applied sciences

Article

Filtered BERT: Similarity Filter-Based Augmentation with Bidirectional Transfer Learning for Protected Health Information Prediction in Clinical Documents

Min Kang [1], Kye Hwa Lee [2] and Youngho Lee [1,*]

[1] Department of Computer Engineering, Gachon University, Sungnam-si 13306, Korea; km8846@gachon.ac.kr
[2] Department of Information Medicine, Asan Medical Center, Seoul 05505, Korea; geffa@amc.seoul.kr
* Correspondence: lyh@gachon.ac.kr

Abstract: For the secondary use of clinical documents, it is necessary to de-identify protected health information (PHI) in documents. However, the difficulty lies in the fact that there are few publicly annotated PHI documents. To solve this problem, in this study, we propose a filtered bidirectional encoder representation from transformers (BERT)-based method that predicts a masked word and validates the word again through a similarity filter to construct augmented sentences. The proposed method effectively performs data augmentation. The results show that the augmentation method based on filtered BERT improved the performance of the model. This suggests that our method can effectively improve the performance of the model in the limited data environment.

Keywords: protected health information; named-entity recognition; transfer learning; augmentation

Citation: Kang, M.; Lee, K.H.; Lee, Y. Filtered BERT: Similarity Filter-Based Augmentation with Bidirectional Transfer Learning for Protected Health Information Prediction in Clinical Documents. *Appl. Sci.* **2021**, *11*, 3668. https://doi.org/10.3390/app11083668

Academic Editors: In Young Choi and Alessandro Di Nuovo

Received: 29 January 2021
Accepted: 16 April 2021
Published: 19 April 2021

1. Introduction

With the advent of the Fourth Industrial Revolution, the medical field is developing by responding most sensibly and rapidly to technological advances [1]. In particular, data analysis and artificial intelligence technology based on clinical medical data are attracting attention because they can be used as clinical decision support systems (CDSSs) that help experts make decisions [2]. A key component of clinical medical data is clinical documents, which are very important for medical data analysis because they contain information written by the clinician [3]. However, it is necessary to de-identify the personal information contained in the clinical document, that is, the protected health information (PHI), to maintain the confidentiality of the patient during secondary use, such as for research and data analysis of the clinical document. In the United States, the Health Insurance Portability and Accountability Act (HIPAA) defined guidelines for the secondary use of medical records, and guidelines for de-identifying medical records were defined accordingly.

Existing PHI removal and de-identification processes are performed manually. In this process, an annotator directly identifies and labels the PHI in a document. However, this method is expensive, and mistakes occur frequently when humans perform annotations directly.

The rule-based system was devised as an automatic annotation method to compensate for the problems of manual annotation. The rule-based system is generally implemented using regular expressions that are portable, fast, and easy to use, as they are standardized in most programming implementations. Shin et al. proposed a de-identification method using regular expressions in a free-text clinical document of an electronic medical record (EMR) in Korea, and obtained high recall and precision [4].

A machine learning system was proposed to solve the labor-intensive problem of the rule-based system and improve its generalization ability. Support vector machine (SVM) [5] and conditional random field (CRF) [6] have been used in previous studies as automatic PHI label identification methods. Although the SVM method is a classic machine learning method, it has been frequently used to identify PHIs. In particular, CRF has received attention because of its promising performance. Aramaki et al. proposed a PHI de-identification system using CRF [7]. Bin et al. proposed WI-deld, a de-identification algorithm based on the CRF algorithm, and achieved a high level of Micro F1-score [8]. However, a machine learning method that shows satisfactory performance requires detailed feature engineering for model tuning, and for this, it must be appropriately preprocessed.

Additionally, artificial neural network (ANN)-based deep learning methods are being actively considered as part of machine learning methods. This deep learning method has the advantage of automatically extracting features without the detailed feature engineering required in machine learning methods. In particular, models that use recurrent neural networks (RNNs) and long short-term memory (LSTM) [9] have attracted significant attention for their high performance, which has not been achieved before. Liu et al. achieved a high level of PHI entity identification performance using an LSTM model [10]. Yang et al. also proposed an anonymization method based on deep learning, using LSTM with a conditional random field [11].

