Systematizing the State of Knowledge in Detecting Privacy Sensitive Information in Unstructured Texts using Machine Learning

Sascha Löbner*, Welderufael B. Tesfay†
Chair of Mobile Business & Multilateral Security
Goethe University Frankfurt
Frankfurt am Main, Germany
ORCID: *0000-0001-9164-1919, †0000-0002-1087-2019

Vanessa Bracamonte‡, Toru Nakamura§
Usable Trust Group
KDDI Research, Inc.
Saitama, Japan
ORCID: ‡0009-0007-3844-5856 , §0000-0003-2530-648X

***Abstract*—Today, vast amounts of private and sensitive data are being shared across a variety of on-line services day-to-day. Recent technologies increasingly simplify the collection, processing and evaluation of these data. This results in numerous threats to the privacy of users. Although there are legal regulations to protect privacy, users are increasingly faced with the challenge of controlling their data to exercise their rights. In order to implement the existing legal framework and to help users protect their privacy, technological solutions are becoming increasingly important. Within the scope of this paper, the research areas of privacy risk detection will be examined in more detail. For this purpose, the state of the art of privacy sensitive information detection is elaborated and then analyzed by means of a specifically developed classification scheme to identify research gaps and trends. As a result, several research gaps and trends have been identified, demonstrating that further research is required to develop user tailored privacy enhancing tools and ensure adequate privacy protection.**

***Index Terms*—Privacy sensitive information, machine learning, privacy, PSI detection.**

I. Introduction

With the advent of machine learning and natural language processing into societys daily activities, the internet continues its massive expansion. Additionally, internet connected devices and services that rely on text processing have become more intertwined with everyday tasks and procedures. While the proliferation of such devices and services is useful to facilitate daily activities, it also presents an unprecedented challenge for user information privacy violation. Common examples for such violations regarding text data are identity theft, bodily and psychological harms and reputation damage of target users. In particular, the unintended access to highly Privacy Sensitive Information (PSI) that users would not want to consciously disclose exacerbates the consequences. But the past has shown that users are often not aware of sharing PSI or do not understand the implications [1].

While text data that is not shared with the public can be technically protected [2] from third parties, with free available text data, e.g., comments on movies, recommendations of products or posts in social media a tremendous amount of data to train PSI detection models exists. Researchers have been proposing various approaches to detect and proactively nudge users regarding the unintended disclosure of PSIs, e.g. by introducing tools that alert the user if personal information is detected and propose privacy friendly transformations [3]. Also, the European Union Data protection Regulation (GDPR) stipulates that the ”processing of personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, and the processing of genetic data, biometric data for the purpose of uniquely identifying a natural person, data concerning health or data concerning a natural persons sex life or sexual orientation shall be prohibited” [4] unless exceptionally required by law and consented by the user. However, this does not stop companies from analyzing unintentionally shared user data. This might be addressed by the AI Act, a proposed European law [5].

To comply with this regulation and increase the accuracy of PSI detection, it is crucial to understand the capabilities of current PSI detection models for text data. Such approaches utilize traditional and emerging technologies in natural language processing and machine learning to identify PSI in different domains from text data.

To the best of our knowledge, the existing research on PSI detection for text data has not been systematically studied. Thus, in this paper, we analyze the literature in the PSI detection from text data, show the capacities of the current models for different PSI types and identify future research works. Furthermore, by cataloguing research outputs of the last decade, we identify research trends and challenges.

II. Methodology

In this section, we explain the methodology we have used to collect the papers and develop the assessment framework.

A. Data Collection

Data collection plays a crucial role in identifying relevant studies and obtaining the necessary information for our analysis. Our process of data collection involves searching through academic databases and search engines to identify relevant peer-reviewed articles, and conference proceedings.

We relied on ACM Digital Library[1], IEEE Xplore[2] and Google Scholar[3]. The papers summarized in this literature synopsis were collected using pertinent key words. A combination of the following search keywords was used: "privacy sensitive information", "personal information", "personal data", "(unstructured) text" and "detection". The keywords "phishing", "malware" and "spam" were used to exclude works that focused on those topics. After combining the results from ACM Digital Library, IEEE Xplore and Google Scholar, we removed duplicates and papers which were not relevant.

B. Scope

The study presented in this paper focuses on approaches that detect the disclosure of privacy sensitive information (PSI) in textual data. We mainly focus on automated machine learning-based approaches. Moreover, the primary focus of our study are social network platforms since these platforms create situation where users can inadvertently divulge PSI which they might regret later. Therefore, the papers considered in this study present various machine learning or natural language processing techniques to detect the disclosure of PSI. Oftentimes, the tools also provide alerting mechanisms to nudge users to reconsider sharing PSI to a wider audience.

Time-wise, we delimited the collection of scientific works published between 2010 and 2022. Based on preliminary assessment, we observed that there was no significant research in this domain prior to 2010. Hence, we delimited the publication time of the papers from 2010 to 2022.

