. *Jama*, 319:388–396.

Seifeddine Mechti, Ayoub Abbassi, Lamia Hadrich Belguith, and Rim Faiz. 2016. An empirical method using features combination for arabic native language identification. In *Proceedings of AICCSA*.

Harshkumar Mehta and Kalpdrum Passi. 2022. Social media hate speech detection using explainable artificial intelligence (xai). *Algorithms*, 15.

Bo Mei, Yinhao Xiao, Hong Li, Xiuzhen Cheng, and Yunchuan Sun. 2017. Inference attacks based on neural networks in social networks. In *Proceedings of HotWeb*, pages 1–6.

Raphael Meier. 2024. Llm-aided social media influence operations. *Large Language Models in Cybersecurity: Threats, Exposure and Mitigation*, pages 105–112.

Wenlong Meng, Guo Zhenyuan, Lenan Wu, Chen Gong, Wenyan Liu, Weixian Li, Chengkun Wei, and Wenzhi Chen. 2025. Rr: Unveiling llm training privacy through recollection and ranking. In *Findings of ACL*.

Durjoy Mistry, Jayonto Dutta Plabon, Bidita Sarkar Diba, Saddam Mukta, and MF Mridha. 2024. Federated learning-based architecture for personalized next emoji prediction for social media comments. *IEEE Access*.

Sandip Modha, Thomas Mandl, Prasenjit Majumder, Shrey Satapara, Tithi Patel, and Hiren Madhu. 2022. Overview of the hasoc subtrack at fire 2022: Identification of conversational hate-speech in hindi-english code-mixed and german language. In *FIRE (Working Notes)*.

Youssef Mohamed, Mohamed Abdelfattah, Shyma Alhuwaider, Feifan Li, Xiangliang Zhang, Kenneth Church, and Mohamed Elhoseiny. 2022. Artelingo: A million emotion annotations of wikiart with emphasis on diversity over language and culture. In *Proceedings of EMNLP*.

Saif Mohammad and Felipe Bravo-Marquez. 2017. WASSA-2017 shared task on emotion intensity. In *Proceedings of WASSA*.

David Moher, Alessandro Liberati, Jennifer Tetzlaff, Douglas G Altman, Prisma Group, and 1 others. 2010. Preferred reporting items for systematic reviews and meta-analyses: the prisma statement. *International journal of surgery*, 8:336–341.

Mainack Mondal, Yabing Liu, Bimal Viswanath, Krishna P Gummadi, and Alan Mislove. 2014. Understanding and specifying social access control lists. In *Proceedings of SOUPS*.

Lachlan Moore, Tatsuya Mori, and Ayako A Hasegawa. 2024. Negative effects of social triggers on user security and privacy behaviors. In *Proceedings of SOUPS*.

Robert Morabito, Sangmitra Madhusudan, Tyler McDonald, and Ali Emami. 2024. Stop! benchmarking large language models with sensitivity testing on offensive progressions. In *Proceedings of EMNLP*.

Shamsuddeen Hassan Muhammad, Idris Abdulmumin, Seid Muhie Yimam, David Ifeoluwa Adelani, Ibrahim Said Ahmad, Nedjma Ousidhoum, Abinew Ali Ayele, Saif Mohammad, Meriem Beloucif, and Sebastian Ruder. 2023. SemEval-2023 task 12: Sentiment analysis for African languages (AfriSenti-SemEval). In *Proceedings of SemEval*.

Cindy Nabila and Andi Idayani. 2022. An analysis of indonesian-eglish code mixing used in social media (twitter). *J-SHMIC: Journal of English for Academic*, 9.

Jai Kruthunz Naveen Kumar, Aishwarya Surani, Harkirat Singh, and Sanchari Das. 2026. Privacy discourse and emotional dynamics in mental health information interaction on reddit. In *Proceedings of the 2026 ACM SIGIR Conference on Human Information Interaction and Retrieval (CHIIR), Seattle, WA, USA*.

Li Nguyen, Christopher Bryant, Sana Kidwai, and Theresa Biberauer. 2021. Automatic language identification in code-switched hindi-english social media text. *Journal of Open Humanities Data*, 7.

Naheem Noah and Sanchari Das. 2021. Exploring evolution of augmented and virtual reality education space in 2020 through systematic literature review. *Computer Animation and Virtual Worlds*, 32(3-4):e2020.

Behnaz Nojavanasghari, Tadas Baltrušaitis, Charles E Hughes, and Louis-Philippe Morency. 2016. Emoreact: a multimodal approach and dataset for recognizing emotional responses in children. In *Proceedings of ICMI*.

Abu Saleh Md Noman, Sanchari Das, and Sameer Patil. 2019. Techies against facebook: understanding negative sentiment toward facebook via user generated content. In *Proceedings of CHI*.

Luis Felipe Ortiz-Clavijo, Carlos Julián Gallego-Duque, Juan Camilo David-Diaz, and Andrés Felipe Ortiz-Zamora. 2023. Implications of emotion recognition technologies: Balancing privacy and public safety. *IEEE Technology and Society Magazine*, 42:69–75.

