Journal of Innovative Engineering and Research (JIER)
Vol.- 5, Issue - 2, October 2022, pp. 1-4 (4 pages)

Data Sensitive Security Model using Transfer Learning for Healthcare Data using Biomedical Natural Language Processing

SOMYA DUBEY AND DR. RAJEEV G. VISHWAKARMA
Department of Computer Science & Engineering, Dr. A. P. J. Abdul Kalam University, Indore
Corresponding Author Email: somydubey16y@gmail.com

Abstract— This paper proposes an intelligent healthcare data security architecture using transfer learning-based biomedical NLP. Digitalization and data collection from medical equipment in healthcare have complicated data analytics. This approach creates a lot of data, which is processed and categorized for an efficient analysis. This information is utilized to build medical prognostics and diagnoses. In healthcare, the security and privacy of so much data is a concern. Data security is uniform. The data's sensitivity depends on the attributes' relevance. Using a single data security architecture is wasteful and redundant. This paper provides an adaptive intelligent security approach based on material sensitivity. Using biomedical natural language processing, transfer learning was used to determine data sensitivity. The Bidirectional Encoder Representations from Transformers (BERT) model is trained on the MIMIC III clinical dataset to capture biomedical text semantics and sensitivity. This sensitivity level determines the data's security model. Based on ICD codes, data was separated into five sensitivity levels, and various MIMIC-III dataset and encryption method attributes were mapped.

Index Terms— Cybersecurity, cyber threats, healthcare, ransomware.

I. INTRODUCTION

Digital medical records changed the healthcare profession throughout the past decade. Electronic medical records have shifted diagnostic and research paradigms (EMR). The EMR's anywhere-and-anytime accessibility has allowed all pertinent data to be consolidated. Patient data includes basic medical history, clinical and doctor data, insurance information, etc. doctor. Parasite sources like remote sensors, biometric or genetic data, and social media data[1] also generate it. Medical equipment, digital health records, apprehending devices, and mobile computers have enhanced healthcare data collection. Researchers may generate new ideas and answer long-standing issues using collected data. Predictive analysis can save lives in emergency scenarios like coronavirus (COVID-19), H1N1, Dengue, and Ebola. This reduces treatment costs and improves efficiency. Big data in healthcare may help prevent epidemics, improve quality of life, speed up clinical trials, and detect drug side effects. Therapeutic decisions can also prevent Preventable Deaths [2].

Regulatory organizations standardized ICD coding to unify the medical profession. In the past decade, several scholars have been interested in ICD-9 disease classification. Larkey and Croft [3] created ICD-9 codes based on discharge reports. As an alternative to this rule-based classifier, which required medical specialists' input, larger datasets were more time-consuming. Learning-based techniques can determine a dataset's underlying distribution using simply learning algorithms. Deep learning techniques can successfully categorize EMR textual data. Julia Medori and Cedric Fairon [7] compared CNN against random forest, SVM, and logistic regression for ICD-9 code classification using radiology information. Researchers found deep learning outperformed an untuned CNN model. Choi et al. [8] predicted heart failure rates using RNNs (RNNs). Lipton et al. [9] classified occasional clinical data (LSTM). S. Baker et al. [10] classified EMR text based on the attention mechanism using biological natural language processing. These NLP algorithms did poorly with non-grammatical material like the healthcare dataset, which has bullet points, telegraphic phrases, and few entire words.

In recent years, various studies have attempted to address the safety and privacy of patient data. A data leak might endanger lives and cause a medical and economic calamity. Many healthcare organizations collect, preserve, and communicate medical data to provide care. This is the most sensitive data, making it a prime target for hackers. Academics have created alternative security models to prevent assaults and leaks [11–13]. [14-16] Throughout the security lifecycle, encryption-based security methods prevent unauthorized access to sensitive healthcare data. Researchers have a huge issue because the encryption key size varies with data. "Data masking" replaces PII in massive volumes of healthcare data. To retrieve relevant data, pseudo-identifiers are used. Swaney and Samrati [17,18] developed k-anonymity data masking to preserve healthcare data privacy. Truta et al. [19] incorporated p-sensitive anonymity to the security system's identification characteristics. These algorithms failed in anonymizing high-dimensional datasets.