However, a difficulty with these studies is that there are only a few appropriately annotated PHI public datasets. This poses challenges in the generalization stage [11,12]. Therefore, it is difficult to create large-scale open datasets, and it risks a patient's privacy. Therefore, technologies that can derive good results using a very small amount of data are required.

To overcome data limitations, data augmentation and transfer learning are mainly used in the traditional machine learning field. Data augmentation is a technique that increases the amount of training data by adding noise to existing data or by generating synthetic data based on the existing data. The field in which the data augmentation technique is most actively applied is the image field. Data augmentation techniques with promising performance have been introduced based on geometric transformation, cropping, rotation, and transformation to deep learning-based data enhancement [13,14]. Additionally, the data augmentation method was considered for time-series data, such as signal data [15]. Likewise, data augmentation was also considered in the natural language processing (NLP) field, where it was used to replace words with synonyms or for inserting and deleting random words, and showed effective and powerful performance improvement in limited data environments [16,17].

Additionally, pre-training [18] and transfer learning [19] have been actively considered as some of the methods to overcome limited data. In the image field, pre-learning has been used as a means to train a network on a large dataset such as ImageNet [20] and to solve other problems with pretrained weights. Furthermore, in the NLP field, pretrained embeddings such as Word2vec [21], GloVe [22], and fastText [23] demonstrate effective features. Embeddings from the language model (ELMO) [24] and bidirectional encoder representations from transformers (BERT) [25] are the most representative examples of using transfer learning in the NLP field. In particular, BERT is attracting attention because it supports fine-tuning and can be applied to various fields of NLP, which require a strong performance [26].

In this study, we propose a filtered BERT augmentation method to overcome limited data. This is to further improve the prediction performance by adding an appropriate augmentation that combines BERT and similarity filters to transfer learning to obtain limited data. We compared the performance of the existing BERT and its PHI prediction with the addition of the filtered BERT-based augmentation proposed using a representative public dataset.

2. Materials and Methods

Section 2 describes the filtered BERT proposed in this study. We first introduce the dataset used in this work in Section 2.1 and explain how data are properly preprocessed in the form of free text in Section 2.2. Section 2.3 describes our proposed filtered BERT augmentation method. Section 2.4 describes the process of fine-tuning BERT to recognize

PHI with datasets created through the augmentation method, and the evaluation metrics are presented in Section 2.5. Figure 1 shows the schematic data pipeline used in this study.

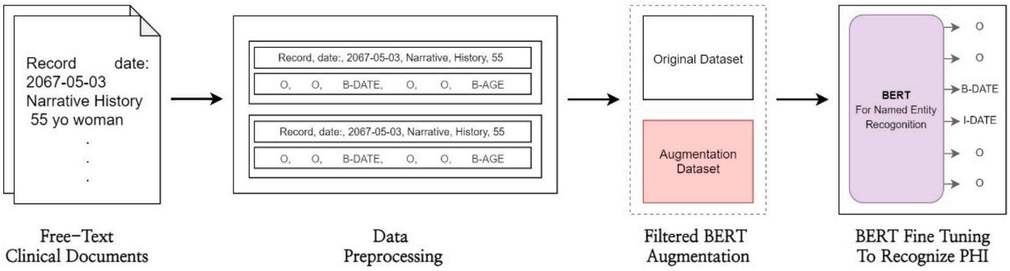

Record date:
2067-05-03
Narrative History
55 yo woman
Free−Text Clinical Documents
Record, date:, 2067-05-03, Narrative, History, 55
O, O, B-DATE, O, O, B-AGE
Record, date:, 2067-05-03, Narrative, History, 55
O, O, B-DATE, O, O, B-AGE
Data Preprocessing
Original Dataset
Augmentation Dataset
Filtered BERT Augmentation
BERT
For Named Entity Recogonition
O
O
B-DATE
I-DATE
O
O
BERT Fine Tuning To Recognize PHI

Figure 1. Filtered BERT augmentation data-pipeline.