Ronghao Pan, José Antonio García-Díaz, Miguel Ángel Rondríguez-García, Francisco García-Sánchez, and Rafael Valencia-García. 2024. Overview of emospeech at iberlef 2024: Multimodal speech-text emotion recognition in spanish. *Procesamiento del Lenguaje Natural*, 73:359–368.

Xudong Pan, Mi Zhang, Shouling Ji, and Min Yang. 2020. Privacy risks of general-purpose language models. In *Proceedings of IEEE S&P*.

Parth Patwa, Gustavo Aguilar, Sudipta Kar, Suraj Pandey, Srinivas PYKL, Björn Gambäck, Tanmoy Chakraborty, Thamar Solorio, and Amitava Das. 2020. Semeval-2020 task 9: Overview of sentiment analysis of code-mixed tweets. In *Proceedings of SemEval*.

Zesis Pitenis, Marcos Zampieri, and Tharindu Ranasinghe. 2020. Offensive language identification in Greek. In *Proceedings of LREC*.

Vyoma Harshitha Podapati, Divyansh Nigam, and Sanchari Das. 2025. Sok: a systematic review of context- and behavior-aware adaptive authentication in mobile environments. In *International Symposium on Human Aspects of Information Security and Assurance*, pages 406–419. Springer.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. Meld: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of ACL*.

Ella Rabinovich, Yulia Tsvetkov, and Shuly Wintner. 2018. Native language cognate effects on second language lexical choice. *Transactions of the Association for Computational Linguistics*, 6:329–342.

Kasturi Rangan Raghavan, Supriyo Chakraborty, Mani Srivastava, and Harris Teague. 2012. Override: A mobile privacy framework for context-driven perturbation and synthesis of sensor data streams. In *Proceedings of SenSys*.

Md Nishat Raihan, Dhiman Goswami, Antara Mahmud, Antonios Anastasopoulos, and Marcos Zampieri. 2023a. SentMix-3L: A novel code-mixed test dataset in bangla-english-hindi for sentiment analysis. In *Proceedings of SEALP*.

Md Nishat Raihan, Umma Hani Tanmoy, Anika Binte Islam, Kai North, Tharindu Ranasinghe, Antonios Anastasopoulos, and Marcos Zampieri. 2023b. Offensive language identification in transliterated and code-mixed bangla. In *Proceedings of BLP*.

Nishat Raihan, Dhiman Goswami, Antara Mahmud, Antonios Anastasopoulos, and Marcos Zampieri. 2024. Emomix-3l: A code-mixed dataset for bangla-english-hindi emotion detection. *LREC-COLING*, page 11.

Tharindu Ranasinghe, Kai North, Damith Premasiri, and Marcos Zampieri. 2022. Overview of the hasoc subtrack at fire 2022: Offensive language identification in marathi. *CEUR Workshop*.

Manikandan Ravikiran, Bharathi Raja Chakravarthi, Anand Kumar Madasamy, Sangeetha S, Ratnavel Rajalakshmi, Sajeetha Thavareesan, Rahul Ponnusamy, and Shankar Mahadevan. 2022a. Findings of the shared task on offensive span identification fromCode-mixed Tamil-English comments. In *Proceedings of DravidianLangTech*.

Manikandan Ravikiran, Bharathi Raja Chakravarthi, Anand Kumar Madasamy, Sangeetha Sivanesan, Ratnavel Rajalakshmi, Sajeetha Thavareesan, Rahul Ponnusamy, and Shankar Mahadevan. 2022b. Findings of the shared task on Offensive Span Identification in code-mixed Tamil-English comments. In *Proceedings of DravidianLangTech*.

Hafiz Muhammad Raza Ur Rehman, Mahpara Saleem, Muhammad Zeeshan Jhandir, Eduardo Silva Alvarado, Helena Garay, and Imran Ashraf. 2025. Detecting hate in diversity: a survey of multilingual code-mixed image and video analysis. *Journal of Big Data*, 12.

Thomas Reitmaier, Electra Wallington, Dani Kalarikalayil Raju, Ondrej Klejch, Jennifer Pearson, Matt Jones, Peter Bell, and Simon Robinson. 2022. Opportunities and challenges of automatic speech recognition systems for low-resource language speakers. In *Proceedings of CHI*.

Julian Risch, Anke Stoll, Lena Wilms, and Michael Wiegand. 2021. Overview of the GermEval 2021 shared task on the identification of toxic, engaging, and fact-claiming comments. In *Proceedings of GermEval*.

Hammad Rizwan, Muhammad Haroon Shakeel, and Asim Karim. 2020. Hate-speech and offensive language detection in roman urdu. In *Proceedings of EMNLP*.

Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Marcos Zampieri, and Preslav Nakov. 2021. Solid: A large-scale semi-supervised dataset for offensive language identification. In *Findings of ACL-IJCNLP*.

Pradeep Kumar Roy and Abhinav Kumar. 2025. Ensuring safety in digital spaces: Detecting code-mixed hate speech in social media posts. *Data & Knowledge Engineering*, 156:102409.

