Information Research - Vol. 31 No. iConf (2026)

Privacy risk assessment method incorporating sensitivity and correlation with empirical study

Ruili Geng, Tiantian Zhang, Sentao Li, and Yishuai Xu*

DOI: https://doi.org/10.47989/ir31iConf64277

Abstract

Introduction. User-generated content (UGC) has emerged as a prominent vector for privacy breaches, especially due to the context-dependence of data sensitivity and vulnerabilities introduced by data correlations. These challenges highlight the growing limitations of traditional assessment methods.

Method. This study proposes a privacy risk quantification method integrating both attribute sensitivity and inter-attribute association, with an experimental validation conducted on the '*Friend Identification*' section of the https://muchong.com. A BERT-BiLSTM-CRF deep learning model is utilized for the automatic identification of attributes from unstructured text. Using a predefined privacy data lexicon, attribute sensitivity is quantified, and pointwise mutual information (PMI) is introduced to measure attribute associations. Combined with a privacy subject identification factor, these elements collectively quantify privacy risk values, followed by risk level classification.

Results. Ablation experiments and manual validation have confirmed the feasibility of the proposed scheme, demonstrating its capability to identify, assess, and classify privacy risks in unstructured textual data with broad applicability.

Conclusion. The study validates the proposed solution theoretically, technically, and empirically, overcoming the limitations of traditional isolated-field evaluation paradigms. The method can be extended to high-sensitivity domains such as healthcare and finance, providing a basis for dynamic, risk-informed classification policies.

Introduction

Online communities serve as important venues for information exchange, dissemination, and sharing. The behavioral patterns of users in these communities are influenced by both environmental and individual factors, and they generally exhibit a greater willingness to disclose personal information. During community interactions, users share information in various forms—such as images, locations, videos, and comments—that often contain sensitive or personally identifiable details. Although users with strong privacy awareness may attempt to reduce information disclosure by concealing personal identifiers or promptly deleting public posts, they still face potential leakage risks, such as associative identification from textual content and inference of sensitive information. Therefore, scientifically quantifying and assessing privacy risks is crucial for mitigating the privacy security risks faced by online community users and optimizing platform privacy protection mechanisms. In user-generated content, unstructured data has become a high-risk vector for privacy leakage. Particularly, the contextual variability of data sensitivity and vulnerabilities in data associativity pose challenges to traditional assessment methods based on structured data or the probability of single-field leakage.

To assess the privacy risks of UGC in online communities, this study proposes a privacy risk quantification and assessment method that integrates sensitivity and association. An experiment was conducted using the '*Friend Identification*' section of https://muchong.com as a case study. Based on the characteristics of textual data and contextual topics, a privacy data lexicon was constructed. The BERT-BiLSTM-CRF deep learning model was employed to identify sensitive data fields from unstructured text, quantify attribute sensitivity, and measure attribute associations using pointwise mutual information (PMI). By incorporating both sensitivity and association, the privacy risk value of each text was calculated, and risk levels were determined according to the types of fields involved. Ablation experiments and manual validation confirmed the feasibility of the proposed method. The findings offer new insights for improving privacy protection policies and enhancing platform privacy governance.

Background literature

Privacy risk refers to various threats faced by personal private information, reflecting the deprivation of individuals' control over their own data (Featherman et al.,2010). Privacy is highly context-dependent, and privacy risks and governance strategies vary across different contexts such as healthcare and finance (Shostack, 2014; Gbongli et al.,2020). Due to the complexity of online community environments, users face the risk of privacy leakage when posting content.

The virtual environment of online communities stimulates users' social needs (Xu et al.,2024). Through information sharing and interaction, users gain emotional support and social recognition, which leads to self-disclosure behaviors and promotes sustained engagement (Posey, et al., 2010). However, excessive use of social media can lead to more relaxed privacy attitudes among users (Tsay-Vogel et al., 2018). Even after large-scale privacy breaches, users' willingness to disclose private information may remain unaffected.

