Designing sensitive Data Detection and Anonymization Model Using BiLSTM for **Amharic Text**

Amare Genetu[*1],Tesfa Tegegne[2], Getasew Abeba[1], Tizazu Bayih[1]

[1]School of computing, Institute of Technology, Woldia university, Woldia, Ethiopia

[2]*ICT4DA Research Center, Bair Dar Institute of Technology, Bahir Dar University, Bahir Dar Ethiopia*

*Corresponding author Email: **amgenetu@gmail.com**

Abstract:

Sensitive information is a type of classified information that shouldn't be disclosed to the public since it may harm the information's owner. These days, a huge amount of information is going to be generated and shared through different platforms. Sensitive information may be disclosed when sharing such information. To protect disclose of sensitive information; applying detection and anonymization tools is a must. Recently it is proved that using the power of machine learning and Natural language it is possible to develop sensitive information detection and anonymization tools. However, such tool is strongly language dependent. And As of our literature review, there is no work attempted for Amharic texts. To address the aforementioned problems, we have proposed a model for detecting and anonymization personal sensitive information. For sensitivity detection use case BI-LSTM is used and it works better with 94% detection accuracy. For anonymization use case, a substitution approach is proposed. And it works accurately according to written substitution rules.

Keywords: Amharic text, Sensitive information, Sensitive information detection, anonymization.

I. Introduction

The amount of digital information produced is growing from time to time. The information content would be; generated by organizations, communicated by someone, stored from a cloud and any storage Medias, published on the Internet, shared, and used for research. In this process the sensitive personal information would be made public. To make the required body access to the information and protect the privacy of individuals at the same time, applying sensitive information anonymization tool is the best choice. The information that is going to be shared to the public may hold sensitive information of organizations, companies, and individuals to list some(Tesfay et al., 2019; Truong et al., 2020). Thus, sensitive information should be protected. Among privacy protection methods, data anonymization is a crucial method for sensitive data protection. The anonymization tool makes the owner of the sensitive information anonymous or un-identified(Goswami & Madan, 2017; Majeed &

Lee, 2021). These makes using sensitive personal information for different purpose like researching possible.

Data anonymization is the technique for altering the data before being shared or published so as to avoid the identification of sensitive attributes(Goswami & Madan, 2017). It is the practice of protecting sensitive information by erasing, di-identifying, or encrypting identifiers that link an individual to stored or shared data(Hassan et al., 2019). By preserving the information as it is, information anonymization methods allow us to make the owner of sensitive personal information anonymous. It is advisable to use a data anonymization method to make Personally Identifiable Information (PII) like names, social security numbers, and addresses anonymous, but the context or meaning of the information should not be changed. The General Data Protection Regulation (GDPR) defines a set of guidelines that protects user data. The GDPR does allow businesses to obtain anonymized data, use it for any purpose, and store it indefinitely as long as all identities are removed. To accomplish so, data Anonymization technologies are required. following a flow chart is presented to show how data Anonymization is done

.

