Journal of Accounting, Finance, Economics, and Social Sciences
Vol 5, No 2, July – December 2020, pp. 10-16
Received: Aug 2020 | Revised: Oct 2020 | Accepted: Dec 2020
DOI: https://doi.org/10.62458/jafess.160224.5(2)10-16

Deep Learning Application – Identifying PII (Personally Identifiable Information) to Protect

Anil K. Makhija*

ABSTRACT

This paper presents application of deep learning and machine learning models in detecting personally identifiable information (PII) in unstructured text (emails). The proposed models use support vector machine (trained using sequential minimal optimization) and long short term memory (LSTM) artificial neural network. Synthetic email dataset has been used to train and validate the proposed models and the outcomes are measured by standard measures of accuracy, precision, recall and F1-score of each of the proposed model. The experimental results on the model that uses support vector machine (trained using sequential minimal optimization) showed most promising results on detecting the personally identifiable information in the email dataset. The LSTM model also showed equally promising results.

Keywords: Personally Identifiable Information, Deep Learning in detecting PII, Machine Learning in detecting PII, Artificial Intelligence in protecting privacy, Protecting Personally Identifiable Information.

1. INTRODUCTION

Technological advances and proliferation of internet and online social network has made the entire world super-connected. Organizations have tremendous focus on provide best in class customer experience and are thus leveraging technology to enable it. In order to provide best in class customer experience and to provide personalized recommendations to the users, organizations gather lot of personal data and information from their customers. This also creates a risk to users' private sensitive information, especially their personally identifiable information (PII) being leaked to users with malicious intents, putting user's privacy at risk. One of the important steps in protecting PII is identifying the PII and protecting it. In recent years, research has focused on applying machine learning algorithms to identify PII. The advances in deep learning present an opportunity to apply deep learning algorithms to identify PII. This research proposes machine learning and deep learning models, to identify the PII in the unstructured text data. Support vector machine, training using sequential minimal optimization model and long sort term memory (LSTM) based models are trained and tested for accuracy, precision, recall and F1 score. Both the models give promising results in detecting PII. The SVM model performance was most promising.

* Anil K. Makhija, B.E., PGDIM, MBA. Lecturer, CamEd Business School.
Email: anil@cam-ed.com

2. RELATED WORK

What is Personally Identifiable Information:

With the technological advances and digital becoming reality for businesses and governments across the globe, personal data of individuals is being collected at an ever-increasing scale. Information about web-searches, browsing history, social relationships, medical history and many other similar data is collected and shared with business organizations, advertisers, government agencies, researchers and so on. A significant portion of this data can be information that can be used to identify the person individually, directly or indirectly (Narayanan & Shmatikov, 2010). Such information is classified as "personally identifiable information" or PII. Some practitioners argue that even when some information can be used to trace an individual's identity when combined with other public information, then also the information in consideration shall be classified as personally identifiable information. Large scale popularity of online social networks has also resulted in significant increase in amount of personal information available on internet (Krishnamurthy & Wills, 2009). Anonymity and international reach of internet create an ideal environment for cyber criminals who employ advanced persistent threat (APT) attacks over the online social network to extract information about organization, about users, and leveraging it for cyber stalking and identity theft

(Louw & von Solms, 2013).

Large scale proliferation and usage of social networking sites (SNS) and focus of businesses in providing user centric services have also contributed to vast amount of PII becoming available over the internet. In authenticating a customer's identity, organizations make extensive use of personally identifiable information. While many of the social networking sites are free, there are multiple instances where PII breaching has been done by these organizations. Organizations specifically do user profiling using this PII and utilize the outcomes of this profiling to make their business models more effective. Large organizations and tech giants outsource the customer PII mining activities to the third-party service providers. These third-party service provider companies are servicing multiple corporate clients and hence user data, containing their personal identifiable information is moving across multiple organizations and entities without they knowing about it (Al-Zaben et al., 2018).

There are several advantages both for the organizations as well as users, as this data helps create data-driven approach in delivery of customer service and in meeting customer expectations, resulting in increased customer satisfaction levels. However, there are many instances of malicious and unauthorized use of this data. As per April 2018 report of The Guardian, more than 50 million Facebook profiles were harvested for Cambridge Analytica, in a major data breach (Cadwalladr & Graham-Harrison, 2018).