To prevent misuse of user-disclosed content, platforms provide clear privacy notices and obtain consent from users regarding data processing activities. Nevertheless, posts in online communities are often publicly accessible, resulting in blurred privacy boundaries and making it difficult for users to maintain effective control over their personal information. Additionally, discrepancies may exist between some platforms' stated privacy policies and their actual practices, leaving users' personal information inadequately protected (Dym et al., 2020).

Therefore, identifying and scientifically evaluating the privacy risks of UGC in online communities can help users understand the potential privacy risks associated with their self-disclosed content. It can also provide a basis for platforms to improve compliance governance strategies for user information.

Privacy risk assessment method incorporating sensitivity and correlation

Assessment method framework

Text content published by users in online communities is collected and undergoes data cleaning to form a basic corpus. A privacy-related lexicon is constructed based on the research context, and manual annotation is performed following the BIO tagging scheme. Deep learning methods are applied to extract sensitive fields from the text. The sensitivity and association of the text fields are then calculated, and their privacy risk values are quantified using a privacy risk measurement algorithm. Finally, risk classification rules are designed to complete the quantitative assessment and grading of privacy risks. The process is illustrated in Figure 1.

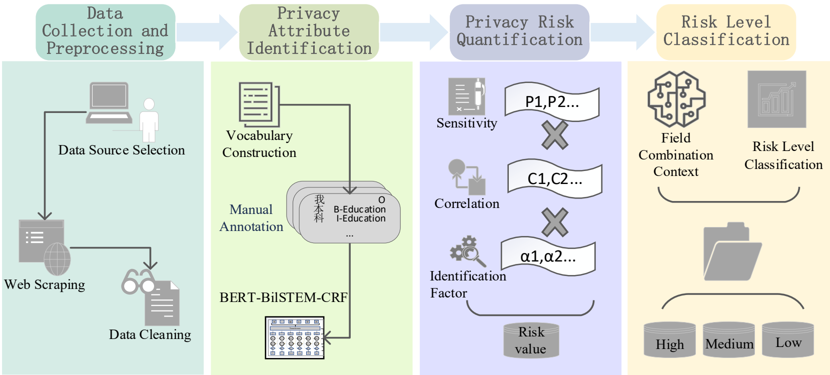

Figure 1. Framework diagram.

Data collection

During the data collection phase, the focus is on target online communities, where both user-published post content and actively disclosed account information (name, age, gender) are retrieved simultaneously.

The preprocessing stage involves multi-step cleaning procedures, including the removal of irrelevant characters, filtering of off-topic text, standardization of text formats, elimination of invalid posts with low character counts, and deletion of duplicate content.

Attribute identification

Using a predefined privacy data lexicon, unstructured text is converted into a structured format. Core attribute categories—such as identifiers, financial, and biometric data—are first derived through data induction. The cleaned data is then split into training, validation, and test sets, and manually annotated using the BIO scheme. A BERT-BiLSTM-CRF model is trained and evaluated in Python using precision, recall, and F1-score (Lample et al., 2016; Devlin et al., 2019). The best-performing model is applied to label non-sample data, producing the final annotated dataset.

Privacy risk quantification

According to risk management theory, the negative outcome of risk is typically measured by the product of the adverse impact caused by a risk event and its probability of occurrence (Bedford & Cooke, 2001). This concept has been applied in contexts such as logistics risk quantification (Nguyen & Wang, 2018). In online virtual communities, where users actively disclose their post content and platforms treat such information as '*already disclosed*,' the probability of leakage in this context refers to the likelihood of identifying the real user's identity through the correlation of sensitive fields. Therefore, this study defines privacy risk in online communities as:

$$Privacy\ risk = disclosure\ loss \times subject\ identifiability\ probability \tag{1}$$

Based on sensitive fields, we construct a privacy text matrix P. Assuming a privacy lexicon $W = \{w_1, w_2, w_3, \cdots, w_j\}$, the privacy text matrix for each textual instance in the dataset is represented as:

$$\mathrm{P} = \begin{bmatrix} w_{1,1} & 0 & \cdots & 0 \\ 0 & w_{2,2} & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & w_{i,j} \end{bmatrix} \tag{2}$$