2.1. Datasets

The i2b2 2014 dataset [27] was used in this study. It is one of the most representative datasets publicly available for PHI anonymization of medical documents. This dataset is a part of Track 1 of the 2014 i2b2/UTHealth natural language processing share task and consists of 1304 longitudinal medical records of 296 diabetic patients. All PHIs were anonymized by the organizer. Each PHI was annotated by three annotators and manually examined for annotation [28].

The dataset was annotated into the i2b2-PHI category, with a more expanded form than the HIPPA-PHI category. Table 1 lists the i2b2-PHI categories.

Table 1. i2b2-PHI categories.

Main Category	Subcategory
NAME	DOCTOR, PATIENT, USERNAME
PROFESSION	
LOCATION	HOSPITAL, COUNTRY, ORGANIZATION, ZIP, STREET, CITY, STATE, LOCATION-OTHER
AGE	
DATE	
CONTACT	PHONE, FAX, EMAIL, URL, IPADDR
ID	MEDICALRECORD, SSN, ACCOUNT, LICENSE, DEVICE, IDNUM, BIOID, HEALTHPLAN, VEHICLE

2.2. Data Preprocessing

We built a pretreatment pipeline for proper PHI recognition. First, tokenization was performed in units of words, and tokenized words were identified using the inside-outside-beginning (IOB) tagging scheme [29]. This method is the same as the existing method found in [10,11]. In the general BERT model, the input data are limited to 512 words. Therefore, each clinical note was divided into 250 words and used as input data. For example, if a clinical document contained 700 words, it was divided into three sets of data: 250, 250, and 200.

2.3. Data Augmentation

2.3.1. Filtered BERT for Augmentation Structure

We propose a filtered BERT method for effective data augmentation of clinical documents that include PHI, by applying a word-similarity-based filtering algorithm. The overall structure of the filtered BERT for the augmentation model is similar to that of the BERT-based augmentation method proposed in a previous study [30]. However, in the augmentation method that predicts the masked word through BERT, we added a filter that checked that word. The filter compares the similarity between the masked word and words in the original sentence. Thus, only words with a certain degree of similarity passed through the filter. The detailed structure of the proposed filtered BERT is shown in Figure 2.

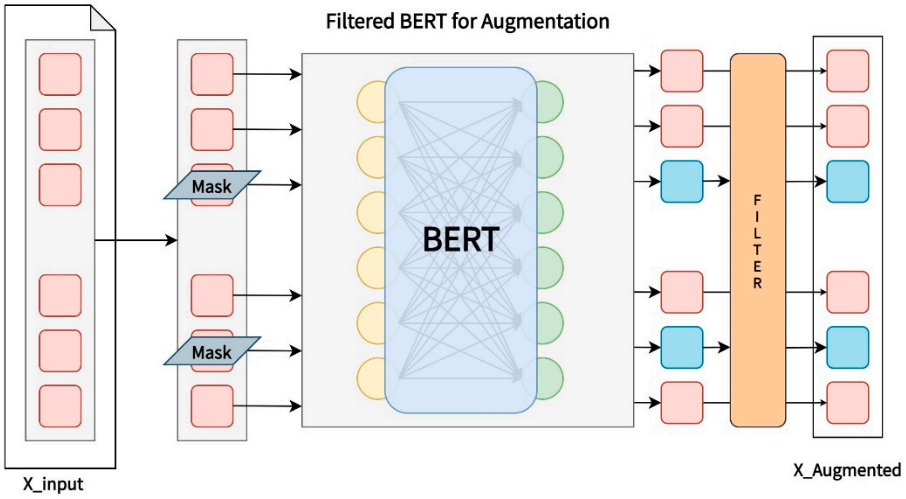
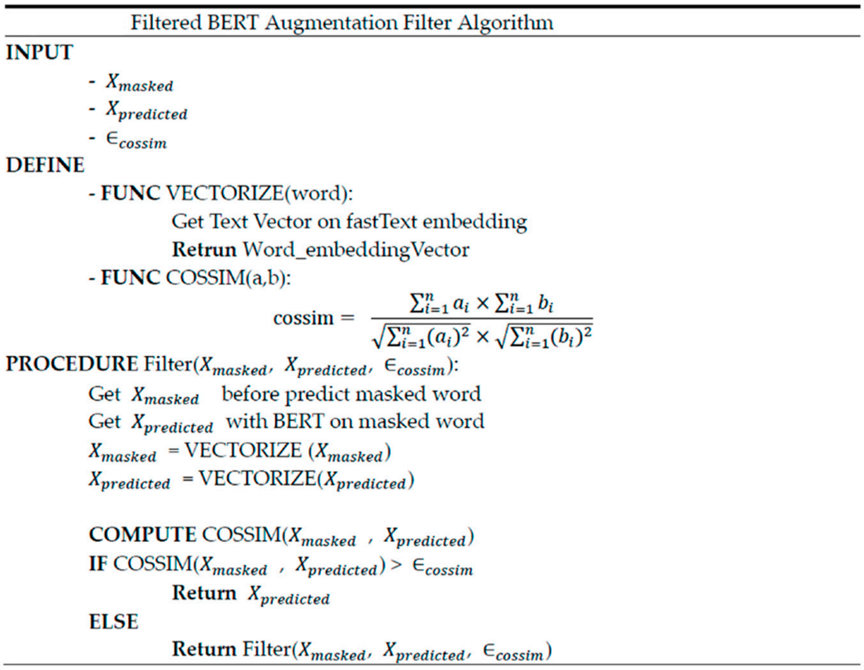