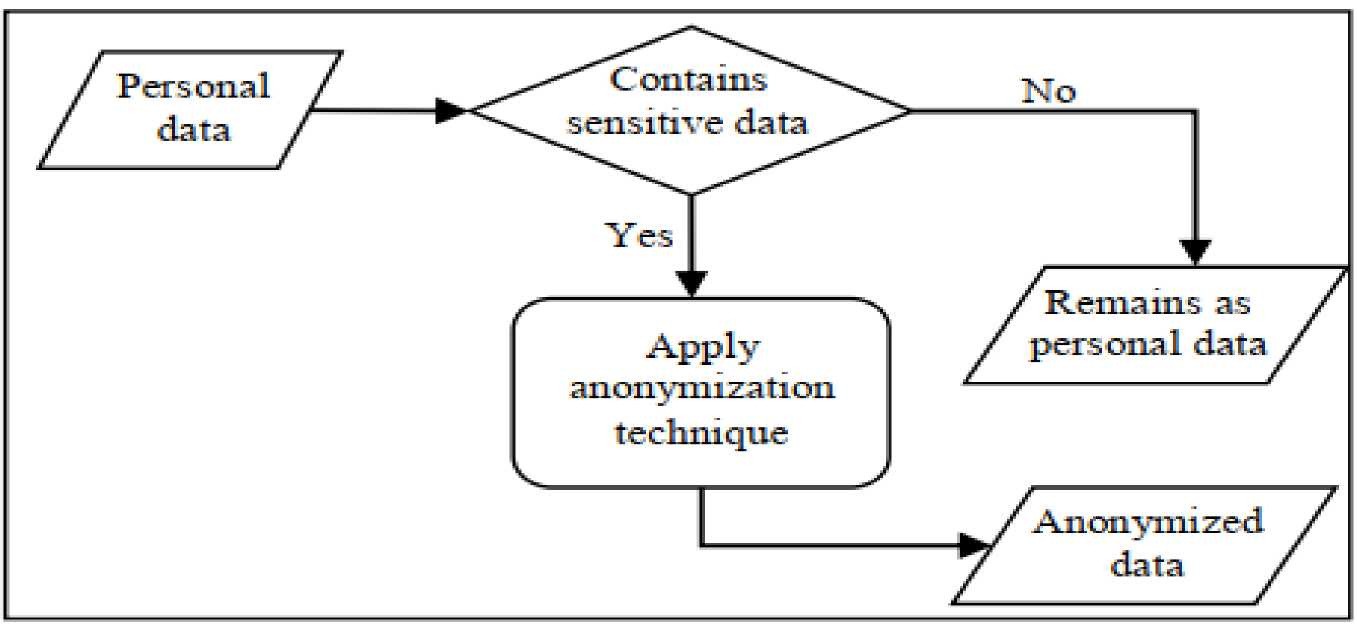

Figure 1: Summary of Data Anonymization concept

The development of text information anonymization tool is language dependent. A tool developed for English cannot work for our local languages like Amharic or Awi. In the works like (Dias et al., 2020; Neerbek et al., 2018; Trieu et al., 2018; Yang & Liang, 2018) it is proved that sensitive data detection and anonymization tools can be implemented using different approaches. However, natural language text processing tasks and structure is language-dependent, which makes it difficult to use sensitive information detector and anonymization tools developed for one language to another due to grammatical structure variation. In addition to this, Amharic is a morphologically rich language. This makes applying a tool that builds for another language to the Amharic language difficult. Manual sensitive information anonymization is expensive, tedious, and time consuming. . For our local language Amharic texts, no works are attempted yet even if the tool is highly important. The aforementioned problems and challenges are motivated us to propose this work

General data protection regulation defines personal information as any information pertaining to an identified natural person (GDPR)(Berhan Taye, 2018; Dove, 2018; S. Tovernic, Z. Hrastic, K. Plantic, A. Sandic, 2018; Yilma & Abraha, 2015). A huge amount of Amharic language texts, which contain sensitive information of individuals had and going to be generated from organization, and individuals. And in this work, we proposed data anonymization model for only this type of sensitive information. Actually, Sensitive information anonymization model development is not a single task. A tool which can detect sensitive information contents is needed, since anonymization is proposed to apply for sensitive information of individuals. So, sensitive information detection model is a must to develop anonymization model. Therefore, in this work we have proposed to do two tasks. First, we have developed sensitive information detection tool. Secondly, we have developed sensitive information anonymization model.

For sensitive information detection, we have applied Bidirectional Long Short-Term Memory (Bi-LSTM). Bi-LSTM networks are a subset of LSTM networks. Bi-LSTMs are made up of two different hidden layers. The first hidden layer processes the input sequence forward. The second hidden layer on the other side processes the sequence backward. This hidden layer enables the output layer to access the past and future background of each point in the series. The LSTM and its bidirectional variants proved to be extremely useful. They can learn how and when they can forget certain information and also, they can learn not to use some gateways in their architecture. Faster learning rate and better performance are the advantages of a Bi-LSTM network(Maslej-Krešňáková et al., 2020). To anonymize the sensitive contents detected we have used a substitution method to substitute the sensitive contents

I. Related Literatures

In the sensitive information detection and anonymization tools researching, many approaches were attempted by different researchers. For sensitive information detection; rule based, conventional machine learning, and deep learning approaches were employed. For the anonymization part commonly, substitution-based approaches are applied. Below some of the research works attempted for other language texts using the aforementioned approaches for sensitive information detection and anonymization are presented.