Data breach can happen both unintentionally as well as intentionally. IN 2013, more than 40 million credit / debit card numbers were stolen from Target's point of sale terminal system. Information leak from 56 million credit cards from Home Depot in 2014, stealing of PII of 79 million customers of Anthem in 2018, and exposure of social security numbers, drivers' license numbers and passport numbers affecting 146 million people due to data breach at Equifax show how alarming the problem of data breaches and breach of PII is (Poyraz et al., 2020).

In order to personalize the experience, which eventually helps in increased sales and better returns, data gathering, storage and analytics is pervasive in all devices, systems, applications, and platforms. This coupled with Internet of Things (IoT) getting integrated into almost all the systems that are used in daily life, and gaps in privacy regulations increases the risk users' privacy breaches (Isaak & Hanna,2018).

How PII is detected and Protected:

An organization's information privacy safeguards have significant influence on how an individual's PII is protected by the organization. There are multiple threat vectors that operate and organization needs to safeguard information privacy from all of those threat vectors in order to keep the PII safe (Posey et al., 2017). Acts such as GDPR – General Data Protection Regulation have specific focus on protecting personally identifiable information (Tikkinen-Piri, 2018)

One of the biggest challenges in protecting the personally identifiable information is identification of whether a piece of information is PII or not. When it comes to organizations that have significant user interaction in the form of email and chats, the personally identifiable information may be contained in the text form in those emails and chats. Similarly, contract documents, agreements, medical records and so on may also contain personally identifiable information in the text form. Hence it is imperative that organizations create robust mechanism to identify if the information contained in a document or email or agreement is PII or not. Only once the company knows that the information under consideration is PII, it can take steps to ensure that the PII is safe and not breached.

Some of the initial research work in this area was related to development of a tool that automatically harvested the identifiers from the user's computer and active directory and then searching various data encodings using fast search algorithms and regular expression matching (Aura et al., 2006). With the advances in technology, specifically artificial intelligence and machine learning, there has been more focus in recent researches on using artificial intelligence and machine learning algorithms to detect personally identifiable information or any private and sensitive information. Artificial intelligence involves intelligent agents (devices) that perceive environment and take action in order to maximize the goal attainment (Ongsulee, 2017). Machine learning involves developing computer systems that can learn automatically and improve with the experience. Machine learning is a method of choice in developing and implementing artificial-intelligence based systems. Rapid growth in our ability to gather huge amount of data (big data) has led to researchers, scientists and practitioners focus on turning to this data to provide insights, and help in developing systems that can learn, predict and

decide based on this data (Jordan & Mitchell, 2015).

Machine learning is classified as supervised learning and unsupervised learning. Supervised learning systems learn mapping from the labelled data and then use it to make predictions. In unsupervised machine learning, unlabeled data discovers information and patterns on its own and utilizes it to make predictions (Ozgur, 2004). When algorithms don't use labeled data and instead, they utilize artificial neural network (ANN) layers, the approach is known as deep-learning. Deep learning is a field that simulates human brain, through the ANN, for analytical learning. Deep learning algorithms require larger amount of data, than a machine learning algorithm, in order to perform well. Compared to machine learning algorithms, deep learning reduces the effort of designing a feature extractor since it obtains high level features directly from the data. In comparison to a machine learning algorithm, training a deep learning algorithm usually takes longer though testing time is shorter for a deep learning algorithm. Another important difference in machine learning and deep learning is that whereas a machine learning algorithm provides explicit details of the results arrival process, it is not so clearly explainable in the case of deep learning algorithm (Xin et al, 2018).

Detecting that a piece of information comes under the purview of privacy is the most significant step in ensuring that this information is not divulged over the internet or otherwise. Research in this area is in very initial stages. There have been few machine-learning based models proposed for identification of PII in emails, specifically email addresses, monetary information, telephone number and addresses. There has been some research on developing automatic learning systems based on Naïve Bayes. There has been some work done on semi-supervised machine learning based detection of personal health information in health records, and development of machine learning based PrivacyBot. Past research in this area also identifies the need and potential to develop deep neural network based and other similar models to detect private sensitive information (Tesfay et al., 2019).

Effectiveness of the machine learning and deep learning algorithms, applied to any context, is assessed by following metrics. The key metrics that are used are Precision, Recall, and F1-score (Apruzzese et al., 2018).