Adaptive Security monitors behaviour and events to reduce risk and respond before an attack. Machine learning and sensitivity ratings are key to adaptive security. This research proposes an AI-based adaptive security strategy for the healthcare business using a transfer learning model to examine text-based healthcare data. The suggested framework leverages MIMIC-III to train the BERT AI model. ICD9 disease codes and other patient variables, such as medical history and insurance, determine data sensitivity.

The major contribution is an adjustable NLP-based biomedical security framework. Label embedding-based attention improves medical data categorization. The BERT

model is used in biological natural language processing to predict transfer learning-based sensitivity. The authors are unaware of any biomedical NLP-based healthcare data security techniques. Big data analytics and Transformer-based NLP help you deal with huge data and make smart decisions in noise and ambiguity. This hybrid technique enhanced model robustness without sacrificing complexity. The paper continues as follows: In section II, we'll examine the NLP transformer model. Section III offers a healthcare data security architecture. In part IV, a simulation study examines the suggested technique's performance.

II. TRANSFORMER MODEL DESCRIPTION

Many When it comes to learning, transfer learning has become a hot topic in recent years because it can be applied across multiple tasks without the requirement for fine-tuning a dataset. BERT, a novel language model based on natural language processing, was proposed by the Google team in 2018. Because typical transfer learning models rely on supervised fine tuning, the model's training process includes both unsupervised pretraining and supervised fine tuning. Using the autoregressive language encoder-only model, BERT's learning capabilities have been improved by preventing tokens from attending future tokens. The learnt weights from a base network trained with a large dataset can typically be applied to the target network by altering the training objectives over a smaller dataset. Transfer learning aims to increase learning characteristics by the use of the appropriate conditional probability distribution, in a target domain DT utilizing the knowledge attained from the source domain DS and Source task TS as the name suggests. The model takes text as its input and then learns from that content in order to produce the desired output text. It makes it possible for the model to apply the exact same model, hyperparameters, loss function, activation function, and other parameters to a variety of different NLP applications [20].
Distance between two arbitrary inputs or output positions increases the processing cost of convolutional neural network-based algorithms such as Extended Neural GPU, ByteNet, and ConvS2S [21]. Learning the dependencies between remote sites is exceedingly difficult because it changes linearly for ConvS2S and logarithmically for ByteNet. The solution to this difficulty was found in transformer design, which provided a resolution tradeoff and limited computation to a fixed number. An explicit location is included in the transformer design since self-attention is non-ordered. relative embedding is used instead of fixed embedding to provide the learned embedding with respect to the offset between "key" and "query." Transform the related logit in the transformer model and apply a scaler "embedding" to it to get the attention weights. All layers benefit from using the same location embedding parameters. The transformer model's layer normalization is positioned outside the residual path, which is a considerable deviation from the standard transformer model. The orthogonal modifications in the structure of the model also make it possible to remove the layer norm bias without severely affecting performance.

III. PROPOSED ADAPTIVE SECURITY MODEL USING TRANSFER LEARNING

Data categorization and security for healthcare data can be broken down into three phases, as shown in Figure 1, which includes preprocessing, training and a security method for data that is sensitive. The MIMIC-III dataset contains around 14,000 numerical codes for 52696 hospital admissions, which represent all possible diagnoses and procedures. " The data is first preprocessed through output labelling, lower-case word conversion, removal of special characters, separation of contractions, number canonization, and word tokenization, as the dataset includes various clinical notes incorporating radiology, nutrition, pharmacy, and insurance details. These preprocessing algorithms generate the discharge summary and related hint notes, which are vectorized collections of notes. When training the model, a new variable is added to the dataset to represent the sensitivity level of the discharge summary data.