C. Limitations

We would like to point out that the search key words, the publication venues where the papers are collected from, as well as the delimitation in years of publication might have impact on the papers that hit our search. Furthermore, we faced accessibility issues with some publication portals such as Scopus. Another limitation is that our work mainly focuses on PSI detection tools for textual data written in English language only. We plan to address research on non-English PSI detection such as Turkish [6], Spanish [7], etc. in our future work.

D. Systematization Criteria

In this subsection, we elaborate on our systematization methodology that we used to develop the evaluation criteria. In general, the evaluation criteria can be split into three categories: (1) Machine learning related; (2) PSI types and (3) Data related. The categories are explained in the following:

1) Machine learning related.: Using this evaluation criteria, we looked at the papers in terms of the type of machine learning they use, the performance of the machine learning algorithms and the validation process.

[1]https://dl.acm.org/

[2]https://ieeexplore.ieee.org/Xplore/home.jsp

[3]https://scholar.google.com/

a) Machine learning in use.: A variety of different ML techniques exist that are used for different tasks in text data processing. The most common tasks are clustering, classification, machine translation and transcription tasks [8], [9]. Analyzing the different ML technologies is important for the understanding the capabilities of PSI detection methods.

b) Performance.: How to evaluate the performance of a model is not a trivial task. While accuracy is one of the most common scores reported, it is likely to hide biases and misclassification, especially for unbalanced data. Thus further evaluation metrics, e.g., F1-Score, precision and recall are taken into account [10]. Accuracy is highly important to this research because it shows how well certain privacy sensitive attributes can be elicited from the text data.

c) Cross Validation.: A k-fold cross validation is used to avoid a lucky split that can obfuscate difficult instances by placing them into the training data, resulting in a higher accuracy score that cannot be achieved after model-deployment [10]. The data is divided into k equal-sized folds, thus k experiments are required. Often, only the mean score from all k-folds is provided [10]. If cross validation was used, the reported performance scores are expected to be robust and reproducible. Although we investigated for other validation methods, we did not find any, thus why we narrowed down the validation method to cross validation.

2) PSI types: PSI types are a major category and an evaluation criteria in itself. We analyse which PSI types are detected in the different models and use the mapping of Chua et al. [11] to further group the PSI types found into six different PSI categories. PSI categories are highly interesting to investigate which PSI types are mostly analysed.

3) Data related: In this subsection, we summarize all criteria that refer to the data set.

a) Data source.: Where data was collected from is important when comparing different PSI studies. On the one hand, if always the same data is used, this increases comparability between the models, on the other hand, if always the same data is used, this is likely to hide biases.

b) Data set availability.: Most of the approaches in this literature review use large amounts of data to train their models. In light of encouraging open research, we considered the public availability of data sets as classification criteria.

c) Size of data.: Size of data is relevant because it provides insights into how likely rare attributes are, compared to other studies. Also the ML technique in use can require a minimum size of data to work at the intended performance.

III. Classification Results and Discussion

In this chapter we provide our results, elicited from table I. According to our methodology, we analyze first the different techniques, second the PSI types and third the data sets used.

A. Techniques for PSI detection

Figure 1 shows the most frequently mentioned techniques for PSI detection. We cluster the different techniques into (1) traditional machine learning techniques, (2) Neural Networks

TABLE I

SELECTION OF REVIEWED PAPERS AND SYSTEMATIZATION CRITERIA.