- Accuracy is defined as ration of correct predictions to the total predictions made. For binary predictions, it can be defined as ratio of sum of true positives and true negatives to the sum of true positives, true negatives, false positives and false negatives (Korotcov et al., 2017).
- Precision is defined as ratio of true positives to total positives (including both true and false positives). It indicates the probability that a predicted true event or label is indeed a true label (Korotcov et al., 2017).
- Recall is the ratio of true positives to sum of true positives and false negatives. It helps one know as to what proportion of actual positives were correctly identified.
- F1-score is harmonic mean of precision and recall and its value is 1 at a perfect precision and prefect recall (Almseidin et al., 2017).

Current Status & Research Question:

As evident from above analysis, privacy is one of the areas which is of extreme importance and faces multiple threat vectors, especially due to technological advances and proliferation of internet, internet of things, and online social networks. Personally-Identifiable-Information (PII) needs to be protected by organizations in order to protect the privacy of their users and customers. One of the important steps in order to protect PII is to identify the PII in an effective way. In recent years, machine learning based models have been applied to identify PII automatically.

Researchers in this area identify the need to use deep learning algorithms to build systems that automatically identify the PII. This research aims to apply deep learning algorithms, such as RNN (recurrent neural network) on emails and / or documents to identify the PII contained in those emails and documents. The effectiveness of those models will be assessed using the measures of accuracy, precision, recall and F1-score.

3. RESEARCH METHODOLOGY & PROPOSED MODELS

Research Data & Research Methodology

This research uses synthetic email dataset, created using mockaroo (Whelan, 2014). The data created is multi class dataset with total of eight classes. A total of 4796 emails are created, out of which 4010 emails contain no personally identifiable information. Remaining 786 emails contain personally identifiable

information. There are 262 emails that contain address information, 223 emails contain credit card numbers, and 241 emails contain name information. There are 60 emails that contain combination of more than one personally identifiable information. Email dataset details are shown in figure 1.

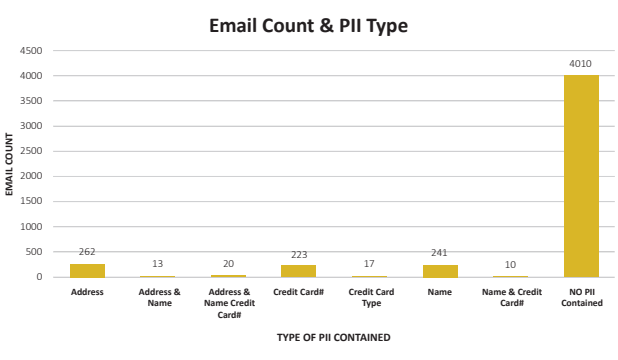

Figure 1: Research Dataset Details – Email Count for Each PII Type

Research uses three models which are based on support vector machine classifier (that uses sequential minimal optimization or SMO) and long short term memory (LSTM) and applies them for multi-class classification (for SMO) as well as binary classification (for SMO and LSTM). Email dataset is split into 80% and 20% buckets. Entire email dataset, which is text dataset, preprocessed and is converted to vector form and then used by SMO and LSTM algorithms. SVMs were designed for binary classification approach though it can be extended to multi-class classification as well and they are relatively insensitive to the relative numbers in each of the class (Druker et al., 1999; Platt, 1998; Mathur & Foody, 2008). LSTM are special kind of RNN (recurrent neural network) that help address the long-term dependency issue faced in RNN (Pienaar & Malekian, 2019). In the models proposed in this research, LSTM is applied once using two hidden LSTM layers and once using three hidden LSTM layers in DL4JMlpclassifier in WEKA. The 80% email dataset bucket is used to train the algorithm and 20% email dataset bucket is used to validate the algorithm. Accuracy, Precision, Recall and F1-scores are recorded for the validations to analyze the results. WEKA is used to build, train, validate and test the proposed models (Eibe et al., 2016; Lang et al, 2019).

Proposed Model(s) Details

To answer the given research question, three models have been proposed in this research paper. First model (SMO-SVM) is trained and validated for both multi-class classification and binary classification. Other two models (LSTM-2HDL, LSTM-3HDL) are trained and validated for binary classification.

Model 1 (SMO-SVM):

(a) Stage 1: Data Pre-processing- Converts email text into a set of numeric attributes that represent word occurrence information of the text contained in the emails.

(b) Stage 2: Applying Sequential Minimal Optimization, or SMO algorithm for training support vector machine

Model 2 (LSTM-2HDL):

(a) Stage 1: Data Pre-processing- Converts email text into a set of numeric attributes that represent word occurrence information of the text contained in the emails.