Figure 2. Detailed structure of filtered BERT augmentation algorithm.

The similarity filter was designed to check whether the words in the vector masked by the BERT model were similar to the original words. Figure 3 shows the detailed processing of the filter algorithm. To apply the BERT model, the words of the original sentence were masked ($\mathrm{X}_{\mathrm{masked}}$) and those predicted through context-based reasoning in the BERT ($\mathrm{X}_{\mathrm{predicted}}$) were converted into word vectors based on fastText [23] word embedding. Cosine similarity was calculated using the converted word vector, and word similarity was measured through this. The calculated cosine similarity had a value from -1 to 1, where -1 meant the masked and predicted words were different in meaning, and 1 meant they were the same. If the calculated cosine similarity was within the preset range, the predicted was finally returned to replace the masked word. Figure 2 shows the detailed processing of the filter algorithm. For example, if $\in_{\mathrm{cossim}}$ is set to 0, and the cosine similarity of $\mathrm{X}_{\mathrm{masked}}$ and $\mathrm{X}_{\mathrm{predicted}}$ is -0.7, it cannot pass the filter, and $\mathrm{X}_{\mathrm{predicted}}$ has to be predicted again. By contrast, if the cosine similarity of $\mathrm{X}_{\mathrm{masked}}$ and $\mathrm{X}_{\mathrm{predict}}$ is 0.5, it passes through the filter to form an augmentation sentence. These algorithms were implemented in Python.



```
Filtered BERT Augmentation Filter Algorithm
INPUT
        - $X_{masked}$
        - $X_{predicted}$
        - $\in_{cossim}$
DEFINE
        - FUNC VECTORIZE(word):
                Get Text Vector on fastText embedding
                Retrun Word_embeddingVector
        - FUNC COSSIM(a,b):
                $\text{cossim} = \frac{\sum_{i=1}^{n} a_i \times \sum_{i=1}^{n} b_i}{\sqrt{\sum_{i=1}^{n}(a_i)^2} \times \sqrt{\sum_{i=1}^{n}(b_i)^2}}$
PROCEDURE Filter($X_{masked}$, $X_{predicted}$, $\in_{cossim}$):
        Get $X_{masked}$ before predict masked word
        Get $X_{predicted}$ with BERT on masked word
        $X_{masked}$ = VECTORIZE ($X_{masked}$)
        $X_{predicted}$ = VECTORIZE($X_{predicted}$)

        COMPUTE COSSIM($X_{masked}$ , $X_{predicted}$)
        IF COSSIM($X_{masked}$ , $X_{predicted}$) > $\in_{cossim}$
                Return $X_{predicted}$
        ELSE
                Return Filter($X_{masked}$, $X_{predicted}$, $\in_{cossim}$)
```

Figure 3. Filter algorithm.