Caroline Sabty, Islam Mesabah, Özlem Çetinoğlu, and Slim Abdennadher. 2021. Language identification of intra-word code-switching for arabic–english. *Array*, 12.

Suleiman Saka and Sanchari Das. 2024. Evaluating privacy measures in healthcare apps predominantly used by older adults. *arXiv preprint arXiv:2410.14607*.

Suleiman Saka and Sanchari Das. 2025. Sok: Reviewing two decades of security, privacy, accessibility, and usability studies on internet of things for older adults. *arXiv preprint arXiv:2512.16394*.

Tanja Samardzic, Yves Scherrer, and Elvira Glaser. 2016. Archimob-a corpus of spoken swiss german. In *Proceedings of LREC*.

Ameni Sassi, Junior Tonga, Stéphanie Poaty, Sanon Steve, Djibrine Idriss Abakar Adjid, Moukhtar Cherif, and Wael Ouarda. 2024. Afridial: African dialect model based on deep learning for sentiment analysis. In *Proceedings of IWCMC*.

Shrey Satapara, Sandip Modha, Thomas Mandl, Hiren Madhu, and Prasenjit Majumder. 2021. Overview of the hasoc subtrack at fire 2021: Conversational hate speech detection in code-mixed language. In *FIRE (Working Notes)*.

Kasthuri Shanmugalingam and Sagara Sumathipala. 2019. Language identification at word level in sinhala-english code-mixed social media text. In *Proceedings of SCSE*.

Abhishek Sharma, Amrita, Sudeshna Chakraborty, and Shivam Kumar. 2022. Named entity recognition in natural language processing: A systematic review. In *Proceedings of DoSCI*.

Sanur Sharma and Anurag Jain. 2020. Role of sentiment analysis in social media security and analytics. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 2017. Membership inference attacks against machine learning models. In *Proceedings of IEEE S&P*. IEEE.

Sunny Shrestha and Sanchari Das. 2022. Exploring gender biases in ml and ai academic research through systematic literature review. *Frontiers in artificial intelligence*, 5:976838.

Sunny Shrestha, Esa Irby, Raghav Thapa, and Sanchari Das. 2022. Sok: A systematic literature review of bluetooth security threats and mitigation measures. In *International Symposium on Emerging Information Security and Applications*, pages 108–127. Springer.

Gudbjartur Ingi Sigurbergsson and Leon Derczynski. 2020. Offensive language and hate speech detection for Danish. In *Proceedings of LREC*.

Paulo Silva, Carolina Gonçalves, Nuno Antunes, Marilia Curado, and Bogdan Walek. 2022. Privacy risk assessment and privacy-preserving data monitoring. *Expert Systems with Applications*, 200:116867.

Shailendra Kumar Singh and KS Manoj. 2017. Importance and challenges of social media text. *International Journal of Advanced Research in Computer Science*, 8:831–834.

KP Soman. 2018. Overview of the second shared task on indian native language identification (inli). *CEUR Workshop*.

Jiayang Song, Yuheng Huang, Zhehua Zhou, and Lei Ma. 2025. Multilingual blending: Large language model safety alignment evaluation with language mixture. In *Findings of NAACL*.

Dama Sravani, Lalitha Kameswari, and Radhika Mamidi. 2021. Political discourse analysis: A case study of code mixing and code switching in political speeches. In *Proceedings of CALCS*.

Cristian-Alexandru Staicu, Sazzadur Rahaman, Ágnes Kiss, and Michael Backes. 2023. Bilingual problems: Studying the security risks incurred by native extensions in scripting languages. In *Proceedings of USENIX*.

Stian Steinbakken and Björn Gambäck. 2020. Native-language identification with attention. In *Proceedings of ICON*.

Jackson Stokes, Tal August, Robert A Marver, Alexei Czeskis, Franziska Roesner, Tadayoshi Kohno, and Katharina Reinecke. 2023. How language formality in security and privacy interfaces impacts intended compliance. In *Proceedings of CHI*.

Yolande Strengers, Lizhen Qu, Qiongkai Xu, and Jarrod Knibbe. 2020. Adhering, steering, and queering: Treatment of gender in natural language generation. In *Proceedings of CHI*.

Adam Sutton, Xi Bai, Kawsar Noor, Thomas Searle, and Richard Dobson. 2025. Named entity inference attacks on clinical llms: Exploring privacy risks and the impact of mitigation strategies. In *Proceedings of PrivateNLP*.

Camilla Arundie Tabe. 2023. Code-mixing and code-switching in cameroon social media. *International Journal of Linguistics and Translation Studies*, 4.

Kunsheng Tang, Wenbo Zhou, Jie Zhang, Aishan Liu, Gelei Deng, Shuai Li, Peigui Qi, Weiming Zhang, Tianwei Zhang, and Nenghai Yu. 2024. Gendercare: A comprehensive framework for assessing and reducing gender bias in large language models. In *Proceedings of CCS*.

Faiza Tazi, Archana Nandakumar, Josiah Dykstra, Prashanth Rajivan, and Sanchari Das. 2023. Sok: Analysis of user-centered studies focusing on healthcare privacy & security. *arXiv preprint arXiv:2306.06033*.