(b) Stage 2: Applying LSTM, a special type of recurrent neural network, with 2 hidden layers (number of outputs in hidden layer 8 and then 4), with training data normalized.

Model 3 (LSTM-3HDL):

(a) Stage 1: Data Pre-processing- Converts email text into a set of numeric attributes that represent word occurrence information of the text contained in the emails.

(b) Stage 2: Applying LSTM, a special type of recurrent neural network, with 3 hidden layers (number of outputs in hidden layer 8 and then 6 and then 4), with training data normalized.

All these models are built, trained, validated and tested using WEKA (Lang et al., 2019).

4. EXPERIMENTAL RESULTS & ANALYSIS

Experiment Design

The proposed models were built and trained using the synthetic email data. Out of total 4796 instances in the dataset, 80% (3837 instances) were used to train the model, whereas remaining 20% (959 instances) were used to validate the models. For first model (SMO-SVM), multi-class labeled data and binary class labeled was used whereas for the other two models (LSTM-2HDL, LSTM-3HDL), binary class labeled data

was used. Training and validation of the models was done using WEKA. Results of the validation and testing are summarized below.

Validation Results

Validation results from WEKA on the 3 models proposed in this research paper are summarized in table 1.

Model	Classification Type	Accuracy (%)	TP	FP	Precision	Recall	F1-Score
SMO-SVM	Multi Class	91.24	0.912	0.349		0.912	
SMO-SVM	Binary	93.64	0.936	0.304	0.941	0.936	0.930
LSTM-2HDL (8,4)	Binary	93.19	0.931	0.329	0.936	0.931	0.924
LSTM-3HDL (8,6,4)	Binary	92.60	0.339	0.339	0.929	0.926	0.918

Table 1: Accuracy, YP, FP, Precision, Recall & F1-Score summary

Results Analysis

Accuracy: As shown in figure 2, model validation results indicate that sequential minimal optimization algorithm based training of SVM (SMO-SVM model) gives that highest level of accuracy (99.6%+) for the validation dataset, amongst the models evaluated. The next best results in terms of accuracy are given by long short term memory (LSTM) based model (LSTM-2HDL model).

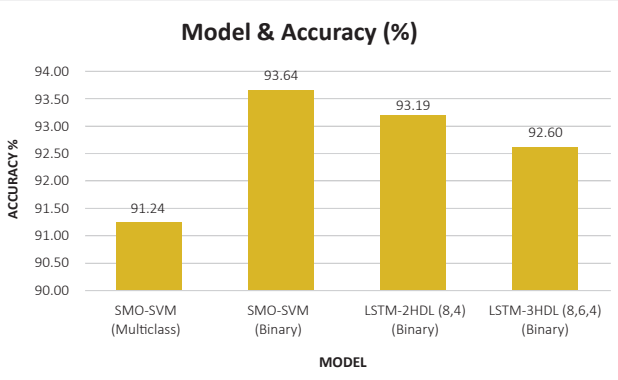

Figure 2: Model Type & Accuracy % Comparison

Precision: As shown in figure 3, model validation results indicate that sequential minimal optimization algorithm based training of SVM (SMO-SVM model) gives that highest level of precision value of 0.941 closely followed by long short term memory (LSTM) based model (LSTM-2HDL model) which gives precision value of 0.936.

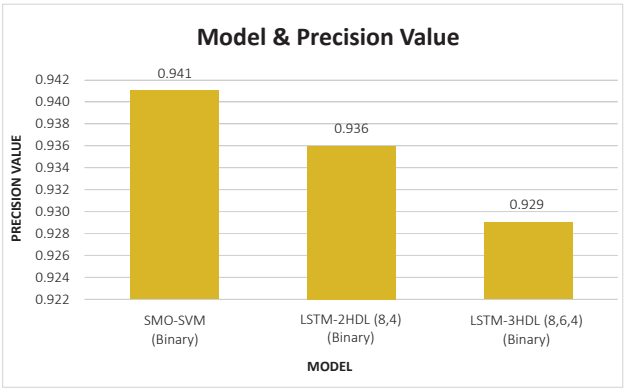

Figure 3: Model Type & Precision Value Comparison

Recall: As shown in figure 4, model validation results indicate that sequential minimal optimization algorithm based training of SVM (SMO-SVM model) gives that highest level of recall value of 0.936 closely followed by long short term memory (LSTM) based model (LSTM-2HDL model) which gives recall value of 0.931.