This study employs PMI to measure attribute correlations, let $M = \{m_1, m_2, m_3, \cdots, m_n\}$denote the complete set of attribute correlations (Church & Hanks,1990). For any two attribute fields $m_p, m_n \in M$, their relationship is expressed in Formula (3):

$$C_{p,n} = PMI(m_p, m_n|K) = log\,\frac{p(m_p, m_n|K)}{p(m_p|K)p(m_n|K)} \tag{3}$$

Compute field correlation coefficients using PMI, where $C_i$denotes the correlation coefficient of field m_i,and ϑ is the filter threshold. Each C_i is derived by taking the $C = \{c_1, c_2, c_3, \cdots, c_j\}$.

$$C_i = \begin{cases} 0 & C_{i,j} < \vartheta \\ C_{i,j} & C_{i,j} \geq \vartheta \end{cases} \tag{4}$$

Let α denote the probability of risk leakage. We mapped the field combinations by referencing the definition of personal data in the GDPR. Assuming the existence of a field combination mapping relationship, an identification factor α is designed to assign values to categorized combinations of data field types, as shown in Formula (5).

$$F(s) = \begin{cases} 2.00, & \text{if } s = \text{‘Identified + semi − identified + sensitive data’} \\ 1.75, & \text{if } s = \text{‘Identified + semi − identified data’} \\ 1.50, & \text{if } s = \text{‘Identified + sensitive data’} \\ 1.25, & \text{if } s = \text{‘Semi − identified + sensitive data’} \\ 1.00, & \text{if } s = \text{‘Identified OR semi − identified’} \\ 0.00, & \text{if } s = \text{‘Sensitive data’} \end{cases}$$

$$\alpha = \mathrm{F}(\mathrm{s_i}), \mathrm{s_i} \in \mathrm{s} \tag{5}$$

Privacy risk level classification

A text with a privacy risk value below the mean is classified as low risk, while one above the mean is considered medium risk. However, since the combination of multiple fields may increase privacy disclosure risk, the risk classification method requires refinement: if a low-risk text involves identified data, it is directly categorized as high risk; if the text only contains semi-identified data and its risk value exceeds the mean, it is classified as high risk—otherwise, it remains medium risk.

Method validation

Data acquisition and preprocessing

We collected an initial sample of 16,425 text entries from https://muchong.com, covering the period from October 2019 to February 2025. After data cleaning, 14,608 valid entries remained.

Textual attribute recognition

(1) Attribute Annotation

To ensure annotation accuracy and consistency, the BIO tagging schema was rigorously implemented. In the sentence ‘*Shanghai Xinzhuang, born in 1992, male, Master's degree, works at a central enterprise, enjoys cycling and sports*’, the following entities are labelled: ‘*Shanghai Xinzhuang*’ is marked as ‘*Address*’, ‘1992’ as ‘*Birthday*’, ‘*male*’ as ‘*Gender*’, ‘*Master's degree*’ as ‘*Degree*’, ‘*central enterprise*’ as ‘*Nature*’, and ‘*cycling*’ and ‘*sports*’ as ‘*Hobby*’.

A randomly selected set of 1,000 samples was manually annotated using the BIO tagging scheme. Within the study's dataset, 32 sensitive attribute fields were identified and categorized into identified, semi-identified, and sensitive data, forming the privacy data lexicon (Table 1). Based on this lexicon, an additional 3,000 entries were annotated for machine learning model training.

Data type	Field
Identified data	Name、Telephone、Email、Wechat、QQ
Semi-identified data	Age、Birthday、Gender、Weight、Ethnicity、Height、Address
Sensitive data	Childbirth、Marriage、Occupation、Education、Degree、Field、Physical、Nature、Company、Hobby、Disease、Genetic、Lifestyle、Smoking、Alcohol、Certificate、Family、Loan、Income、Major

Table 1. Attribute field type categorization.

(2) Model Training

From the dataset of 4,000 entries, 3,200 texts were randomly selected as the training set, 400 as the validation set, and another 400 as the experimental test set.