2.3.2. Filtered BERT-Based Clinical Documents Augmentation

We applied an optimized pre-training model for filtered BERT-based data augmentation. Bio-Clinical BERT [31], which was pretrained using bio and clinical data, was applied to the BERT model that predicted the masked word vector. It was confirmed in a previous study that the BERT model pretrained with a corpus suitable for the data showed good performance. To calculate the cosine similarity at the filter stage, fastText-based embedding was used when vectorization was performed by the embedding vector. In this case, fastText embedding, trained using the BioWordVec corpora [32], was used.

2.4. Named Entity Recognition with BERT

2.4.1. Tokenization and Labeling for the BERT Model

To train the BERT model, words were re-tokenized using Wordpiece [33]. Wordpiece tokenizes long words into multiple subparts. If a token was part of a precedent token, two marks (##) were attached to the front of the token to indicate its continuity. Additionally, two special tokens, CLS and SEP, were added to express the beginning and end of a sentence. To prevent the loss of IOB-labeled words during the re-tokenization process, the label of the tokenized word followed the IOB label before tokenization.

2.4.2. Fine-Tuning BERT

In this study, a PHI entity recognition model based on the i2b2 dataset was constructed to verify the performance of the proposed filtered BERT augmentation method. We used a fine-tuned pretrained BERT model that showed a good named entity recognition (NER) performance in a previous study [34]. The structure of the BERT model constructed in this study was a structure in which a token classification layer for entity name recognition was added to the pretrained BERT embedding. Figure 4 shows the BERT architecture and fine-tuning method.

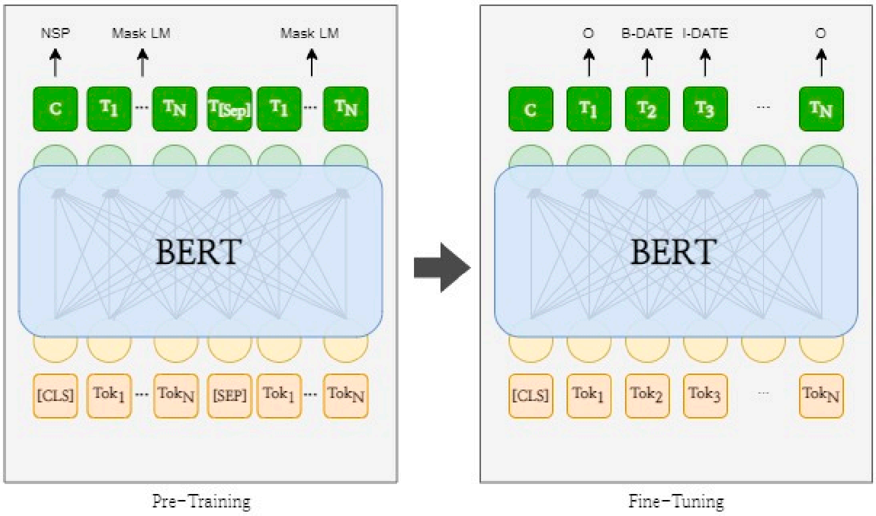


Figure 4. BERT architecture and fine-tuning process.

A PyTorch version of the publicly available BERT implementation was used in the experiment. A BERT parameter consisting of 12 layers, 768 hidden layers, and 12 heads was used. We used a pretrained bio-clinical BERT. The Adam Optimizer [35] was used for model training, and the learning rate was 3×10^{-5} (0.00003). Training was repeated for five epochs.

2.5. Evaluation

Various evaluation indices were used to evaluate the performance of the constructed model. We considered several performance evaluation indicators based on the accuracy of the confusion matrix [36], calculated as (correct prediction)/(total number of data). This is a method that is often used in general deep learning classification problems, but it is difficult to properly evaluate the performance of the model's data with severe data imbalance problems such as NER. Accordingly, we used (1) precision, (2) recall, and (3) F1-Score, which is obtained through recall and precision, to evaluate the performance of the model. True positive (*TP*) refers to the case where a PHI label is predicted as a PHI label, whereas a false positive (*FP*) refers to a case in which a normal label O is predicted as a PHI label. True negative (TN) is where a normal label O is predicted as a normal label, while false negative (*FN*) refers to the case where the PHI label is predicted as a normal label O. The F1-Score was obtained by calculating the harmonic average of recall and precision. This made it suitable for evaluating the performance of the model even when the data were unevenly distributed. Therefore, in this study, recall, precision, and F1-Score were used as evaluation indicators to ensure accurate model evaluation.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{1}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{2}$$