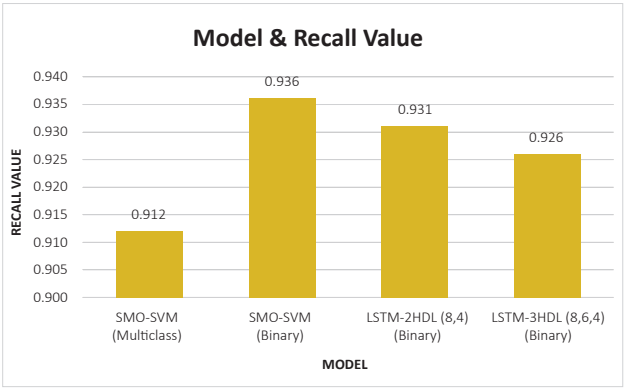

Figure 4: Model Type & Recall Value Comparison

F1-Score: As shown in figure 5, model validation results indicate that sequential minimal optimization algorithm based training of SVM (SMO-SVM model) gives that highest level of F1-score of 0.930 closely followed by long short term memory (LSTM) based model (LSTM-2HDL model) which gives recall value of 0.924.

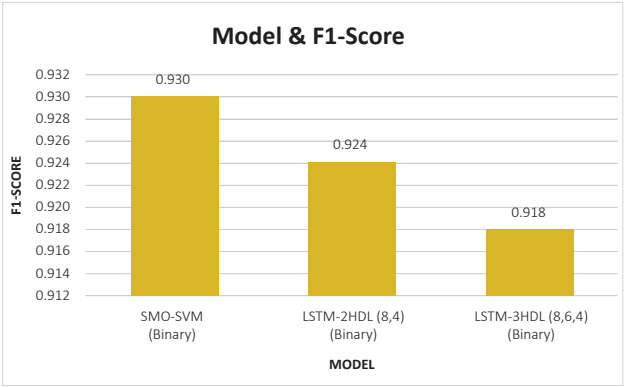

Figure 5: Model Type & F1-score Comparison

5. CONCLUSION & FUTURE WORK

It is evident that organizations can achieve 90%+ predictability in identifying emails (unstructured text) that contain personally identifiable information. Both machine learning and deep learning approaches demonstrate promising results in identifying personally identifiable information (PII) contained within unstructured text (in the form of emails). Organizations can leverage the proposed models to analyze text information to flag any information that is outbound (or inbound) and contains PII. The instances where information that was PII but not flagged can still be covered using a policy framework and putting accountability on the employees handling such information. However, a first level classification using a machine learning or deep learning model will help organizations improve their compliance to various laws and regulations, such as GDPR, which require organizations to classify and protect personally identifiable information. The accuracy demonstrated by the models proposed in this research was between 93 to 94%. Since protecting PII is gaining importance across the industry spectrum, further work can be done in creating models that provide even higher levels of accuracy in detecting personally identifiable information.

6. REFERENCES

Almseidin, M., Alzubi, M., Kovacs, S., & Alkasassbeh, M. (2017). Evaluation of machine learning algorithms for intrusion detection system. 2017 IEEE 15th International Symposium on Intelligent Systems and Informatics (SISY). doi:10.1109/sisy.2017.8080566

Al-Zaben, N., Hassan Onik, M. M., Yang, J., Lee, N.-Y., & Kim, C.-S. (2018). General Data Protection Regulation Complied Blockchain Architecture for Personally Identifiable Information Management. 2018 International Conference on Computing, Electronics & Communications Engineering (iCCECE). doi:10.1109/iccecome.2018.8658586

Apruzzese, G., Colajanni, M., Ferretti, L., Guido, A., & Marchetti, M. (2018). On the effectiveness of machine and deep learning for cyber security. 2018 10th International Conference on Cyber Conflict (CyCon). doi:10.23919/cycon.2018.8405026

Aura, T., Kuhn, T. A., & Roe, M. (2006). Scanning electronic documents for personally identifiable information. Proceedings of the 5th ACM Workshop on Privacy in Electronic Society - WPES '06. doi:10.1145/1179601.1179608

Cadwalladr, C., & Graham-Harrison, E. (2018). Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach. The guardian, 17, 22.