Sensitive information detection and anonymization tool researching had started a long time ago. The first research work had been attempted by using a set of a sensitive keyword as a feature and texts having one or more of those keywords detected as sensitive, and a text that does not hold any of the sensitive keywords categorized as non-sensitive (Pecherle et al., 2011). After a year natural language processing techniques like Named Entity Recognition had been attempted. As compared to the keyword matching based approaches, it had been performing well. Next to this, a combination of natural language and machine learning techniques have been used hand-crafted features (Alzhrani et al., 2016; Briand et al., 2018; Jose et al., 2017). Recently deep learning approaches is attempted and it provides state-of-the-art detection performance(Dias et al., 2020; García-Pablos et al., 2020; Geetha et al., 2020).

(Pecherle et al., 2011) have proposed a rule-based sensitive information detection model for detecting files containing sensitive information at data storage. Their proposed model detects a file as sensitive if it contains one or more of the defined keywords. However, this model can't consider the semantics or context of the keywords. In (Aldayel & Alhussain, 2017) a model for mobile applications has been proposed for mobile users to warn users of possibly privacy leaks in mobile applications by analyzing the sensitivity of user's input, that means keywords. The authors used privacy-related keywords using Natural Language Processing (NLP) methods as a feature for checking whether the user input is sensitive or not, however, it is also a keyword searching-based detection.

Another work (Xu et al., 2018) have proposed a sensitivity and information domain classification model for the Tibetan language. They have applied SVM with sensitive word vocabulary as a feature using cosine similarity measures by considering the location of sensitive words and they achieved good results, this work does not consider context and semantics. The authors of the work (Dias et al., 2020) have proposed a sensitive data detection task for Portuguese language texts. They have experimented with Bidirectional Long-Short Term Memory (Bi-LSTM). After all they have reported that Bi-LSTM have performed well with an accuracy of 83.01%. In the work (Xu et al., 2019), the authors apply Convolutional neural Network (CNN) for detecting sensitive information from unstructured texts to fill the training time gap of Recurrent Neural Network (RNN) based models for Chinese text. And finally, they have proved that their approach minimizes the training time of the model.

To extract or identify the part of the information to be anonymized two methods which are knowledge base (Dictionary) and Named Entity Recognition (NER) can be used. By developing language specific dictionary or knowledge base, the sensitive terms to be anonymized can be detected. The other method is using Named Entity Recognition. By developing NER tools, it is possible to tag and identify sensitive terms or sensitivity trigger words for anonymization. The authors in(Lee et al., 2017) have developed health data anonymization model that can make the health data of patients to privet privacy disclose. Their evaluation result showed that the proposed method provides good result. The other in have proposed anonymization framework for health data, and when this work compared with (Lee et al., 2017) it perform better with minimal information loss. The authors in (Ruch et al., 2000) designed a system for removing identifiers in medical records, with a success rate of about 99% using natural language processing tools to tag the Information to be anonymized. They have used replace operation to anonymized the extracted terms or words. The work in (Hassan et al., 2019) proposed semantic based anonymization models using word embedding techniques to detect sensitive contents from the given text document. The authors in this paper have been used Named Entity Recognition to extract sensitive terms. Their evaluation result shows that, their proposed model works with 67% recall.

I. Methodology

In this work, we use experimental research methodology. The proposed work have passed through the following phases; data acquisition, data annotation, preprocessing, Model development, and evaluation. The dataset was collected using three basic sources and techniques. First, we have collected from al-ain news (አል ዓይን ኒውስ) website, Ethio-WikiLeaks, and tweeter. Secondly, we have used other language sensitive information datasets by translating to Amharic language text which is published at GitHub for research. Thirdly, we have prepared an additional dataset by constructing Amharic sentences which is sensitive personal information using the vocabularies we have collected. To annotate the dataset, we have prepared annotation guidelines based on personal data protection regulation draft document of Ethiopia drafted by Ministry of Innovation and Technology (MInT). Having this guideline, dataset annotators and taggers were invited. After all the dataset was reviewed by MInT office workers.