Work	Techniques tested	PSI type	Cross val.	Data source	Open access
[12]	BERT, LSTM	Personal, network identity, secret and credential, financial	None	Pastebin	Yes
[13]	LSTM, CNN	Health, political, and ethnic origin	None	Al-Ain news website, Ethio-WikiLeaks, tweeter, other	Partially
[14]	Word embedding		None	Wikipedia	NA
[15]	CRF, rule-based method	Date, name, age, contact, ID, location, profession	None	i2b2 NLP clinical	Yes
[16]	SVD, K-means, J48, Adaboost, KNN, SVM, RF, Spectral Clust.	Diagnosis	None	Twitter	NA
[17]	LogR, Simple Logistic, J48, FT, RF, NB, NNge, Heuristics	Location, psychological	None	myPersonality	No
[18]	SVM, LSTM	Contact information, credit history	None	Created with Mockaroo.	NA
[19]	SVM, NB, LR, LSTM	Sensitive to non-sensitive tweets.	5-fold	Twitter	NA
[20]	CNN, LSTM	Demographic, location, credit history, family/relationships, diagnosis, professional career	None	Medhelp, Amazon, Hotel reviews, POI reviews, Psychiatric forums, Twitter, Stack overflow	NA
[21]	LDA	Self-disclosure in COVID-19 related tweets	None	Twitter	Yes
[22]	RNN, LSTM	Personal, professional, health	10-fold	Twitter	No
[23]	BERT	Diagnosis	5-fold	Twitter	No
[24]	SVM, NB, GBM	Personal, professional, health related tweets	10-fold	Twitter	No
[25]	LSTM, CNN, DNN	Identity code	5-fold	MIDV dataset, DIQA, Ghega	Yes
[26]	CNN, LSTM	Financial account, contact information, computer device details, identity code	None	Synthetically generated, Enron, Trec07, ADCG SS14 Challenge corpora	Partially
[27]	Naive binary classifier	Contact number, email, birthdate, home town, current town, job details, relationship status, interests, religious views, political views	None	Survey/None	NA
[28]	LSTM	13 sensitivity classes from P3P (categories)	None	Reddit (by Googles BigQuery) + synthetic augmentation	Partially
[29]	NB, SVM, LogR, C4.5, fastText, CNN, RNN, LSTM	Binary: contains private information	10-fold	Sina Weibo	No
[30]	CNN, LogR, SVM, NB, kNN, RF, DT	Lexicon-based privacy disclosure features Privacy Word (PW)	5-fold	22 Datases	Yes
[31]	SVM, NB, GBM	Family/Relationship, professional career, financial account, contact information, password/pin, diagnoses, behavior, demographics	10-fold	Twitter	NA
[32]	LDA, FastText, CNN, Bi-GRU	Diagnoses	None	CLEF dataset (Reddit)	Yes
[33]	SVM, RF, CNN	Demographics, location, ethnicity, credit history	5-fold	News discourse dataset	No
[34]	rule-based matching	Self-disclosure of personal information	None	Twitter	Yes
[35]	NB, SVM, RF, TF-IDF	Location, medical conditions, personal attacks, stereotypes, family, association, personal details, PII, security	10-fold	Twitter	No
[36]	Heuristics, SVM, Word embedding	social security number, IP address	10-fold	Twitter	No
[24]	SVM, NB, GBM	Personal (personal/family-related information), professional (occupational, financial status, business oriented information), health-related (disease-related information, psychological status)	Yes	Twitter	No
[37]	K-Means, LDA, NLP	Privacy dictionary	None	Twitter	No
[38]	ANN, DNN, LSTM, TF-IDF, n-gram	Dengue disease	None	Twitter	No
[39]	NN, CNN	Name, ID-card number, telephone number, etc.	None	Apple Daily News	No
[40]	LDA, Privacy ontology, AdaBoost	location, medical, drug/alcohol, emotion, personal attacks, stereotyping, family or other associations, personal details, PI information	10-fold	Twitter	No
[41]	Rule-based method	Phrases in clinical narratives which express a sensitive condition about the patient	None	i2b2 data, Medical forums	Yes
[42]	RNN	Tweets including 23 cyber-keywords identified with the privacy information security classification (PISC) model	None	Twitter	No
[43]	CNN	Location, health, emotion, sociality, others	None	Microblog	No
[44]	TF-IDF, RNN, Bag of words	Health&medical, work, drugs&alcohol, obscenity, religion, politics, racism, family&personal, relationship, sexual orientation, travel, school life, entertainment	None	Twitter	No
[45]	NN	Health, relationship, finance	None	Twitter	No
[46]	BERT, LR, NB, SVM, RF, XGBoost	Political, religion, sexual orientation, drug_alcohol, racism, school life, family_personal, obscenity, entertainment, travel, relationship, medical, work	None	Twitter	No
[47]	NB, CNN, LSTM, LSTM+CNN	Diagnoses, emotion	None	Health blog	No
[48]	LR, Decision Trees, SVM, NB, RF	Sexual orientation, religious or philosophical beliefs, location or travel, political opinions, racial or ethnic, relationship, health, children, alcohol or drugs, family, personal attacks, association membership, emotions, personal information	10-fold	Twitter	No
[49]	Rule-based matching	Age, race, sexual orientation, location, affiliation, money, relationship, experience, interests, feelings, opinion	None	News comments	Yes

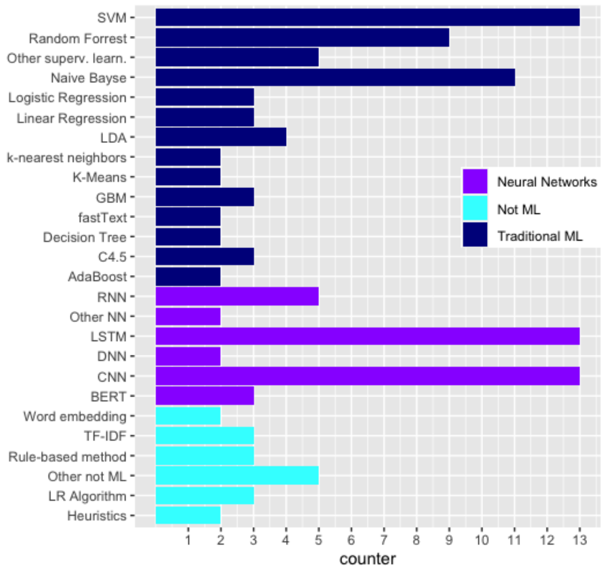

Fig. 1. PSI detection approaches

and (3) all other techniques that do not require ML. We consider all models mentioned in each paper. In total, we found 64 techniques related to traditional machine learning, 38 techniques related to neuronal networks and 18 techniques not using ML algorithms. Most of the ML algorithms are used for data classification but also a view for clustering tasks.