Faiza Tazi, Archana Nandakumar, Josiah Dykstra, Prashanth Rajivan, and Sanchari Das. 2024. Sok: Analyzing privacy and security of healthcare data from the user perspective. *ACM Transactions on Computing for Healthcare*, 5(2):1–31.

Faiza Tazi, Sunny Shrestha, Junibel De La Cruz, and Sanchari Das. 2022. Sok: An evaluation of the secure end user experience on the dark net through systematic literature review. *Journal of Cybersecurity and Privacy*, 2(2):329–357.

Daniela Teodorescu, Tiffany Cheng, Alona Fyshe, and Saif Mohammad. 2023. Language and mental health: Measures of emotion dynamics from text as linguistic biosocial markers. In *Proceedings of EMNLP*.

Joel Tetreault, Daniel Blanchard, and Aoife Cahill. 2013. A report on the first native language identification shared task. In *Proceedings of BEA*.

Douglas Trajano, Rafael H Bordini, and Renata Vieira. 2024. Olid-br: offensive language identification dataset for brazilian portuguese. *Language Resources and Evaluation*, 58:1263–1289.

Blase Ur and Yang Wang. 2013. A cross-cultural framework for protecting user privacy in online social media. In *Proceedings of WWW*.

Nishant Vishwamitra, Keyan Guo, Farhan Tajwar Romit, Isabelle Ondracek, Long Cheng, Ziming Zhao, and Hongxin Hu. 2024. Moderating new waves of online hate with chain-of-thought reasoning in large language models. In *Proceedings of IEEE S&P*.

Paul Voigt and Axel Von dem Bussche. 2017. The eu general data protection regulation (gdpr). *A Practical Guide, 1st Ed., Cham: Springer International Publishing*, 10:10–5555.

Fanfan Wang, Heqing Ma, Rui Xia, Jianfei Yu, and Erik Cambria. 2024. SemEval-2024 task 3: Multimodal emotion cause analysis in conversations. In *Proceedings of SemEval*.

Shuo Wang and Richard O Sinnott. 2017. Protecting personal trajectories of social media users through differential privacy. *Computers & Security*, 67:142–163.

Sze-Meng Jojo Wong and Mark Dras. 2009. Contrastive analysis and native language identification. In *Proceedings of ALTA*.

Liang Xiao, Fei-Peng Guo, and Qi-Bei Lu. 2018. Mobile personalized service recommender model based on sentiment analysis and privacy concern. *Mobile Information Systems*, 2018.

Yunze Xiao, Yujia Hu, Kenny Tsu Wei Choo, and Roy Ka-Wei Lee. 2024. ToxiCloakCN: Evaluating robustness of offensive language detection in Chinese with cloaking perturbations. In *Proceedings of EMNLP*.

Honghui Xu, Wei Li, Daniel Takabi, Daehee Seo, and Zhipeng Cai. 2025. Privacy-preserving multimodal sentiment analysis. *IEEE Internet of Things Journal*.

Wentao Xu. 2024. Characterization of political polarized users attacked by language toxicity on twitter. In *Proceedings of CSCW*.

Svetlana N Yanushkevich, Shawn C Eastwood, Martin Drahansky, and Vlad P Shmerko. 2018. Understanding and taxonomy of uncertainty in modeling, simulation, and risk profiling for border control automation. *The Journal of Defense Modeling and Simulation*, 15:95–109.

Zipeng Ye and Wenjian Luo. 2025. Llms are privacy erasable. In *Findings of EMNLP*.

Zheng Xin Yong, Ruochen Zhang, Jessica Forde, Skyler Wang, Arjun Subramonian, Holy Lovenia, Samuel Cahyawijaya, Genta Winata, Lintang Sutawika, Jan Christian Blaise Cruz, Yin Lin Tan, Long Phan, Long Phan, Rowena Garcia, Thamar Solorio, and Alham Fikri Aji. 2023. Prompting multilingual large language models to generate code-mixed texts: The case of south East Asian languages. In *Proceedings of CALCS*.

Lin Yue, Weitong Chen, Xue Li, Wanli Zuo, and Minghao Yin. 2019. A survey of sentiment analysis in social media. *Knowledge and information systems*, 60:617–663.

Reza Zafarani and Huan Liu. 2013. Connecting users across social media sites: a behavioral-modeling approach. In *Proceedings of SIGKDD*, pages 41–49.

Marcos Zampieri, Alina Maria Ciobanu, and Liviu P Dinu. 2017. Native language identification on text and speech. In *Proceedings of BEA*.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019a. Predicting the type and target of offensive posts in social media. In *Proceedings of NAACL*.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019b. SemEval-2019 task 6: Identifying and categorizing offensive language in social media (OffensEval). In *Proceedings of SemEval*.

Marcos Zampieri, Shervin Malmasi, Yves Scherrer, Tanja Samardžić, Francis Tyers, Miikka Silfverberg, Natalia Klyueva, Tung-Le Pan, Chu-Ren Huang, Radu Tudor Ionescu, Andrei M. Butnaru, and Tommi

Jauhiainen. 2019c. A report on the third VarDial evaluation campaign. In *Proceedings of VarDial*.

Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Çağrı Çöltekin. 2020. SemEval-2020 task 12: Multilingual offensive language identification in social media (OffensEval 2020). In *Proceedings of SemEval*.

Marcos Zampieri, Damith Premasiri, and Tharindu Ranasinghe. 2024. A federated learning approach to privacy preserving offensive language identification. In *Proceedings of LREC-COLING*.

Alisa Zezulak, Faiza Tazi, and Sanchari Das. 2023. Sok: Evaluating privacy and security concerns of using web services for the disabled population. In *7th Workshop on Technology and Consumer Protection (ConPro'23)*.

Jinxue Zhang, Jingchao Sun, Rui Zhang, Yanchao Zhang, and Xia Hu. 2018. Privacy-preserving social media data outsourcing. In *Proceedings of INFOCOM*. IEEE.

Ruochen Zhang, Samuel Cahyawijaya, Jan Christian Blaise Cruz, Genta Winata, and Alham Aji. 2023. Multilingual large language models are not (yet) code-switchers. In *Proceedings of EMNLP*.

Wei Zhang and Alexandre Salle. 2023. Native language identification with large language models. *arXiv preprint arXiv:2312.07819*.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28.

Marc A Zissman, Terry P Gleason, Deborah M Rekart, and Beth L Losiewicz. 1996. Automatic dialect identification of extemporaneous conversational, latin american spanish speech. In *Proceedings of ICASSP*.

8 Venue and Task-wise Paper Selection Details

Venue	Initial Count	Final Count
NDSS	16	1
CCS	68	2
USENIX	689	3
S&P	19	3
SOUPS	19	4
CSCW	43	5
ICWSM	71	7
CHI	165	8
EMNLP	1513	36
ACL	1272	41
Other Journals and Conferences	107	93
Total	**3982**	**203**

Table 5: Venue-wise # of the Publications Collected

Task	Paper Count
Sentiment Analysis	16
Emotion Detection	14
Offensive Language Identification	19
Code-Mixing	39
Native Language Identification	29
Dialect Identification	24
Privacy	62
Total	**203**

Table 6: Task-wise # of the Peer-reviewed Publications

9 Additional Survey Findings and Experimental Results

Table 7 shows the survey results of Datasets, Benchamark Competitions, and Computational Methodology in Social Media NLP Tasks.

Table 8 shows Accuracy, Precision and Recall across all the models.

10 Social Media Privacy Risk Analysis

Tables 9, 10, 11, 12, 13, and 14 show the detailed Analysis of Social Media Privacy Risks across different tasks.

11 Social Media Privacy Risk Mitigation Analysis

Table 15 shows the detailed analysis of mitigation strategies of privacy concerns in social media.