We set Epochs = 30 and monitored changes in F1, precision, and recall. As shown in Figure 2, the model reached optimal and stable performance at Epoch = 10. The model from this epoch was selected to automatically identify privacy attribute entities in the remaining text data.

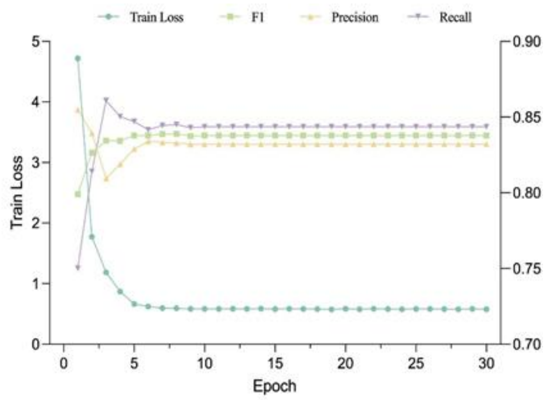

Figure 2. Results of 30 epochs of model training.

F1, R, and P metrics collectively reveal the model’s performance characteristics across distinct entity categories (see Table 2).

Category	Precision	Recall	F1-score
Address	0.8986	0.7649	0.8264
Education	0.7500	0.8049	0.7765
Gender	0.8889	0.6729	0.7660
Hobby	0.8122	0.8212	0.8167
Weight	0.9136	0.9250	0.9193
Birthday	0.7892	0.8421	0.8148
Degree	0.9524	0.9607	0.9565
Wechat	0.7111	0.7619	0.7356
Height	0.9073	0.9538	0.9300
Marriage	0.9104	0.9531	0.9313
Smoking	0.8947	1.0000	0.9444
QQ	0.5714	0.5455	0.5581
Major	0.7008	0.6364	0.6667
Fields	0.5000	1.0000	0.6667
Occupation	0.7966	0.7966	0.7966
Lifestyle	0.7143	0.7143	0.7143
Alcohol	0.9286	1.0000	0.9630
Family	0.7963	0.8269	0.8113
Email	0.6667	0.4000	0.5000
Physical	0.8667	0.6842	0.7647
Nature	0.8542	0.9318	0.8913
Income	0.3000	0.3750	0.3333
Loan	0.8333	0.8333	0.8333
Genetic	0.0000	0.0000	0.0000
Childbirth	0.9091	0.7692	0.8333
Company	0.4091	0.6429	0.5000
Age	0.8333	0.8824	0.8571
Telephone	0.0000	0.0000	0.0000
Name	0.0000	0.0000	0.0000
Disease	0.0000	0.0000	0.0000
Certificate	0.0000	0.0000	0.0000
Ethnicity	0.0000	0.0000	0.0000
micro avg	0.8480	0.8363	0.8421
macro avg	0.8519	0.8363	0.8411

Table 2. Results of the BERT-BiLSTM-CRF model in different entity categories.

Privacy attribute matrix

The privacy value of a single field in an individual record is converted into a vector matrix, resulting in a privacy attribute sensitivity matrix, as shown in Table 3.

	w_1	w_2	w_3	…	w_{31}	w_{32}
S_1	0.000	0.000	0.000	…	0.000	0.000
S_{108}	0.589	0.000	0.000	…	0.000	0.000
S_{5083}	0.000	0.569	0.600	…	0.215	0.000
S_{9402}	0.000	0.100	0.000	…	0.000	0.000
S_{13065}	0.000	0.000	0.000	…	0.000	1.700
S_{14608}	0.000	0.569	0.000	…	0.000	0.000

Table 3. Attribute sensitive matrix.

Attribute correlation identification

PMI was utilized to measure attribute associations across all 14,608 annotated text entries, generating an attribute association adjacency matrix, as presented in Table 4.

	Age	Birthday	Gender	...	Name
Age	0.000	0.000	0.146	...	0.204
Birthday	0.000	0.000	0.272	...	0.175
Gender	0.146	0.272	0.000	...	0.353
...	...	...	...	...	...
Name	0.204	0.175	0.353	...	0.000

Table 4. Attribute association adjacency matrix (partial).