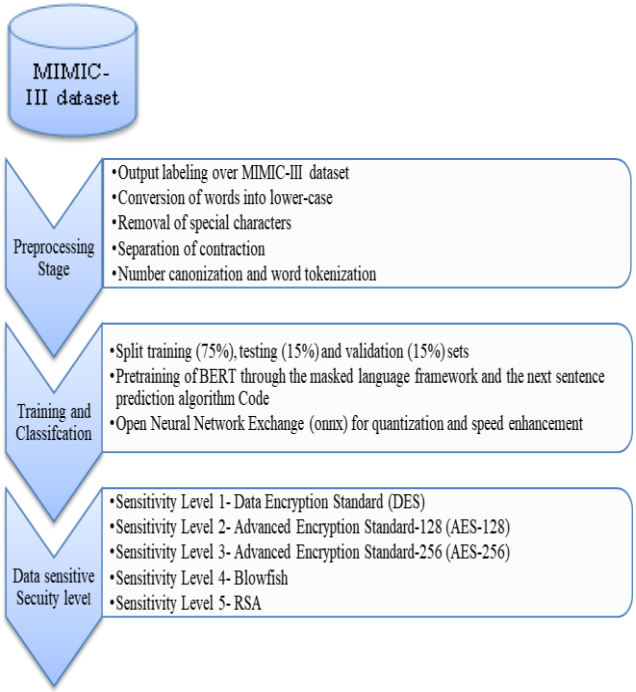

Figure 1. Flow graph of the proposed security framework

The next step of the suggested research is to train the transformer model for multi-level categorization. Training (75 percent), testing (15 percent), and validation were the three subgroups formed from the preprocessed and combined data (15 percent). Optimizing the cross-entropy function ensures good training results. To train the BERT transformer model, we employ a masked language framework and an algorithm for predicting the following sentence. The instructor pushing and maximum likelihood training principles are incorporated into the proposed strategy. To get the best inference speed, use onnxruntime, which quantizes the entire model before running it. Using the Open Neural Network Exchange (onnx), it quantizes the pre-trained model

and produces a model that can be run in a single line of code using the onnxruntime cross-platform inference framework. As a machine-learning accelerator, it allows hardware, drivers, and operating systems from many manufacturers to work together to improve performance. For the same precision, the quantized models have a lower latency when compared to the regular transformer models (due to graph optimization). The model is built using the TensorFlow and Keras frameworks, with L2 regularization and dropouts, to increase training efficiency. In this case, the Adam optimizer is utilized to improve training efficiency.

The transformer model's data sensitivity level is then utilized to determine the optimum security framework for the particular data. This research proposes an adaptive security framework that provides several levels of security to data with varying levels of sensitivity, ranging from level 1 to level 5. In this scenario, we use a high level of encryption for data with a high level of sensitivity and a lower level of protection for data with a lower level of sensitivity. Data Encryption Standard (DES), Advanced Encryption Standard-128 (AES-128), Advanced Encryption Standard-256 (AES-256), Blowfish, and RSA are the security levels employed in this article to correlate to each data sensitivity level [22].

IV. RESULT AND DISCUSSION

In the experimental investigation, we employed the MIMIC-III dataset to confirm the validity of our technique. Patients admitted to the Beth Israel Deaconess Medical Center's intensive care unit between 2001 and 2012 were assigned ICD-9 codes as labels. The pre-processed dataset contains 8,922 distinct codes for 52,724 discharge summaries. With the help of the PyTorch v1.5.0 and Pytorch-Lightning frameworks, and a python package called Transformersv3.0.2, the proposed transformer model and training techniques were constructed. Cross entropy loss and teacher training were employed for efficient training performance, whereas GPT2 was used in the decoder for efficient training. The encoder of the transformer is exported to the onnx model, which considerably reduces the complexity of the model and the amount of memory required. Multipartite graph-based keyword extraction and FlashText library are utilized to improve reasoning and understanding between words in the story to provide context awareness, while the PyTorch Lightning library provides a range of options for customization. At each time step the most likely hypotheses are preserved, and the hypothesis with the highest overall probability is selected. This model has been used to reduce the risk of hidden high probability word sequences.