Tasks	Datasets	Competitions	Computational Methodology
Sentiment Analysis	SentMix-3L (Raihan et al., 2023a), Bambara-French (Konate and Du, 2018), Yelp (Zhang et al., 2015), Sentiment140 (Go et al., 2009)	SemEval (Muhammad et al., 2023; Barnes et al., 2022; Patwa et al., 2020), EvalITA (Barbieri et al., 2016), SIGHAN (Lee et al., 2024)	Feature Engineering, Deep Learning, Word2Vec (Barbieri et al., 2016), Dependency Graph (Muhammad et al., 2023), Transformer (Lee et al., 2024; Muhammad et al., 2023; Barnes et al., 2022; Patwa et al., 2020; Raihan et al., 2023a), LLM Prompting (Lee et al., 2024; Raihan et al., 2023a), LSTM (Konate and Du, 2018), CNN, SVM, Naive Bayes (Barbieri et al., 2016; Konate and Du, 2018), LAPT, TAPT, Ensembles, Sentence Transformer (Barnes et al., 2022)
Emotion Detection	EmoMix-3L (Raihan et al., 2024), MELD (Poria et al., 2019), ArtELingo (Mohamed et al., 2022), EmoReact (Nojavanasghari et al., 2016)	SemEval (Wang et al., 2024; Kumar et al., 2024), WASSA (Giorgi et al., 2024; Mohammad and Bravo-Marquez, 2017), IberLEF (Pan et al., 2024)	SVM, LSTM (Wang et al., 2024; Mohammad and Bravo-Marquez, 2017), LLM Fine-tuning (Kumar et al., 2024), Transformer (Kumar et al., 2024; Giorgi et al., 2024; Pan et al., 2024; Raihan et al., 2024), LLM Prompting (Wang et al., 2024; Giorgi et al., 2024; Raihan et al., 2024), CNN, Word2Vec (Mohammad and Bravo-Marquez, 2017), LoRA (Pan et al., 2024)
Offensive Language Identification	OffMix-3L (Goswami et al., 2023), COLD (Deng et al., 2022), RUHSOLD (Dewani et al., 2023), KOLD (Jeong et al., 2022), DK_{HATE} (Sigurbergsson and Derczynski, 2020), OGTD (Pitenis et al., 2020), OLID (Zampieri et al., 2019a), SOLID (Rosenthal et al., 2021), OLID-BR (Trajano et al., 2024), TB-OLID (Raihan et al., 2023b)	SemEval (Zampieri et al., 2019b, 2020), HASOC (Ghosh et al., 2023; Ranasinghe et al., 2022; Mandl et al., 2019, 2020), TRAC (Kumar et al., 2018, 2020), GermEval (Risch et al., 2021), DravidianLangTech (Ravikiran et al., 2022a; Chakravarthi et al., 2021; Ravikiran et al., 2022b; B et al., 2024)	Bi-LSTM (Zampieri et al., 2019b; Ghosh et al., 2023; B et al., 2024), Transformer (Goswami et al., 2023; B et al., 2024; Zampieri et al., 2019b, 2020; Mandl et al., 2019; Ranasinghe et al., 2022; Kumar et al., 2020; Ravikiran et al., 2022a; Chakravarthi et al., 2021), LLM Prompting (Goswami et al., 2023), LR (Ghosh et al., 2023; B et al., 2024; Chakravarthi et al., 2021), Naive Bayes (Ghosh et al., 2023), SVM (Zampieri et al., 2019b; Ghosh et al., 2023; Mandl et al., 2020; Kumar et al., 2018, 2020; B et al., 2024), n-gram (Mandl et al., 2020), KNN, Deep Learning (Ranasinghe et al., 2022), CNN (Kumar et al., 2018, 2020; Zampieri et al., 2019b; Ghosh et al., 2023; B et al., 2024), LSTM, TF-IDF (Chakravarthi et al., 2021; Kumar et al., 2018, 2020; Mandl et al., 2020), RNN (B et al., 2024)
Code-Mixing	English-Manipuri (Lamabam and Chakma, 2016), Indonesian-English (Nabila and Idayani, 2022), English-Kannada (Lakshmi and Shambhavi, 2017), English-Cameroon-French (Tabe, 2023), Sinhala-English (Shanmugalingam and Sumathipala, 2019), AR-EN CS LID Corpus (Sabty et al., 2021), Hinglish, Tenglish, Tamlish (Chandu et al., 2018)	ICON (Balouchzahi et al., 2022), FIRE (Chakravarthi et al., 2020b, 2023), CALCS (Aguilar et al., 2018; Sravani et al., 2021; Yong et al., 2023), HASOC (Modha et al., 2022; Satapara et al., 2021)	Transformer (Balouchzahi et al., 2022; Chakravarthi et al., 2020b, 2023; Modha et al., 2022; Satapara et al., 2021), TF-IDF (Balouchzahi et al., 2022; Satapara et al., 2021; Lakshmi and Shambhavi, 2017), CNN (Chakravarthi et al., 2020b; Satapara et al., 2021), SVM (Chakravarthi et al., 2023; Aguilar et al., 2018; Shanmugalingam and Sumathipala, 2019), KNN, Bi-LSTM (Chakravarthi et al., 2023), LSTM (Aguilar et al., 2018; Balouchzahi et al., 2022), LLM Prompting (Yong et al., 2023), n-gram (Lamabam and Chakma, 2016; Lakshmi and Shambhavi, 2017), BoW (Lakshmi and Shambhavi, 2017), LR, Decision Tree, Naive Bayes (Shanmugalingam and Sumathipala, 2019)
Native Language Identification	Reddit-L2 (Rabinovich et al., 2018), INLI 2013 (Anand Kumar et al., 2017), INLI 2017 (Soman, 2018)	NLI (Tetreault et al., 2013; Malmasi et al., 2017), INLI (Anand Kumar et al., 2017; Soman, 2018)	n-gram (Koppel et al., 2005), TF-IDF (Wong and Dras, 2009), Lexical and Syntactic Feature (Mechti et al., 2016; Gebre et al., 2013), SVM (Jarvis et al., 2013), LR (Di Nuovo et al., 2020), Ensemble and Meta Classifier (Zampieri et al., 2017), Transformer (Steinbakken and Gambäck, 2020), LLM (Zhang and Salle, 2023)
Dialect Identification	AfriDial (Sassi et al., 2024), MPCA (Malmasi et al., 2015), North African Dialect (Berrimi et al., 2020), TIMIT (Zissman et al., 1996), TwitterAAE (Blodgett et al., 2016), DIALECTBENCH (Faisal et al., 2024), ArchiMob (Samardzic et al., 2016), ITDI-FDI (Aepli et al., 2022)	VarDial (Malmasi et al., 2016a; Zampieri et al., 2019c; Gaman et al., 2020; Chifu et al., 2024), NADI (Abdul Mageed et al., 2024, 2023, 2022; Abdul-Mageed et al., 2021; Abdul Mageed et al., 2020)	n-gram, TF-IDF (Abdul Mageed et al., 2023, 2020; Abdul-Mageed et al., 2021; Abdul Mageed et al., 2022; Gaman et al., 2020), Ensemble (Abdul Mageed et al., 2023, 2020; Abdul-Mageed et al., 2021; Abdul Mageed et al., 2022; Malmasi et al., 2016a), LLM Prompting (Abdul Mageed et al., 2023, 2022; Chifu et al., 2024), Transformer (Abdul Mageed et al., 2024; Gaman et al., 2020; Chifu et al., 2024; Zampieri et al., 2019c), CNN, SVM (Gaman et al., 2020; Zampieri et al., 2019c; Malmasi et al., 2016a), LoRA (Chifu et al., 2024), Naive Bayes (Malmasi et al., 2016a; Zampieri et al., 2019c), LSTM, Meta Classifier, Bi-LSTM (Zampieri et al., 2019c)