1) Classification models: Most papers, such as [18] follow a similar structure, applying a classification model. Most of the models found, rely on data that was collected from twitter. E.g., Wang [44] used snowball crawling to find users that are not celebrities or public accounts. The classification models have in common that to allocate text data to a certain group, labeled data is required. This labeling is achieved by either using pre-labeled data [18] or by generating the labeles. If the data is not labeled, the label-generation can be done by using Natural Language Processing (NLP). E.g., Casillo et al. [30] use keyword search utilizing a privacy dictionary to pre-label the data into privacy categories. Mittal et al. [46] also use keyword based privacy categorization to assign the tweets to a certain privacy category. Mehdy et al. [45] label their twitter data in disclosure or non-disclosure using 3 human annotators from Amazon Mechanical Turk per post. Also Umar et al. [49] used Amazon Mechanical Turk for labeling into 12 classes. Since data collected from social media such as Twitter is full of shortened words, colloquial language, abbreviations and emoticons [31], pre-processing is used to generate numeric attributes from the text data. Besides others, date/time, phone numbers and URLs are usually replaced. Another common step is tokenization where punctuation and insignificant words are removed, resulting in lexical text only [30], [47]. This allows the computation of word occurrence and vectorization of data. Especially for resolving hashtags and abbreviations, the text analysing tool GATE was used that was developed specifically for twitter posts [44]. But also syntactic and semantic structures are analyzed, providing common sequences of tokens and co-occurrence of dependency tags [30], [47]. More precisely, sentences with different syntactic form but the same meaning are represented in dependency parse trees [45]. Moreover, some papers, e.g. Mehdy et al. [47] take supplemental features or meta data into account. Such features can be number of pronouns, emotional tone or number of negations found in the text and are used to better separate classes. Also part of pre-processing but already nested with the classification is the use of pre-trained layers, e.g., Bidirectional Encoder Representations from Transformers (BERT) that allows transfer learning and custom NLP tasks [45].

Finally, a classification method is applied, e.g, LSTM [18], [47], SVM [18], [31], CNN [30], [47], NB [31], [47], GBM [30], [31], DNN [47]. The classification can be multiclass, classifying into different privacy categories [18] or binary, classifying into disclosure or non-disclosure of private data [30], [47]. Another application for binary classification is text classification into positive or negative emotions [46]. Geetha et al. [31] use a stacked classification where the sensitive keywords are used as features for reclassification to find the best performing classifier. Interesting is also the approach of Mehdy et al. [45] where two output layers are used to do simultaneously a binary disclosure/non-disclosure classification and a multiclass-classification into the three PSI classes health, finance and relationship.

2) Classification Performance: In general, it is not a trivial task to compare the performance of different classification models that exhibit differences in pre-processing and are trained and tested on different data sets. Although we analyzed the performance measurements of accuracy, precision, recall and F1-Score that were not always all reported in each paper, we want to provide rather an overall trend then comparing performance metrics in detail, especially because some of the papers already provide a detailed comparison of different techniques. First, as mentioned above, it is necessary to differentiate between binary and multiclass-classifications.

For binary classifications that predict for a given text whether PSI is disclosed or not disclosed, a very good performance was reported by Kopeykina and Savchenko [25] who achieved with their fully connected model a precision of 1.0, recall of 0.97, F1-score of 0.98 and error rate of 0.015. The results are validated with a 5-fold cross-validation approach.

For multiclass-classifications where a text is classified into several Privacy Sensitive Categories a high performing models was provided by [31] who achieve using naive bayes for a three classifying into (1) personal (accuracy: 0.8339, precision 0.695, recall, (2) professional and (3) health data with an average accuracy of 0.7524, recall of 0.736 and precision of 0.724. It is important to mention that the paper also uses a 10-fold cross validation approach and provides detailed scores of precision, recall and accuracy for each of the three classes. Utilizing a CNN Xu and Xu [43] achieve for a five class problem classifying into (1) location, (2) emotion, (3) health, (4) sociality, and (5) other, a precision of 0.9018. The class accuracy is for the best class 0.9698 and for the worst 0.8583. The results are validated in a 10-fold cross-validation. Further results with very high qulaity to mention were achieved by Guo et al. [12] who achieved with their combination BERT+BiLSTM+Attention matrix a precision of 0.9872, recall of 0.9958 and F1-score of 0.9915. Unfortunately, no information about cross validation is provided. In their approach, 4 word clouds, containing personal, network identity, secret and credential, and financial information are created. Overall, we find that if different ML techniques are compared, neuronal networks outperform traditional ML (e.g., [43], [47]). Future paper should focus on providing a sufficient validation of results, e.g., using cross validation and ensure sufficient scores for each class in multi class classification.