Based on the adjacency matrix and Formula (4), the correlation coefficients of the 32 attributes were computed (Table 5). Genetic and Disease showed the strongest correlation, suggesting that their mention often coincides with disclosure of other information, indicating higher privacy risk. Attributes like Name, Certificate, and Income also had high correlations, reflecting users' frequent and active disclosure of these personal fields.

Here, $c_i$denotes the correlation coefficient of field m_i,and ϑ represents the filter threshold. The correlation coefficient is derived by taking the maximum association strength value for field m_i, yielding the correlation coefficient array $C = \{c_1, c_2, c_3, \cdots, c_j\}$.

Category	C	Category	C	Category	C	Category	C
Age	0.325	Birthday	0.353	Gender	0.353	Ethnicity	0.394
Education	0.509	Degree	0.384	Fields	0.517	Major	0.566
Height	0.318	Weight	0.331	Physical	0.701	Occupation	0.357
Family	0.503	Nature	0.408	Hobby	0.397	Telephone	0.413
Wechat	0.456	QQ	0.456	Email	0.582	Company	0.380
Address	0.336	Marriage	0.515	Disease	0.982	Childbirth	0.529
Genetic	0.982	Lifestyle	0.629	Smoking	0.642	Alcohol	0.675
Loan	0.558	Certificate	0.876	Income	0.465	Name	0.876

Table 5. Field attribute association coefficient.

Privacy risk assessment

Attribute sensitivity refers to data involving personal or privacy-related information, such as gender, age, religion, etc. Attribute association, on the other hand, describes whether these attributes are interconnected in some way or how they are contextually linked in UGC.

In this study, the correlation between attributes is determined using correlation coefficients. A higher correlation coefficient between one attribute and another indicates a stronger inter-attribute association.

Analysis reveals a strong positive correlation (Pearson r = 0.977) between overall data sensitivity and attribute association. Sensitivity values are relatively dispersed, while association values are more concentrated. As shown in Figure 3, this demonstrates an intrinsic link between data field sensitivity and attribute characteristics.

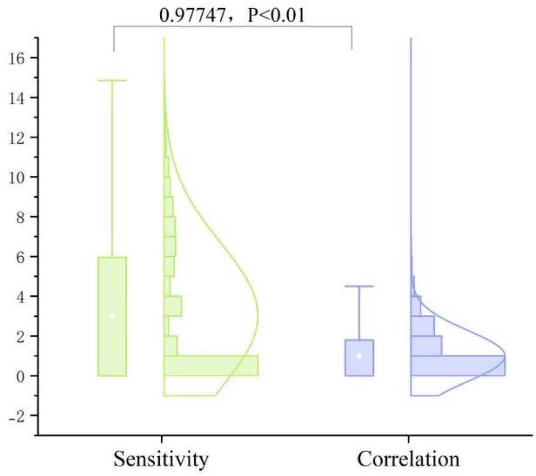

Figure 3. Sensitivity-Correlation relationship diagram.

Risk values were computed according to the established privacy risk assessment scheme, with each text classified into predefined risk levels. Results show 6,903 high-risk texts (47.25%), 457 medium-risk (3.13%), and 7,248 low-risk items (49.62%). The distribution of risk levels across the dataset is presented in Figure 4, where horizontal lines and dots indicate medians and 95% confidence intervals per risk level, respectively.