$$\text{F1 Score} = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{3}$$

3. Results

3.1. Data Augmentation

In this study, we created an augmentation text by transforming certain words in the original document using filtered BERT. Five words were substituted in all instances, and only those with a cosine similarity greater than 0 were passed through the filter and selected. Table 2 shows some examples of the original and augmented instances entered into the filtered BERT. It was confirmed that the word "denies" in the instance was replaced with "denied", a similar word (past tense).

Table 2. Examples of instances created as a result of filtered augmentation.

	Instance							
None	He O	has O	h/o O	drug/ETOG O	abuse O	but O	denies O	X O
Filtered BERT Augmentation	He O	has O	h/o O	drug/ETOG O	abuse O	but O	denied O	X O

We implemented data augmentation in the above form for all samples of training data; that is, the size of the augmented dataset was equal to that of the training data.

3.2. Results of Named Entity Recognition

In this study, we compared the results of the model when it was fine-tuned using an augmented instance, using only training data before augmentation, to the performance of the NER when the model was fine-tuned using the augmented instance through augmentation. Table 3 shows the results of the NER performance evaluation before and after the application of filtered BERT.

Table 3. Named entity recognition performance evaluation result.

	Precision	Recall	F1-Score
None	0.8811	0.8880	0.8845
Filtered BERT Augmentation	0.9265	0.9201	0.9233

To further elaborate on the results, we evaluated the detailed performance of each PHI tag, which is shown in Table 4. We evaluated the precision, recall, and F1-score for each label, and the support indicates the number of labels.

Table 4. Prediction performance by PHI entity before filtered BERT augmentation.

Tags	Precision		Recall		F1-Score		Support
	Before	After	Before	After	Before	After	
DOCTOR	0.73	0.88	0.76	0.90	0.75	0.89	2525
PATIENT	0.80	0.89	0.74	0.89	0.76	0.89	2275
USERNAME	0.95	0.98	0.85	0.91	0.90	0.95	167
PROFESSION	0.08	0.29	0.26	0.34	0.12	0.31	135
HOSPITAL	0.70	0.84	0.63	0.76	0.66	0.80	1665
COUNTRY	0.00	0.00	0.00	0.00	0.00	0.00	133
ORGANIZATION	0.00	0.05	0.00	0.17	0.00	0.07	88
ZIP	0.92	0.98	0.77	0.99	0.84	0.99	416
STREET	0.28	0.86	0.32	0.74	0.30	0.79	173
CITY	0.46	0.69	0.49	0.52	0.47	0.59	404
STATE	0.60	0.83	0.94	0.84	0.73	0.83	227
LOCATION-OTHER	0.00	0.00	0.00	0.00	0.00	0.00	16
AGE	0.92	0.95	0.86	0.96	0.89	0.95	621
DATE	0.98	0.99	0.98	0.98	0.98	0.98	13,024
PHONE	0.77	0.90	0.69	0.86	0.73	0.88	665
MEDICALRECORD	0.96	0.95	0.95	0.98	0.95.	0.97	2046
DEVICE	0.00	0.00	0.00	0.00	0.00	0.00	31
IDNUM	0.89	0.91	0.84	0.76	0.86	0.83	675

4. Discussion

For the secondary use of clinical documents, it is important to de-identify PHIs. Data augmentation and transfer learning methods can be used as effective methods to overcome the problem of a limited dataset. They can create a high-performance model when there is a problem with a small number of publicly available datasets for de-identification.

A major novelty of this study lies in the process of verifying the masked words predicted by BERT through the similarity filter during the data augmentation process. PHI is a set of individual data points, such as a person's name or registration number, which is difficult to augment with the existing medical data knowledge base, such as the unified medical language system (UMLS). Therefore, it is difficult to apply knowledge-based augmentation methods. Additionally, BERT effectively predicts a masked word through context-based inference, but there is a risk that the word may be predicted as a word with a completely different meaning without being verified. We filtered words with completely opposite meanings by verifying the words for a second time through the similarity filter based on fastText embedding, which was trained in advance.