3) Clustering: In general, clustering is used to find similarities in texts. Blose et al. [21] use Latent Dirichlet Allocation (LDA) for topic generation of self-disclosing twitter posts during the COVID-19 pandemic. They generate 6 topics, and analyse the ratio of each at different times during the pandemic. De Luca [37] uses k-means for clustering data into 5 and 20 clusters. De Luca identified difficulties in using unsupervised learning in privacy models.

Overall, we found clustering be barely used in PSI detection. Nevertheless, it can be a method of choice if new privacy cluster need to be built (see Blose et al. [21]).

B. Types of Privacy Sensitive Information

Works in this area focus on different types of privacy-related information. Table II shows examples of the type of information detected in the analyzed papers, classified by category and type. The classification of category and type has been adapted from [11]. The examples are elicited from the papers analysed in table I and were mapped to the respective types. We have added the type psychological and addiction in the category medical health, because many PSI examples were found. We have deleted types if no example was found in table I.

As the table shows, the definition of what is considered private information or PSI can vary considerably, and can depend on the objectives of each approach. The type of information detected can also be generally classified into specific pieces of information and topics. Specific pieces of information refers to information such as phone numbers, addresses, government identification numbers (social security, passport, etc.) or credit card numbers. Since this type of information generally follows some pattern, it is usually identified by keyword search approaches [22]. Topics on the other hand, refer to information that does not necessarily fit a pattern that can be easily searched. The topics described by Wang et al. [19] have been used to classify private information revealed in social network site posts. Although a specific topic can contain specific keywords, detecting whether the information belongs to a privacy sensitive topic can be heavily influenced by subjective opinions, since it relies on human labeling.

As mentioned before, although the works in this area focus on the detection of private information, in some cases the information that is detected cannot be strictly classified as privacy sensitive information. Some examples are the topics of ”Personal attacks” or ”Obscenity” included in the classification by Wang et al. [19], which are described as sensitive but may or may not reveal private information. In a similar way, works in this area can include the detection of non-privacy sensitive information and still classify it as such. To some extent, the introduction of regulation such as the EU's GDPR has influenced the type of data and topics that is considered private or privacy-sensitive in the area of However, not only regulation but also factors such as cultural context can influence what is considered privacy sensitive information. Other type of information included in works in this area are pieces of information such as for example device ID, which by themselves are not necessarily private information, but which could be considered so if it is related or linked to other personal information about an individual. Other seemingly unrelated information could identify an individual if it is collected over a long period of time [50]. This indicates that a strict categorization of what consists private or privacy sensitive information, which would apply to every information detection approach, is not ideal or even possible. Instead, clearer definitions of the type of information included in the detection would serve to better evaluate the results of each approach. In the works reviewed, the information detected is sometimes defined only in general terms or the keywords used are not specified.

Besides the probability for PSI leakage depends on the emotional state. Mittal et al. [46] find that a negative emotional state, such as fear or anger increases the overall probability of PSI leakage. But they also find that a positive emotional state can increase the leakage of specific PSI types such as family or entertainment. These types can in our categorization be found in the category lifestyle-behaviour (see II).

C. Data sets

All data sets from the papers investigated, contain user generated data. 21 of the analyzed 39 paper use twitter as their main source for user generated text data. 10 paper use user generated comments from, e.g. blogs, articles or reviews, for products or hotels. A minority of papers as shown in table I utilizes synthetically generated data. Most of the data is not referenced in the model or not for free use. One reason for that is that most of the Twitter data was achieved by using scrapers, collecting the data over several month.

IV. Research Trends and Gaps

In this section, we provide further details in our observations regarding the research trends and challenges in this arena.

A. Research Trends

1) Deep learning: Deep learning techniques, e.g., CNNs and RNNs, are highly effective for detecting PSI in textual data. In recent years, more and more papers have used DNNs to detect PSI. We foresee this trend might continue as the research and development of deep learning intensifies.

2) Transfer learning: In some of the latest papers on auto-detecting PSI, transfer learning has shown to be effective. This can be particularly useful when labeled PSI data is limited.