Table 7: Survey Result of Datasets, Benchamark Competitions, and Computational Methodology in Social Media NLP Tasks

Tasks	Finetuning									Privacy-Preserving Finetuning								
	XLM-R			GPT-2			FLAN-T5			XLM-R			GPT-2			FLAN-T5		
	Acc	Prec	Rec	Acc	Prec	Rec	Acc	Prec	Rec	Acc	Prec	Rec	Acc	Prec	Rec	Acc	Prec	Rec
Sentiment Analysis	0.84	0.84	0.84	0.86	0.86	0.86	0.50	0.25	0.50	0.84	0.84	0.84	0.85	0.85	0.85	0.50	0.25	0.50
Emotion Detection	0.64	0.60	0.64	0.63	0.58	0.63	0.59	0.58	0.59	0.61	0.57	0.61	0.60	0.56	0.60	0.48	0.23	0.48
Offensive Language Identification	0.84	0.84	0.84	0.86	0.86	0.86	0.79	0.79	0.79	0.82	0.82	0.82	0.84	0.84	0.84	0.72	0.72	0.72
Code-Mixing	0.65	0.62	0.65	0.62	0.58	0.62	0.64	0.67	0.64	0.61	0.57	0.61	0.58	0.54	0.58	0.63	0.65	0.63
Native Language Identification	0.58	0.58	0.58	0.50	0.55	0.50	0.20	0.41	0.20	0.38	0.44	0.38	0.32	0.40	0.32	0.17	0.13	0.17
Dialect Identification	0.67	0.62	0.67	0.60	0.56	0.60	0.50	0.36	0.50	0.64	0.60	0.64	0.50	0.49	0.50	0.47	0.31	0.47

Table 8: Experimental Results for Privacy-Preserving NLP (Accuracy, Precision, Recall)

	Data Collection and Usage													
Reference	User Opinion	Demographic	Emotion State	Vulnerability Leakage	Behavioral Pattern	Toxic Content	Harassment Risk	Context Misinterpretation	Mixed Context	Identity Risk	L1 Marking	Region Marking	Speech Sensitivity	Ethnic Trace
Xiao et al. (2018) Sharma and Jain (2020) Lohar et al. (2021)	•	•	•	•	•	•	•	•	•	•		•	•	
Raihan et al. (2024) Teodorescu et al. (2023) Arango et al. (2024) Zampieri et al. (2024)	•	•	•	•	•	•	•	•	•	•	•	•	•	•
Sigurbergsson and Derczynski (2020) Rosenthal et al. (2021) Deng et al. (2022)	•	•		•	•	•	•	•	•	•	•	•	•	•
Hidayatullah et al. (2022)	•							•	•	•	•	•	•	
Staicu et al. (2023) Bierner (2001) Goldin et al. (2018) Jauhiainen et al. (2019)	•	•	•					•	•	•	•	•	•	
Jørgensen et al. (2015) Huang (2015) Fleisig et al. (2024) Barot et al. (2024)	•	•	•	•		•		•	•	•	•	•	•	•

Table 9: Analysis of Social Media Privacy Concerns (Data Collection and Usage). Each row of reference is associated to this concern of the corresponding tasks mentioned in Table 1.

	Data Preprocessing and Anonymization																
Reference	Name-Entity Leak	Quote Leakage	Stylistic Markers	Emotion Cues	Intensity Tag	Mood Triggers	Slang Trace	Sensitive Pairing	Script Leakage	Phonetic Spelling	Spelling Errors	Grammatical Patterns	Syntax Clues	Accent Trace	Word Uniqueness	Phrasing Style
Lohar et al. (2021) Xiao et al. (2018)		•	•	•	•	•	•	•		•	•	•	•		•	•
Teodorescu et al. (2023)	•	•	•	•	•	•	•	•	•		•	•	•	•	•	•
Leonardelli et al. (2021) Jeong et al. (2022) Sigurbergsson and Derczynski (2020) Rosenthal et al. (2021)	•	•	•	•	•		•	•			•	•			•	•
Hidayatullah et al. (2022)			•				•	•		•	•	•	•		•	•
Berzak et al. (2017)	•		•					•				•	•	•		•
Malmasi et al. (2015) Ferragne and Pellegrino (2007) Mahmud et al. (2023)	•	•	•	•	•	•	•	•		•	•	•	•	•	•	•

Table 10: Analysis of Social Media Privacy Concerns (Data Preprocessing and Anonymization). Each row of reference is associated to this concern of the corresponding tasks mentioned in Table 1.