When analyzing the results of the study and comparing BERT before and after augmentation, the overall performance showed significant improvements. Labels, including DOCTOR, PATIENT, and USERNAME, showed a performance improvement of more than 5%. Furthermore, in the case of labels with fewer classes such as PROFESSION, ORGANIZATION, and STREET, we could see an immense performance improvement. However, in the case of labels such as LOCATION-OTHER and DEVICE, it was observed that sufficient data for learning was not secured because the number of classes was too small.

Although the absolute number of datasets affects the occurrence of this problem, the fact that the configured dataset was unbalanced had a significant impact. The data augmentation method, which increases the absolute number of datasets, was effective in improving the overall performance; however, because the number of major classes is augmented together, there may be a limit to the learning of the minor classes.

The learning problem of unbalanced data was solved by the hybrid method of rule-based systems and machine learning in previous studies [8,11,37]. Additionally, an oversampling method, performed to resolve the unbalanced class, was also considered. This is because these methods are based on deep learning, such as the variational autoencoder (VAE), which is considered an effective solution to the unbalanced data problem [38]. If the oversampling method is applied in the future, it will be possible to present a model with superior performance.

The generalization ability of the proposed model can be considered as a limitation of this study. Since the training and testing data consisted of data collected from the same institution, applying it to other types of clinical documents may raise questions about the generalization of the model. Therefore, further evaluation using other organizations and other types of data sources is required.

Future work will focus on generalizing the results and the methodology, using more samples and samples of the same type of clinical documents data from other institutions.

Author Contributions: Conceptualization, M.K.; methodology, M.K.; project administration, Y.L.; software, M.K.; supervision, Y.L.; validation, K.H.L.; writing—original draft, M.K.; writing—review and editing, K.H.L. and Y.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Ministry of Science and ICT (MSIT), Korea, under the Information Technology Research Center support program (IITP-2020-2017-0-01630) supervised by the Institute for Information and Communications Technology Promotion (IITP).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: We used datasets released as part of the 2014 i2b2 challenge. Currently, it is part of the n2c2 dataset and can be used after requesting appropriate access from the DBMI portal, at https://portal.dbmi.hms.harvard.edu/ (Accessed on 1 January 2021).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Melo, C.; de Melo, J.A.G.; Araújo, N.M.F. Impact of the Fourth Industrial Revolution on the Health Sector: A Qualitative Study. *Healthc. Inform. Res.* **2020**, *26*, 328–334. [CrossRef]
2. Park, Y.T.; Kim, Y.S.; Yi, B.K.; Kim, S.M. Clinical Decision Support Functions and Digitalization of Clinical Documents of Electronic Medical Record Systems. *Healthc. Inform. Res.* **2019**, *25*, 115–123. [CrossRef]
3. Mujtaba, G.; Shuib, L.; Idris, N.; Hoo, W.L.; Raj, R.G.; Khowaja, K.; Shaikh, K.; Nweke, H.F. Clinical Text Classification Research Trends: Systematic Literature Review and Open Issues. *Expert Syst. Appl.* **2019**, *116*, 494–520. [CrossRef]
4. Shin, S.-Y.; Park, Y.R.; Shin, Y.; Choi, H.J.; Park, J.; Lyu, Y.; Lee, M.-S.; Choi, C.-M.; Kim, W.-S.; Lee, J.H. A De-Identification Method for Bilingual Clinical Texts of Various Note Types. *J. Korean Med. Sci.* **2015**, *30*, 7–15. [CrossRef]
5. Corinna, C.; Vapnik, V. Support-Vector Networks. *Mach. Learn.* **1995**, *20*, 273–297.
6. Lafferty, J.; McCallum, A.; Pereira, F.C. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of the Eighteenth International Conference on Machine Learning*; Morgan Kaufmann Publishers Inc.: Burlington, MA, USA, 2001; pp. 282–289.
7. Aramaki, E.; Imai, T.; Miyo, K.; Ohe, K. Automatic Deidentification by Using Sentence Features and Label Consistency. i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data. 2006, Volume 2006, pp. 10–11. Available online: http://luululu.com/paper/2006-i2b2/i2b2-deid.pdf (accessed on 29 January 2021).
8. He, B.; Guan, Y.; Cheng, J.; Cen, K.; Hua, W. CRFS Based De-Identification of Medical Records. *J. Biomed. Inform.* **2015**, *58*, S39–S46. [CrossRef]
9. Hochreiter, S. Long Short-Term Memory. *J. Neural Comput. Schmidhuber* **1997**, *9*, 1735–1780. [CrossRef]
10. Liu, Z.; Yang, M.; Wang, X.; Chen, Q.; Tang, B.; Wang, Z.; Xu, H. Entity Recognition from Clinical Texts Via Recurrent Neural Network. *BMC Med Inform. Decis. Mak.* **2017**, *17*, 67. [CrossRef]
11. Yang, X.; Lyu, T.; Li, Q.; Lee, C.Y.; Bian, J.; Hogan, W.R.; Wu, Y. A Study of Deep Learning Methods for De-Identification of Clinical Notes in Cross-Institute Settings. *BMC Med Inform. Decis. Mak.* **2019**, *19*, 232. [CrossRef]
12. Yue, X.; Zhou, S. Phicon: Improving Generalization of Clinical Text De-Identification Models Via Data Augmentation. *arXiv* **2020**, arXiv:2010.05143.
13. Shorten, C.; Khoshgoftaar, T.M. A Survey on Image Data Augmentation for Deep Learning. *J. Big Data* **2019**, *6*, 60. [CrossRef]
14. Mikołajczyk, A.; Grochowski, M. Data Augmentation for Improving Deep Learning in Image Classification Problem. In Proceedings of the 2018 International Interdisciplinary PhD Workshop (IIPhDW), Świnoujście, Poland, 9–12 May 2018.