TABLE II
CLASSIFICATION OF TYPES OF INFORMATION DETECTED, BASED ON [11]

Category	Type	Examples
Lifestyle-behaviour	Beliefs	Religion, religious views, philosophical beliefs
	Preferences or interests	Entertainment, membership
	Behavior	Personal attacks, obscenity, racism, stereotypes, school life, travel
	Family/Relationship	Children, spouse, family, interpersonal relationships, relationship status
	Politics	Politics
Social-economic	Ethnicity	Ethnic origin, race
	Demographics	Height, weight, name, date, time, age, date of birth, sexual orientation
	Professional career	Business, profession, job, affiliation, job details
Tracking	Contact information	email address, phone number, address, Home town
	Location	Location
	Computer device details	IP address (IPv4, IPv6), MAC address, UUID, URL
Financial	Credit history	Monetary information, money
	Financial account	Bank account no., credit card no.
Authenticating	Passwords and pin	Password, encryption key
	Identity Code	Passport, license card, social sec. no.
Medical-health	Diagnoses	disease, medical condition, blood type, eating disorders, specific disease, pregnancy, Covid-19
	Addiction	Drugs, alcohol
	Psychological	Psychological status, emotion

B. Research Gaps

During the course of this study, we observed the following gaps in the field. We elaborate these gaps so that future researchers would take up the challenge and fill the void.

1) Data quality and availability: The availability and quality of data is essential for training ML models. However, we observed that many researchers do not publicly avail their data sets. This can be in part due to the fact that the research is dealing with PSI hence publishing the data set might threaten the privacy of participants.

2) Real-time detection: The focus in this paper is to synthesize research works in detecting PSI. As such, the real-time detection of sensitive information is increasingly important, especially in scenarios such as online social networks. However, it is our observation that many of the approaches do not consider real-time detection and protection. Moreover, generalizability remains largely unaddressed. PSI detection tools should, at best, work across different platforms and scenarios. Ensuring that models are robust and can handle different types of data is crucial for real-world applications.

3) Interpretability and explainability: ML models are often considered black boxes, mainly because it is difficult to comprehend how they arrive at their decisions. ML-based PSI detection tools need to be interpretable and explainable to gain trust by users in the results they produce. There is a gap in addressing the interpretability and explainability aspects in the PSI detection approaches.

4) Multimodal detection: Multimodal ML models that can detect sensitive information in text, images, and other data types are largely lacking. As PSI can be divulged in either or a combination of text, images, and other forms of data, developing multimodal systems becomes of paramount importance to safeguard the privacy of users. These models would provide a more comprehensive protection of PSI, allowing for greater protection and compliance with data protection regulations.

5) Ethics and fairness: Ethics and fairness are becoming integral parts of responsible AI development. Ensuring that ML models that detect PSI are ethical and fair is crucial. Models should not discriminate against individuals based on their race, sexual orientation, religious or political beliefs, gender, or other factors.

6) Regulatory compliance: Adhering to regulatory frameworks and standards is essential when dealing with PSI. Thus, ensuring that ML models meet regulatory requirements is important not only for compliance but also for developing solutions that can be adopted in real-world applications.

V. CONCLUSIONS

In summary, PSI detection using machine learning is a rapidly growing research field that has the potential to have a significant impact on enhancing the privacy of users online. Machine learning models have demonstrated to be quite promising in detecting PSI, and there has been remarkable progress in developing effective models for this purpose. Results in this paper indicate that PSI detection tools based on deep neural networks have demonstrated to be more performant than other methods. However, there are still significant challenges and research gaps that need to be addressed in this field. These include the lack of standardized datasets, the need for more research on small datasets, the limited research on multimodal detection and sensitive information detection including in non-English languages, the need for more research on transferability across domains and adversarial attacks, and the need for more research on real-world deployment, explainability, and privacy-preserving methods. Addressing these challenges and research gaps will require collaboration among researchers from different disciplines and the development of new methods and techniques. Overall, PSI detection using machine learning is a promising research field that has the potential to help protect privacy in a wide range of applications, including social media networks, but more research is needed to ensure the development of explainable, usable, effective, ethical and legally compliant solutions.

REFERENCES

[1] R. LaRose and N. J. Rifon, "Promoting i-safety: effects of privacy warnings and privacy seals on risk assessment and online privacy behavior," *Journal of Consumer Affairs*, vol. 41, no. 1, 2007.

[2] S. Löbner, B. Gogov, and W. B. Tesfay, "Enhancing privacy in federated learning with local differential privacy for email classification," in *International Workshop on Data Privacy Management*. Springer, 2022, pp. 3–18.

[3] V. Bracamonte, S. Pape, and S. Loebner, "all apps do this: Comparing privacy concerns towards privacy tools and non-privacy tools for social media content," *Proceedings on Privacy Enhancing Technologies*, vol. 3, pp. 57–78, 2022.

[4] EU, "Regulation (eu) 2016/679 of the european parliament and of the council on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (general data protection regulation)," 2016.