Drucker, H., Donghui Wu, & Vapnik, V. N. (1999). Support vector machines for spam categorization. IEEE Transactions on Neural Networks, 10(5), 1048–1054. doi:10.1109/72.788645

Eibe Frank, Mark A. Hall, and Ian H. Witten (2016). The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques", Morgan Kaufmann, Fourth Edition, 2016

Isaak, J., & Hanna, M. J. (2018). User Data Privacy: Facebook, Cambridge Analytica, and Privacy Protection. Computer, 51(8), 56–59. doi:10.1109/mc.2018.3191268

J. C. Platt, "Sequential minimal optimization: A fast algorithm for training support vector machines," in Advances in Kernel Method: Support Vector Learning, Scholkopf, Burges, and Smola, Eds. Cambridge, MA: MIT Press, 1998, pp. 185–208

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. Science, 349(6245), 255–260. doi:10.1126/science.aaa8415

Korotcov, A., Tkachenko, V., Russo, D. P., & Ekins, S. (2017). Comparison of Deep Learning With Multiple Machine Learning Methods and Metrics Using Diverse Drug Discovery Data Sets. Molecular Pharmaceutics, 14(12), 4462–4475. doi:10.1021/acs.molpharmaceut.7b00578

Krishnamurthy, B., & Wills, C. E. (2009). On the leakage of personally identifiable information via online social networks. Proceedings of the 2nd ACM Workshop on Online Social Networks - WOSN '09. doi:10.1145/1592665.1592668

Lang, S., Bravo-Marquez, F., Beckham, C., Hall, M., & Frank, E. (2019). WekaDeeplearning4j: A deep learning package for weka based on Deeplearning4j. Knowledge-Based Systems. doi:10.1016/j.knosys.2019.04.013

14. Louw, C., & von Solms, S. (2013). Personally identifiable information leakage through online social networks. Proceedings of the South African Institute for Computer Scientists and Information Technologists Conference on - SAICSIT '13. doi:10.1145/2513456.2513467

Mathur, A., & Foody, G. M. (2008). Multiclass and Binary SVM Classification: Implications for Training and Classification Users. IEEE Geoscience and Remote Sensing Letters, 5(2), 241–245. doi:10.1109/lgrs.2008.915597

Narayanan, A., & Shmatikov, V. (2010). Myths and fallacies of "personally identifiable information." Communications of the ACM, 53(6), 24. doi:10.1145/1743546.1743558

Ongsulee, P. (2017). Artificial intelligence, machine learning and deep learning. 2017 15th International Conference on ICT and Knowledge Engineering (ICT&KE). doi:10.1109/ictke.2017.8259629

Ozgur, A. (2004). Supervised and unsupervised machine learning techniques for text document categorization. Unpublished Master's Thesis, İstanbul: Boğaziçi University.

Pienaar, S. W., & Malekian, R. (2019). Human Activity Recognition using LSTM-RNN Deep Neural Network Architecture. 2019 IEEE 2nd Wireless Africa Conference (WAC). doi:10.1109/africa.2019.8843403

Posey, C., Raja, U., Crossler, R. E., & Burns, A. J. (2017). Taking stock of organisations' protection of privacy: categorising and assessing threats to personally identifiable information in the USA. European Journal of Information Systems, 26(6), 585–604. doi:10.1057/s41303-017-0065-y

Poyraz, O.I., Canan, M., McShane, M. et al. Cyber assets at risk: monetary impact of U.S. personally identifiable information mega data breaches. Geneva Pap Risk Insur Issues Pract 45, 616–638 (2020). https://doi.org/10.1057/s41288-020-00185-4

Tesfay, W. B., Serna, J., & Rannenberg, K. (2019). PrivacyBot: Detecting Privacy Sensitive Information in Unstructured Texts. 2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS). doi:10.1109/snams.2019.8931855

Tikkinen-Piri, C., Rohunen, A., & Markkula, J. (2018). EU General Data Protection Regulation: Changes and implications for personal data collecting companies. Computer Law & Security Review, 34(1), 134–153. doi:10.1016/j.clsr.2017.05.015

Whelan, M. (2014). Creating Test Data with Mockaroo. https://www.michael-whelan.net/creating-test-data-with-mockaroo/.

Xin, Y., Kong, L., Liu, Z., Chen, Y., Li, Y., Zhu, H., ... & Wang, C. (2018). Machine learning and deep learning methods for cybersecurity. IEEE Access, 6, 35365-35381.