In order to train and evaluate the suggested system, Google Cloud Compute Engine is used. An NVIDIA Tesla T4 virtual machine also existed. There is a 0.1 multiplier applied to all embeddings: token, segment, and position. In the original transformer, the dropout probability of the self-attention layer and all fully linked layers is preserved at 0.1. FastT5 model's Adam optimizer training rate is set to 0.001 per second. It is only possible to train the model for two epochs, each with an identical number of iterations, due to GPU memory constraints. When compared to the same reference, the precision is on par with the hypothesis's 4-gram precision. In order to evaluate the suggested classification model's performance, the F1 score was used. The comparison is depicted in table 1.

Table 1: Classification of MIMIC clinical notes

Source	Methods	No. of records	F1 score
Perotte et al.,[22]	Hierarchal SVM	22000	39.5
Liendo et.al. [23]	CNN	32000	76.2
Gehrmann et.al. [24]	LP 3-grams	16000	34.6
Proposed work	BERT Transformer over onnxruntime	46000	80.2

The performance of the proposed classification model is reflected from the improved F1 score in comparison with the SVM, CNN and LP 3-grams technique. The decision-making efficiency of the proposed framework is shown in fig.3. The 5 classes represent five encryption techniques corresponds to every medical data packet.

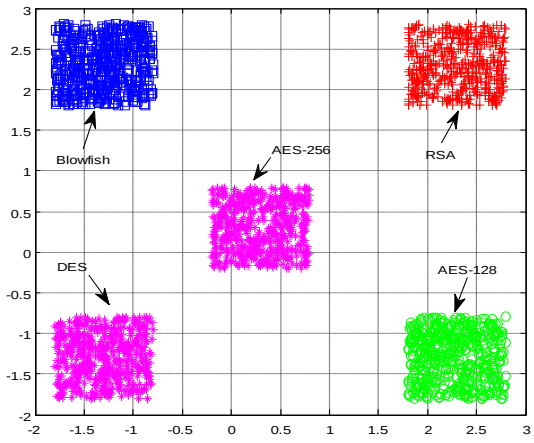

Figure 2. Encryption technique Allotment

V. CONCLUSION

Using the BERT Transformer model, researchers in this work suggest a data-sensitive adaptive intelligent security system. The MIMIC-III dataset, which comprises patient data, illness characteristics, symptoms, medical reports, and electronic medical records, is secured using the approach indicated for biomedical natural language processing. Data sensitivity levels range from 1 to 5, and the properties of healthcare data are categorised accordingly. The target security model in this study is a set of five different encryption methods, which can be used depending on the level of sensitivity. For each medical data packet, an intelligent model, trained on 31500 samples, chooses the most appropriate encryption scheme. In terms of performance evaluation, the proposed technique outperforms the standard categorization framework. The simulation results illustrate the superiority of the suggested

method over the conventional categorization framework. The computational complexity is decreased, and the system's efficiency is enhanced by using low-dimensional security for data with low sensitivity. In addition, the system's security is enhanced by employing the highest security technique for the most sensitive data. Deep learning models applied to a real-world dataset could improve the findings in the future.