As depicted from Figure 2, the proposed generalized architecture contains four components that are preprocessing, word representation, model development, and model testing. The input is the annotated dataset and it is going to be preprocessed using a text preprocessing algorithm. Next to this, the preprocessed dataset is converted to numeric vectors using word2vec word embedding technique to the dataset ready for the proposed model development. After all, we have built, train and evaluated the proposed mode l

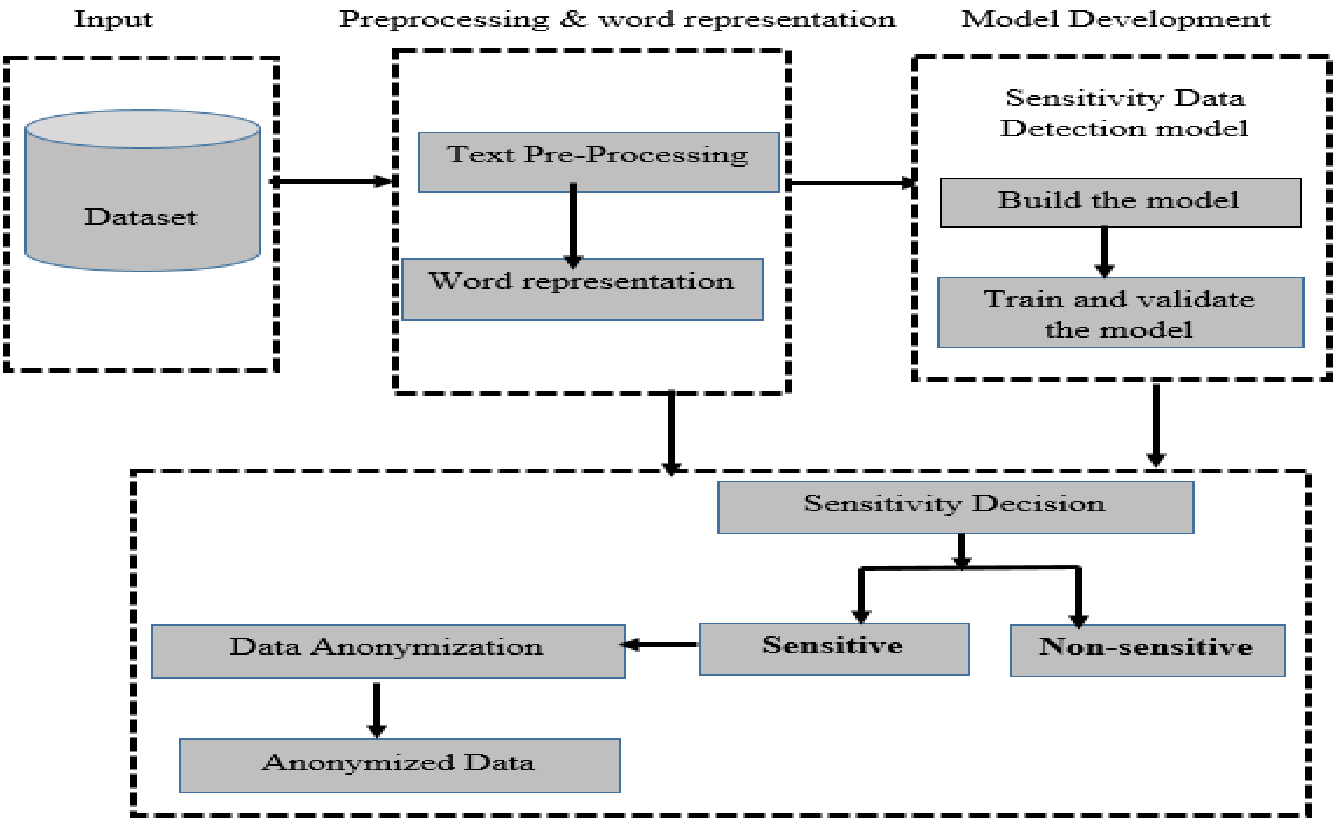

Figure 2: Proposed model

In the model development phase, for sensitive information detection have used we have used Bi-LSTM deep learning algorithm. Bi-LSTM is made up of two different hidden layers. The first hidden layer processes the input sequence forward. The second hidden layer on the other side processes the sequence backward. In this work, these two hidden layer capability of Bi-LSTM makes it better for capturing the context of words in the input text correctly. To extract the entities to be anonymized. we have used Named Entity tagging. On behalf of that asset of rules are implemented to substitute the sensitive terms or entities to be anonymized. For example all the name of persons in sensitive information assets were seated to be substituted by **"ግለሰቡ/ቧ".**