	Data Visibility and User Profiling Risks																
Reference	Sentiment Pattern	Group Clustering	Profile Inference	Behavioral Modeling	Profile Tracking	Exposure Risk	Group Tagging	Group Inference	Speech Style	Cultural Assumption	Educational Bias	Forced Clustering	Region Profiling	Cultural Tagging	Group Labelling	L1 Trace
Lohar et al. (2021) Sharma and Jain (2020)	•	•	•	•	•	•	•	•	•				•		•	
Leonardelli et al. (2021) Rosenthal et al. (2021)	•	•	•	•		•	•	•	•	•		•	•	•	•	
Hidayatullah et al. (2022) Zhang et al. (2023)	•								•	•						•
Jauhiainen et al. (2019) Goldin et al. (2018) Aoyama and Schneider (2024)		•	•		•					•	•		•	•		•

Table 11: Analysis of Social Media Privacy Concerns (Data Preprocessing and Anonymization). Each row of reference is associated to this concern of the corresponding tasks mentioned in Table 1.

	Computational Risks in NLP										
Reference	Gradient Leakage	Model Inversion	Adversarial Attack	Memory Retention	Model Overfitting	LLM Data Leakage	Lossy Tokenization	Embedding Noise	Misclassification	L1 Hallucination	Dialect-Mixing
Arango et al. (2024) Xiao et al. (2024) Morabito et al. (2024)			•		•	•	•	•	•	•	
Mahmud et al. (2023)					•		•	•	•		•

Table 12: Analysis of Social Media Privacy Concerns (Computational Risks in NLP). Each row of reference is associated to this concern of the corresponding tasks mentioned in Table 1.

	Bias, Fairness, and Discrimination Risks																
Reference	Polarity Skewness	Cultural Bias	Language Bias	Labeling Bias	Gender Bias	Underrepresented Group	Minority Flag	Informality Bias	Code-switching Flag	Monolingual Bias	Label Noise	Context Loss	Majority L2 Focus	Underrepresented L1 Bias	Sociolect Bias	Wrong Annotation	Accent Bias
Konate and Du (2018) Barbieri et al. (2016)	•		•	•		•		•	•	•	•	•	•	•		•	
Fleisig et al. (2024)		•	•	•	•	•	•	•		•		•	•	•	•		•

Table 13: Analysis of Social Media Privacy Concerns (Bias, Fairness, and Discrimination Risks). Each row of reference is associated to this concern of the corresponding tasks mentioned in Table 1.

	Regulatory Compliance and Ethics														
Reference	Data Reuse	Purpose Mismatch	Right to be Forgotten	HIPAA	Longitudinal Use	Defamation Liability	Ethical Data Review	Data Consent	Speech Control	Transparency	Local Rights	Opt in-out policy	Multilingual Consent	GDPR	CCPA
Xiao et al. (2018) Raihan et al. (2023a)	•	•					•	•		•		•	•		
Mohamed et al. (2022) Raihan et al. (2024) Teodorescu et al. (2023)	•			•	•		•	•	•	•	•	•	•	•	•
Zampieri et al. (2024) Jeong et al. (2022)	•		•			•	•	•	•	•	•	•		•	
Hidayatullah et al. (2022) Garg et al. (2018)	•	•					•		•						
Jauhiainen et al. (2019) Nguyen et al. (2021)	•	•					•		•	•					
Huang (2015) Fleisig et al. (2024) Faisal et al. (2024) Barot et al. (2024)	•						•	•		•	•	•	•	•	•

Table 14: Analysis of Social Media Privacy Concerns (Regulatory Compliance and Ethics). Each row of reference is associated to this concern of the corresponding tasks mentioned in Table 1.

	Mitigation Strategy														
Reference	Differential Privacy	Federated Learning	SMPC	xAI	Homomorphic Encryption	K-Anonymity	L-Diversity	Name-Entity Masking	Data Aggregation	Data Regularization	Phonetic Normalization	Topic Filtering	SHAP	LIME
Identity Protection (Lohar et al., 2021; Teodorescu et al., 2023; Leonardelli et al., 2021; Hidayatullah et al., 2022; Berzak et al., 2017; Malmasi et al., 2015)				•				•	•			•		
Model Explainability (Xiao et al., 2018; Mohamed et al., 2022; Zampieri et al., 2024; Hidayatullah et al., 2022; Jauhiainen et al., 2019; Huang, 2015)	•	•						•	•	•		•		
Secured Computation (Arango et al., 2024; Xiao et al., 2024; Mahmud et al., 2023; Mehta and Passi, 2022; Gangarde et al., 2021; Alloghani et al., 2019)			•	•	•	•	•	•	•	•	•	•	•	•
Data Refinement (Jeong et al., 2022; Sigurbergsson and Derczynski, 2020; Berzak et al., 2017; Ferragne and Pellegrino, 2007)	•	•		•				•	•	•	•	•		

Table 15: Mitigation Strategies of Social Media Privacy Concerns