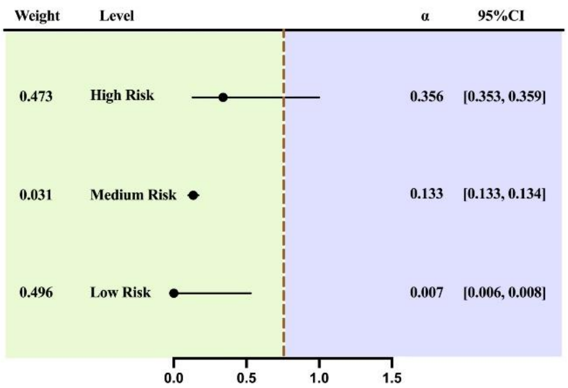

Figure 4. Distribution of risk levels

Evaluation effectiveness validation

(1) Ablation experiment

This study employs Formula (6) to conduct ablation experiments, further examining the impact of four modules—privacy sensitivity(S), attribute association(C), privacy breach loss(P), and the probability of privacy disclosure(L)—on the identification results.

$$F/F` = \{ x \mid x \in F \land x \notin F` \} \quad (6)$$

Here, F represents the complete set of modules, and F/F′ denotes the set of remaining modules after removing a specific module. The results of the ablation experiments are shown in Table 6.

	F/F`S	F/F`C	F/F`P	F/F`L
Precision	0.876	0.826	0.654	0.825
Recall	0.834	0.813	0.698	0.817
F1	0.854	0.82	0.665	0.821

Table 6. Ablation experiment results for each module.

Results show that P most significantly influences risk assessment outcomes. Although sensitivity and association are sub-dimensions of privacy value, ablating either has limited impact, as risks remain quantifiable. The privacy breach loss module sustains identification accuracy in virtual academic communities, confirming the feasibility of this method.

(2) Reliability analysis

This study performed a manual consistency check using the following criterion due to the limited number of medium-risk texts: if the text involved personal privacy leakage, it was classified as high risk; otherwise, as low risk.

Two privacy governance researchers independently evaluated 500 randomly selected texts. The Dice coefficient between manual labels and model results exceeded 0.9, with 472 texts consistent, confirming the scheme's reliability (ZHOU et al., 2024).

Discussion

Privacy leakage loss manifests as the synergistic interplay between sensitivity and correlation

This study reveals that in user-generated content within online communities, privacy breach loss is not determined solely by the sensitivity of isolated fields, but rather results from the synergistic effect of sensitivity and association. Focusing solely on the sensitivity analysis of individual fields—such as the inherent sensitivity of basic information like address, education, or income—can identify elementary threats but fails to capture the privacy risks arising from combinations of multiple fields.

When highly sensitive fields (e.g., ID numbers) coexist with low sensitivity yet strongly associated fields (e.g., educational background or employer), the resulting loss to the privacy subject increases. This is corroborated by correlation analysis, which shows a significant positive relationship between sensitivity and association: when high-sensitivity information is exposed, users tend to simultaneously disclose multiple associated fields across dimensions. When highly sensitive fields appear, users often provide more personal information (e.g., age, gender, and educational background) for social needs, which can lead to privacy leaks unintentionally.

Reconceptualizing the essence of leakage probability

For privacy management, the reconceptualization of the probability of leakage in this study holds significant academic value. This method redefines the core meaning of privacy leakage probability, transforming it into the likelihood of identifying the privacy subject. If data cannot be linked to a specific individual, its disclosure does not constitute a substantive privacy risk. Only when data can be associated with a specific natural person does leakage pose an actual risk.

In the context of UGC in online virtual communities, the concept of leakage probability shifts from '*whether the data will be leaked*' to '*whether the disclosed information can be linked to an individual's identity*.' The theoretical foundation of this shift lies in revealing the generative mechanism of privacy risk: risk arises from the ability of combinations of fields to identify the subject.

Conclusions

This study introduces a privacy risk assessment method for unstructured online community text, integrating attribute sensitivity and association via a BERT-BiLSTM-CRF recognition model and PMI. The approach enables quantitative risk evaluation and classification. Experiments show consistent performance in ablation tests and manual validation, confirming real-world applicability. Our study also has certain limitations: (1) Due to the scarcity of samples in some fields and the diversity of expression forms, the recognition performance for certain attributes was poor. (2) We did not further explore whether privacy protection awareness differs across different user groups. Future work will employ data augmentation and repeated sampling to further investigate variations in UGC privacy risks based on users' disciplines, educational backgrounds, and ages.