Result and discussion

I. Result and discussion

For sensitive information detection task, we have used 8K sentences. The dataset was divided into training, validation, and testing using 80:10:10 splitting ratio. For evaluating the performance of the anonymization model, we have used 1k personal sensitive information which is tagged by domain experts. The dataset contains sentences with a maximum length of 34 words and a minimum length of 5 words.

The hyper-parameter setups presented at Table 1 was selected for experimentation for the sensitive information detection part after trying many setups. The best setup which performs well was selected as our model hyper-parameter for sensitivity detection task

Table 1: Experimentation Hyper-parameter setups

No.	Hyper-parameter		Setup 1 (Experiment 1)	Setup 2 (Experiment 2)	Setup 3 (Experiment 3)
1.	Batch size	Training	128	64	64
		Validation	64	32	16
		Testing	64	32	16
2.	Learning rate		0.001	0.0001	0.0005
3.	Dropout		0.2	0.4	0.5
4.	Activation function		Sigmoid	Sigmoid	Sigmoid
5.	Epoch		15	25	10
6.	Optimizer		Adam	Nadam	SGD
7.	Sequence length		34	34	34
8.	Embedding dimension		64	128	200

The second hyperparameter setup with 64, 32, and 32 batch size for training, validation, and testing respectively, 0.0001 learning rate, 0.4 dropout, and 25 epoch performs better. At experimentation we understand that the aforementioned hyperparameter value makes the proposed to enhance training and prediction/testing performance. Batch sizes which have too small value makes the proposed model totake long training time and it creates noise for our model since providing too many batches to our model makes each batch a noisy representation of the whole dataset. Large Batch size makes the proposed model to have poor generalizing power and this makes the model performance lower. The intermediate batch size 64/32/32 makes the proposed model to perform well. Actually the 64/32/32 being the intimidate batch size depend the size of the dataset used, it may be too low for large dataset or too high for very small dataset. However, this work dataset is not either very small or very large, so we have concluded that 64/32/32 is an intermediate batch size for our case. Below at Table 2 the performance of the Bi-LSTM for sensitivity detection at the three experiments is presented

.

Table 2: Experimentation result

Experiment	Accuracy
Experiment 1(hyper-parameter 1)	88%
Experiment 2(hyper-parameter 2)	95%
Experiment 3(hyper-parameter 3)	85%

The learning rate with too small values forces the loss function not to be reduced to the optimal range. This too small learning rate makes the loos not reduced enough to enhance the performance of the model under developed. Due to this the 0.0001 learning rate value works well for this work. The significance of using dropout layer with a specified value is for controlling model overfitting. When we are not using a dropout layer and using small value of dropout the proposed model was overfit. For this work, 0.1-0.3 dropout values do not prevent overfitting correctly and due to this the accuracy of the model was reduced. However, 0.4 dropout value eliminates the proposed model overfitting problem. Similarly, too small and too large epochs have reduced the model accuracy. Too small epochs make the proposed model not to learn the training dataset enough and too large epochs makes the model to over learn. Both cases were reduced the accuracy of the model going developed. Below the training and validation loss and accuracy of Bi-LSTM at the selected hyper-parameter setup is presented