REFERENCES

1. S. Sharathkumar and G. Jagadamba, "Adaptive content-aware access control of EPR resource in a healthcare system," 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI), Udupi, pp. 205-210, 2017.
2. A. Dev Mishra and Y. Beer Singh, "Big data analytics for security and privacy challenges," 2016 International Conference on Computing, Communication and Automation (ICCCA), Noida, pp. 50-53, 2016.
3. Leah S Larkey and W Bruce Croft. Automatic assignment of icd9 codes to discharge summaries. Technical report, Technical report, University of Massachusetts at Amherst, Amherst, MA, 1995.
4. John P Pestian, Christopher Brew, Paweł Matykiewicz, Dj J Hovermale, Neil Johnson, K Bretonnel Cohen, and Włodzisław Duch. A shared task involving multi-label classification of clinical free text. In Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing, pages 97–104. Association for Computational Linguistics, 2007.
5. Berthier Ribeiro-Neto, Alberto HF Laender, and Luciano RS De Lima. An experimental study in automatically categorizing medical documents. Journal of the Association for Information Science and Technology, 52(5):391– 401, 2001.
6. Stephane M Meystre, Guergana K Savova, Karin C Kipper-Schuler, John F Hurdle, et al. Extracting information ´ from textual documents in the electronic health record: a review of recent research. Yearb Med Inform, 35(128):44, 2008.
7. Julia Medori and Cedrick Fairon. Machine learning and features selection for semi-automatic icd-9-cm encoding. ´ In Proceedings of the NAACL HLT 2010 Second Louhi Workshop on Text and Data Mining of Health Documents, pages 84–89. Association for Computational Linguistics, 2010.
8. Edward Choi, Andy Schuetz, Walter F Stewart, and Jimeng Sun. Using recurrent neural network models for early detection of heart failure onset. Journal of the American Medical Informatics Association, page ocw112, 2016.
9. Zachary C Lipton, David C Kale, Charles Elkan, and Randall Wetzell. Learning to diagnose with LSTM recurrent neural networks. arXiv preprint arXiv:1511.03677, 2015.
10. S. Baker, A. Korhonen, Initializing neural networks for hierarchical multilabel text classification, in: BioNLP 2017, Association for Computational Linguistics, Vancouver, Canada, pp. 307–315, 2017.
11. Zhang R, Liu L. Security models and requirements for healthcare application clouds. In: IEEE 3rd international conference on cloud computing. 2010.
12. G. Ghinita, "Privacy for location-based services synthesis," Lectures on Information Security, Privacy, and Trust, University of Massachusetts, Boston, Tech. Rep., April 2013.
13. X. Liang, R. Lu, L. Chen, X. Lin, and X. Shen, "Pec: A privacy preserving emergency call scheme for mobile healthcare social networks," Communications and Networks, Journal of, vol. 13, no. 2, pp. 102–112, April 2011.
14. M. A. D. Mashima, D. Bauer and D. Blough, "User-centric identity management architecture using credential-holding identity agents," Digital Identity and Access Management: Technologies and Frameworks, IGI Global, December 2012.
15. F. Paci, R. Ferrini, A. Musci, K. Steuer, and E. Bertino, "An interoperable approach to multifactor identity verification," IEEE Computer, vol. 42, no. 5, pp. 50–57, 2009.
16. R. Lu, X. Lin, and X. Shen, "Spoc: A secure and privacy preserving opportunistic computing framework for mobile-healthcare emergency," Parallel and Distributed Systems, IEEE Transactions on, vol. 24, no. 3, pp. 614–624, March 2013.
17. Sweeney L. Achieving k-anonymity privacy protection using generalization and suppression. Int J Uncertain Fuzziness Knowl Based Syst. 10:571–88, 2002.
18. Samrati P. Protecting respondents' identities in microdata release. IEEE Trans Knowledge Data Eng. 13:1010–27, 2001.
19. Truta TM, Vinay B. Privacy protection: p-sensitive k-anonymity property. In: Proceedings of 22nd international conference on data engineering workshops. pp. 94, 2006.
20. J. Devlin et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *NAACL-HLT*. 2019.
21. G. Lample and A. Conneau. "Cross-lingual Language Model Pretraining". In: ArXiv abs/1901.07291, 2019.
22. M. Ali Shaik, Dhanraj Verma , Santosh Pawar, Ritesh Yadav, "Predicting Diabetes through Machine Learning, " Journal of Innovative Engineering and Research, Vol. 4, Issue 2, pp. 6-10, 2021.