[5] Artificial Intelligence Act, "Proposal for a regulation of the european parliament and the council laying down harmonised rules on artificial intelligence and amending certain union legislative acts," *EUR-Lex*, 2021.

[6] S. E. Erol, Ş. SAĞIROĞLU, and U. DEMİREZEN, "Detecting personal health data disclosures in turkish social data," *International Journal of Information Security Science*, vol. 11, no. 2, pp. 69–84, 2022.

[7] A. García-Pablos, N. Perez, and M. Cuadros, "Sensitive data detection and classification in spanish clinical text: Experiments with bert," *arXiv preprint arXiv:2003.03106*, 2020.

[8] C. M. Bishop, *Pattern recognition and machine learning*. springer, 2006.

[9] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, http://www.deeplearningbook.org, accessed: 01/01/2021.

[10] J. D. Kelleher, B. Mac Namee, and A. D'arcy, *Fundamentals of machine learning for predictive data analytics: algorithms, worked examples, and case studies*. MIT press, 2020.

[11] H. N. Chua, J. S. Ooi, and A. Herbland, "The effects of different personal data categories on information privacy concern and disclosure," *Computers & Security*, vol. 110, p. 102453, 2021.

[12] Y. Guo, J. Liu, W. Tang, and C. Huang, "Exsense: Extract sensitive information from unstructured data," *Computers & Security*, vol. 102, p. 102156, 2021.

[13] A. Genetu and T. Tegegne, "Designing sensitive personal information detection and classification model for amharic text," in *2021 International Conference on Information and Communication Technology for Development for Africa (ICT4DA)*. IEEE, 2021, pp. 54–58.

[14] F. Hassan, D. Sánchez, J. Soria-Comas, and J. Domingo-Ferrer, "Automatic anonymization of textual documents: detecting sensitive information via word embeddings," in *TrustCom/BigDataSE*. IEEE, 2019.

[15] H. Yang and J. M. Garibaldi, "Automatic detection of protected health information from clinic narratives," *Journal of biomedical informatics*, vol. 58, pp. S30–S38, 2015.

[16] S. R. M. Castillo and Z. Chen, "Using transfer learning to identify privacy leaks in tweets," in *Collaboration and Internet Computing (CIC), 2016 IEEE 2nd International Conference on*. IEEE, 2016, pp. 506–513.

[17] J. Alemany, E. Del Val, and A. García-Fornes, "Empowering users regarding the sensitivity of their data in social networks through nudge mechanisms." in *HICSS*, 2020, pp. 1–10.

[18] A. K. Makhija, "Deep learning application–identifying pii (personally identifiable information) to protect."

[19] Q. Wang, H. Xue, F. Li, D. Lee, and B. Luo, "# donttweetthis: Scoring private information in social networks," *Proceedings on Privacy Enhancing Technologies*, vol. 2019, no. 4, pp. 72–92, 2019.

[20] N. Mehdy, C. Kennington, and H. Mehrpouyan, "Privacy disclosures detection in natural-language text through linguistically-motivated artificial neural networks," in *SPNCE 2019, Tianjin, China, April 13–14, 2019, Proceedings 2*. Springer, 2019, pp. 152–177.

[21] T. Blose, P. Umar, A. Squicciarini, and S. Rajtmajer, "Privacy in crisis: A study of self-disclosure during the coronavirus pandemic," *arXiv preprint arXiv:2004.09717*, 2020.

[22] R. Geetha, S. Karthika, N. Pavithra, and V. Preethi, "Tweedle: sensitivity check in health-related social short texts based on regret theory," *Procedia Computer Science*, vol. 165, pp. 663–675, 2019.

[23] P. Umar, C. Akiti, A. Squicciarini, and S. Rajtmajer, "Self-disclosure on twitter during the covid-19 pandemic: A network perspective," in *ECML PKDD 2021, Bilbao, Spain, September 13–17, 2021, Proceedings, Part IV 21*. Springer, 2021, pp. 271–286.

[24] R. Geetha, S. Karthika, and P. Kumaraguru, "Tweet-scan-post: a system for analysis of sensitive private data disclosure in online social media," *Knowledge and Information Systems*, vol. 63, pp. 2365–2404, 2021.

[25] L. Kopeykina and A. Savchenko, "Photo privacy detection based on text classification and face clustering," in *Proceedings of ITNT-IPERS 2020*, vol. 2665, 2020, pp. 171–176.

[26] A. Truong, A. Walters, and J. Goodsitt, "Sensitive data detection with high-throughput neural network models for financial institutions," *arXiv preprint arXiv:2012.09597*, 2020.

[27] A. Srivastava and G. Geethakumari, "Measuring privacy leaks in online social networks," in *2013 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*. IEEE, 2013.

[28] A. Pogiatzis and G. Samakovitis, "Using bilstm networks for context-aware deep sensitivity labelling on conversational data," *Applied Sciences*, vol. 10, no. 24, p. 8924, 2020.