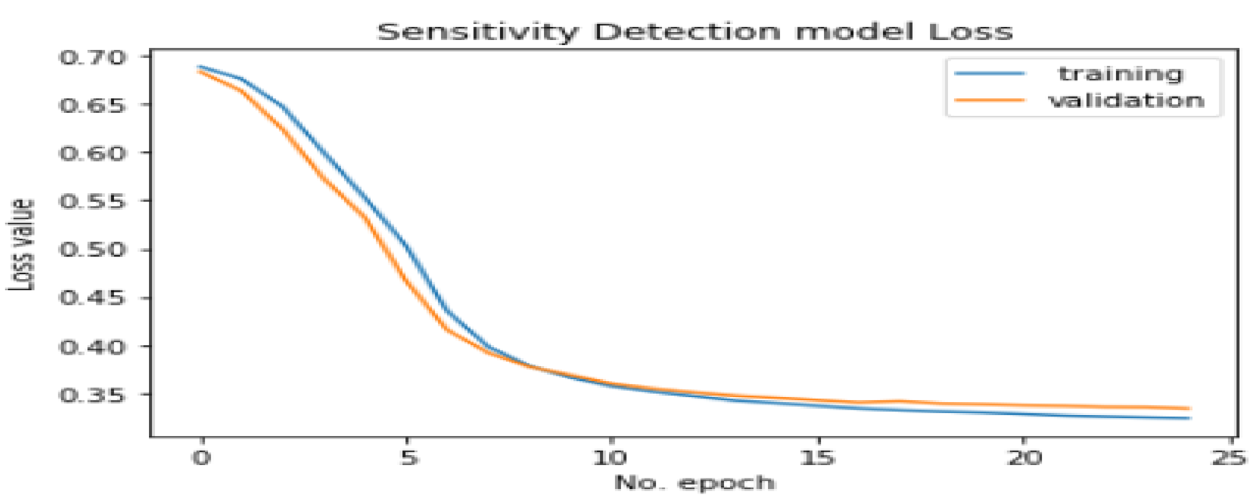

Figure 3: Sensitivity detection model training loss

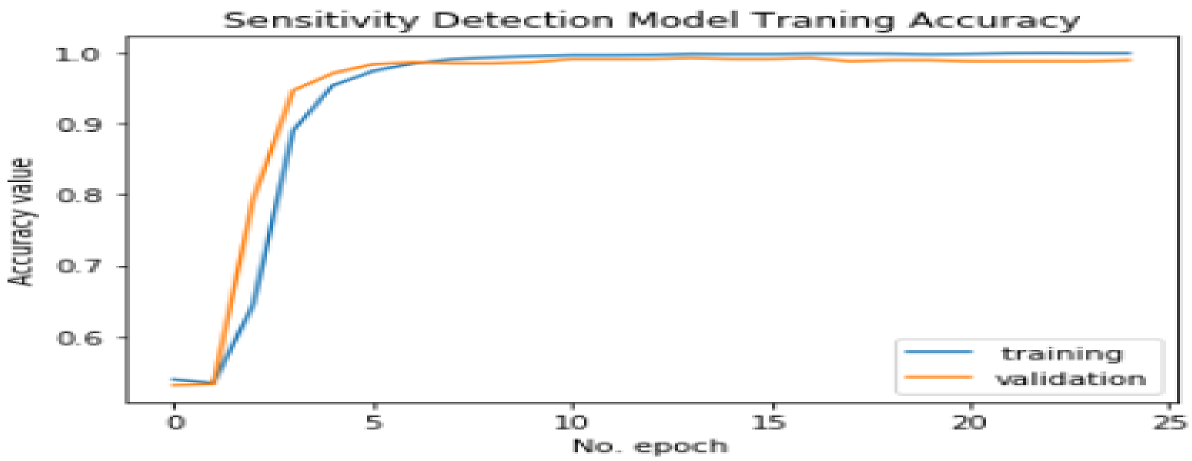

Figure 4: Sensitivity detection model training accuracy

As we can see from figures 3 and 4, the training and validation loss was high at the initial epochs, however, it has decreased as the epoch increases. When we see the training and validation accuracy, it has improved as the epoch increase. Overall figure 3 and 4 shows that the model learns the training dataset correctly as the learning iteration increases. When we see the performance of anonymization model, it works based on the sensitive content substitution rules and substitution terms or representations. Those we understood that anonymization errors or quality limitations are mostly comes from the NER tagging. However, the anonymization model works accurately according to written substitution rules. In the sensitive information detection part, the Bi-LSTM model provide us a better result with 95% detection accuracy. This is because of Bi-LSTMs are made up of two different hidden layers. The first hidden layer processes the input sequence forward. The second hidden layer on the other side processes the sequence backward. These two hidden layer capability of Bi-LSTM makes it better for capturing the context of words in the input sentences. Generally, we understood that Bi-LSTM is good for preserving word context.