Acknowledgements

This paper is supported by the General Project of National Social Science Foundation of China (Project No. 22BTQ074) and the Innovative Talent Support Program for Philosophy and Social Sciences in Higher Education Institutions of Henan Province, China (NO. 2025-CXRC-24).

About the authors

Ruili Geng is an Associate Professor in the School of Information Management, ZhengZhou University, ZhengZhou, China. She received their PhD from Peking University in China. Her research interests include information behaviour and privacy management. She can be contacted at gengruili@zzu.edu.cn

Tiantian Zhang is a master student in the School of Information Management, ZhengZhou University, ZhengZhou, China. Her research interest is privacy management. She can be contacted at ttianzhang06@163.com

Sentao Li is a PhD student in the School of Information Management, ZhengZhou University, ZhengZhou, China. His research interests include information behaviour and privacy management. He is the corresponding author and can be contacted at lisentao99@gs.zzu.edu.cn

Yishuai Xu is a PhD student in the Department of Library and Information Science, Universiti Malaya, Kuala Lumpur, Malaysia. His research interests include machine learning, information retrieval and personalized recommendation. He can be contacted at yishuai@um.edu.my

References

Bedford, T., & Cooke, R. (2001). Probabilistic risk analysis: foundations and methods. Cambridge University Press.

Church, K., & Hanks, P. (1990). Word association norms, mutual information, and lexicography. Computational linguistics, 16(1), 22-29. https://dl.acm.org/doi/10.3115/981623.981633

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of NAACL-HLT 2019 (pp. 4171-4186), Minneapolis, USA, June 2-7, 2019. https://doi.org/10.18653/v1/N19-1423

Dym, B., & Fiesler, C. (2020). Social norm vulnerability and its consequences for privacy and safety in an online community. Proceedings of the ACM on Human-Computer Interaction, 4(CSCW2), 1-24. https://doi.org/10.1145/3415226

Featherman, M. S., Miyazaki, A. D., & Sprott, D. E. (2010). Reducing online privacy risk to facilitate e‐service adoption: the influence of perceived ease of use and corporate credibility. Journal of services marketing, 24(3), 219-229. https://doi.org/10.1108/08876041011040622

Gbongli, K., Xu, Y., Amedjonekou, K. M., & Kovács, L. (2020). Evaluation and Classification of Mobile Financial Services Sustainability Using Structural Equation Modeling and Multiple Criteria Decision-Making Methods. Sustainability, 12(4), 1288. https://doi.org/10.3390/su12041288

Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., & Dyer, C. (2016). Neural architectures for named entity recognition. arXiv preprint arXiv:1603.01360.

https://doi.org/10.48550/arXiv.1603.01360

Nguyen, S., & Wang, H. Y. (2018). Prioritizing operational risks in container shipping systems by using cognitive assessment technique. Maritime Business Review, 3 (2), 185-206. https://doi.org/10.1108/MABR-11-2017-0029

Posey, C., Lowry, P. B., Roberts, T. L., & Ellis, T. S. (2010). Proposing the online community self-disclosure model: the case of working professionals in France and the UK who use online communities. European journal of information systems, 19(2), 181-195. https://doi.org/10.1057/ejis.2010.15

Shostack, A. (2014). Threat modeling: Designing for security. John wiley & sons.

Tsay-Vogel, M., Shanahan, J., & Signorielli, N. (2018). Social media cultivating perceptions of privacy: A 5-year analysis of privacy attitudes and self-disclosure behaviors among Facebook users. New media & society, 20(1), 141-161. https://doi.org/10.1177/1461444816660731

Xu, X., Liu, J., & Liu, J. H. (2024). The effect of social media environments on online emotional disclosure: tie strength, network size and self-reference. Online Information Review, 48(2), 390-408. https://doi.org/10.1108/OIR-04-2022-0245

Zhou, L., Schellaert, W., Martínez-Plumed, F., Moros-Daval, Y., Ferri, C., & Hernández-Orallo, J. (2024). Larger and more instructable language models become less reliable. Nature, 634(8032), 61-68. https://doi.org/10.1038/s41586-024-07930-y