[29] Y. Xu, X. Meng, Y. Li, and X. Xu, "Research on privacy disclosure detection method in social networks based on multi-dimensional deep learning," *Comput. Mater. Contin*, vol. 62, pp. 137–155, 2020.

[30] F. Casillo, V. Deufemia, and C. Gravino, "Detecting privacy requirements from user stories with nlp transfer learning models," *Information and Software Technology*, vol. 146, p. 106853, 2022.

[31] R. Geetha, S. Karthika, and P. Kumaraguru, "will i regret for this tweet?twitter users behavior analysis system for private data disclosure," *The Computer Journal*, vol. 65, no. 2, pp. 275–296, 2022.

[32] B. Tébar and A. Gopalan, "Early detection of eating disorders using social media," in *2021 IEEE/ACM Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE)*. IEEE, 2021, pp. 193–198.

[33] P. Umar, A. Squicciarini, and S. Rajtmajer, "A study of self-privacy violations in online public discourse," in *2020 IEEE International Conference on Big Data (Big Data)*. IEEE, 2020, pp. 1041–1050.

[34] T. Blose, P. Umar, A. Squicciarini, and S. Rajtmajer, "A study of self-disclosure during the coronavirus pandemic," *First Monday*, 2021.

[35] L. Deodhar, D. M. Divakaran, and M. Gurusamy, "Analysis of privacy leak on twitter," in *GLOBECOM 2017-2017 IEEE Global Communications Conference*. IEEE, 2017, pp. 1–6.

[36] Y. Karimi, A. Squicciarini, and S. Wilson, "Automated detection of doxing on twitter," *Proceedings of the ACM on Human-Computer Interaction*, vol. 6, no. CSCW2, pp. 1–24, 2022.

[37] P. De Luca, "Detecting private information in large social network using mixed machine learning techniques," 2019.

[38] S. Amin, M. I. Uddin, S. Hassan, A. Khan, N. Nasser, A. Alharbi, and H. Alyami, "Recurrent neural networks with tf-idf embedding technique for detection and classification in tweets of dengue disease," *Ieee Access*, vol. 8, pp. 131 522–131 533, 2020.

[39] Y. Lin, G. Xu, G. Xu, Y. Chen, and D. Sun, "Sensitive information detection based on convolution neural network and bi-directional lstm," in *2020 IEEE 19th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)*. IEEE, 2020.

[40] A. Caliskan Islam, J. Walsh, and R. Greenstadt, "Privacy detective: Detecting private information and collective privacy behavior in a large social network," in *Proceedings of the 13th Workshop on Privacy in the Electronic Society*. ACM, 2014, pp. 35–46.

[41] P. Jindal, C. A. Gunter, and D. Roth, "Detecting privacy-sensitive events in medical text," in *Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*. ACM, 2014, pp. 617–620.

[42] R. Geetha, S. Karthika, and S. Mohanavalli, "Tweet classification using deep learning approach to predict sensitive personal data," in *Advances in Electrical and Computer Technologies: Select Proceedings of ICAECT 2019*. Springer, 2020, pp. 171–180.

[43] X. Huajun and X. Yabin, "Research on privacy disclosure detection in microblog," in *2017 3rd IEEE International Conference on Computer and Communications (ICCC)*. IEEE, 2017, pp. 1479–1486.

[44] Q. Wang, "Towards the understanding of private content–content-based privacy assessment and protection in social networks," Ph.D. dissertation, University of Kansas, 2020.

[45] A. Mehdy and H. Mehrpouyan, "A multi-input multi-output transformer-based hybrid neural network for multi-class privacy disclosure detection," *arXiv preprint arXiv:2108.08483*, 2021.

[46] M. Mittal, M. R. Asghar, and A. Tripathi, "Do my emotions influence what i share? analysing the effects of emotions on privacy leakage in twitter," in *2020 TrustCom*. IEEE, 2020, pp. 1228–1235.

[47] A. Mehdy and H. Mehrpouyan, "A user-centric and sentiment aware privacy-disclosure detection framework based on multi-input neural network," in *PrivateNLP 2020, in Houston, Texas, USA.*, 2020.

[48] W. B. Tesfay, J. Serna, and K. Rannenberg, "Privacybot: detecting privacy sensitive information in unstructured texts," in *2019 sixth international conference on social networks analysis, management and security (SNAMS)*. IEEE, 2019, pp. 53–60.

[49] P. Umar, A. Squicciarini, and S. Rajtmajer, "Detection and analysis of self-disclosure in online news commentaries," in *The World Wide Web Conference*, 2019, pp. 3272–3278.

[50] W. B. Tesfay, J. Serna, and S. Pape, "Challenges in detecting privacy revealing information in unstructured text." in *PrivOn@ ISWC*, 2016.