I. Conclusion and recommendation

In this work, we have developed a model that can detect and anonymize sensitive information form Amharic text. The experimentation proves that Bi-LSTM Algorithm achieves best result (95% accuracy) for detecting sensitive information contents. For the anonymization part, rules we have developed provide us promising anonymization result. However, we understood that high quality NER is required to anonymize sensitive information using rule-based approaches. As a future work we recommend to study deep learning algorithms for automatic anonymization of sensitive information without the need of writing man coded rules.

Acknowledgement

Thanks are to the Almighty God who did everything in the first place. For he is God, the Creator of all things. Next, we would like to thank my collogues, they tirelessly devoted their precious time to give constructive comments for correcting both the technical and non-technical errors and misconceptions that had to occur through this research work. Secondly, we would like to thank the dataset annotators for giving their precious time to annotate this work dataset. Thirdly, we would like to thank Minister of Innovation and Technology worker, Mr. Lema misgan for sharing information protection regulations and giving time for reviewing the annotation guidelines and the dataset. Finally, we would to thank my family for their moral and financial support.

REFERENCES

Aldayel, M., & Alhussain, M. (2017). Enhanced identification of sensitive user inputs in mobile applications. *ICISSP 2017 - Proceedings of the 3rd International Conference on Information Systems Security and Privacy, 2017-Janua*(Icissp), 506–515. https: // doi.org /10.5220/0006238405060515

Alzhrani, K., Rudd, E. M., Boult, T. E., & Chow, C. E. (2016). Automated big text security classification. *IEEE International Conference on Intelligence and Security Informatics: Cybersecurity and Big Data, ISI 2016*, 103–108. https://doi.org/10.1109/ISI.2016.7745451

Berhan Taye, R. T. (2018). Privacy and Personal Data Protection in Ethiopia. *International ICT Policy for East and Southern Africa(CIPESA), 1*(September), 26.

Briand, A., Zacharie, S., Jean-Louis, L., & Meurs, M. J. (2018). Identification of sensitive content in data repositories to support personal information protection. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics): Vol. 10868 LNAI*. Springer International Publishing. https:// doi.org/10.1007 /978-3-319-92058-0_86

Dias, M., Boné, J., Ferreira, J. C., Ribeiro, R., & Maia, R. (2020). Named entity recognition for sensitive data discovery in portuguese. *Applied Sciences (Switzerland), 10*(7). https://doi.org /10.3390/app10072303

Dove, E. S. (2018). The EU General Data Protection Regulation: Implications for International Scientific Research in the Digital Era. *Journal of Law, Medicine and Ethics, 46*(4), 1013–1030. https://doi.org /10.1177 /1073110518822003

García-Pablos, A., Perez, N., & Cuadros, M. (2020). Sensitive data detection and classification in Spanish clinical text: Experiments with BERT. *LREC 2020 - 12th International Conference on Language Resources and Evaluation, Conference Proceedings*, 4486–4494. http://arxiv.org/abs

/2003.03106

Geetha, R., Karthika, S., & Mohanavalli, S. (2020). *Learning Approach to Predict Sensitive*. *December*. https://doi.org/10.1007/978-981-15-5558-9

Goswami, P., & Madan, S. (2017). Privacy preserving data publishing and data anonymization approaches: A review. *Proceeding - IEEE International Conference on Computing, Communication and Automation, ICCCA 2017*, *2017-Janua*, 139–142.https://doi.org/10.1109/CCAA.2017.8229787

Hassan, F., Sanchez, D., Soria-Comas, J., & Domingo-Ferrer, J. (2019). Automatic anonymization of textual documents: Detecting sensitive information via word embeddings. *Proceedings - 2019 18th IEEE International Conference on Trust, Security and Privacy in Computing and Communications/13th IEEE International Conference on Big Data Science and Engineering, TrustCom/BigDataSE 2019*, 358–365. https :// doi.org /10.1109 /Trust Com / BigDataSE.2019.00055

Jose, J. M., Hauff, C., Altıngövde, I. S., Song, D., Borlund, P., Kekalainen, J., Yimaz, E., Gaber, M., Kruschwitz, U., Russell-Rose, T., Albakour, D., Watt, S., He, D., Can, F., Sorumunen, E., Tait, J., & Goker, A. (2017). Advances in Information Retrieval. *39th European Conference on IR Research, ECIR 2017, Aberdeen, UK, April 8-13, 2017, Proceedings*, *10193 LNCS*, V–VI. https://doi.org/10.1007/978-3-319-56608-5

Lee, H., Kim, S., Kim, J. W., & Chung, Y. D. (2017). Utility-preserving anonymization for health data publishing. *BMC Medical Informatics and Decision Making*, *17*(1), 1–12. https://doi.org/10.1186/s12911-017-0499-0

Majeed, A., & Lee, S. (2021). Anonymization Techniques for Privacy Preserving Data Publishing: A Comprehensive Survey. *IEEE Access*, *9*, 8512–8545. https://doi.org /10.1109/ ACCESS.2020.3045700

Maslej-Krešňáková, V., Sarnovský, M., Butka, P., & Machová, K. (2020). Comparison of deep learning models and various text pre-processing techniques for the toxic comments classification. *Applied Sciences (Switzerland)*, *10*(23), 1–26. https:// doi.org / 10.3390 / app 10238631

Neerbek, J., Assent, I., & Dolog, P. (2018). Detecting Complex Sensitive Information via Phrase Structure in Recursive Neural Networks. *Springer International Publishing AG*, *2*, 1–13. https://doi.org/10.1007/978-3-319-93040-4

Pecherle, G., Gyorödi, C., Gyorödi, R., Andronic, B., & Ignat, I. (2011). New method of detection and wiping of sensitive information. *Proceedings - 2011 IEEE 7th International Conference on Intelligent Computer Communication and Processing, ICCP 2011*, 145–148. https:// doi.org / 10.1109/ ICCP.2011.6047859

Ruch, P., Baud, R. H., Rassinoux, A., Bouillon, P., & Robert, G. (2000). Infermsome Inf . Thatinoconcerin. *Medical Informatics*, 729–733.

S. Tovernic, Z. Hrastic, K. Plantic, A. Sandic, M. B. (2018). Solution for Detecting Sensitive Data inside a Data Lake. *MIPRO 2018*, 1284–1288.

Tesfay, W. B., Serna, J., & Rannenberg, K. (2019). PrivacyBot: Detecting Privacy Sensitive Information in Unstructured Texts. *2019 6th International Conference on Social Networks Analysis, Management and Security, SNAMS 2019*, 53–60. https:// doi.org /10.1109 / SNAMS.2019.8931855

Trieu, L. Q., Tran, T. N., Tran, M. K., & Tran, M. T. (2018). Document Sensitivity Classification for Data Leakage Prevention with Twitter-Based Document Embedding and Query Expansion. *Proceedings - 13th International Conference on Computational Intelligence and Security, CIS 2017*, *2018-Janua*, 537–542. https://doi.org/10.1109/CIS.2017.00125

Truong, A., Walters, A., & Goodsitt, J. (2020). Sensitive data detection with high-throughput neural network models for financial institutions. *ArXiv*.

Xu, G., Qi, C., Yu, H., Xu, S., Zhao, C., & Yuan, J. (2019). Detecting sensitive information of unstructured text using convolutional neural network. *Proceedings - 2019 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery, CyberC 2019*, 474–479. https:// doi.org / 10.1109 / Cyber C.2019.00087

Xu, G., Yu, Z., & Qi, Q. (2018). Efficient sensitive information classification and topic tracking based on tibetan web pages. *IEEE Access*, *6*, 1–10. https:// doi.org / 10.1109/ ACCESS.2018.2870122

Yang, Z., & Liang, Z. (2018). Automated identification of sensitive data from implicit user specification. *Cybersecurity*, *1*(1), 1–15. https://doi.org/10.1186/s42400-018-0011-x

Yilma, K., & Abraha, H. (2015). The Internet and Regulatory Responses in Ethiopia: Telecoms, Cybercrimes, Privacy, E-commerce, and the New Media. *Mizan Law Review*, *9*(1), 108. https://doi.org/10.4314/mlr.v9i1.4