15. Um, T.T.; Pfister, F.M.; Pichler, D.; Endo, S.; Lang, M.; Hirche, S.; Fietzek, U.; Kulić, D. Data Augmentation of Wearable Sensor Data for Parkinson's Disease Monitoring Using Convolutional Neural Networks. *ICMI* **2017**, *17*, 216–220.
16. Kobayashi, S. Contextual Augmentation: Data Augmentation by Words with Paradigmatic Relations. *arXiv* **2018**, arXiv:1805.06201.
17. Wei, J.; Zou, K. Eda: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks. *arXiv* **2019**, arXiv:1901.11196.
18. Erhan, D.; Courville, A.; Bengio, Y.; Vincent, P. Why Does Unsupervised Pre-Training Help Deep Learning? In Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, Sardinia, Italy, 13–15 May 2010.
19. Shao, L.; Zhu, F.; Li, X. Transfer Learning for Visual Categorization: A Survey. *IEEE Trans. Neural Netw. Learn. Syst.* **2014**, *26*, 1019–1034. [CrossRef] [PubMed]
20. Deng, J.W.; Dong, R.; Socher, L.; Li, L.K.; Li, F.F. Imagenet: A Large-Scale Hierarchical Image Database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009.
21. Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.; Dean, J. Distributed Representations of Words and Phrases and Their Compositionality. *J. Adv. Neural Inf. Process. Syst.* **2013**, *26*, 3111–3119.
22. Pennington, J.; Richard, S.; Manning, C.D. Glove: Global Vectors for Word Representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014.
23. Joulin, A.; Edouard, G.; Piotr, B.; Matthijs, D.; Hérve, J.; Mikolov, T. Fasttext. Zip: Compressing Text Classification Models. *arXiv* **2016**, arXiv:1612.03651.
24. Peters, M.E.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; Zettlemoyer, L. Deep Contextualized Word Representations. *arXiv* **2018**, arXiv:1802.05365.
25. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-Training of Deep Bidirectional Transformers for Language Understanding. *arXiv* **2018**, arXiv:1810.04805.
26. Yang, S.; Yoo, S.; Jeong, O. Denert-Kg: Named Entity and Relation Extraction Model Using Dqn, Knowledge Graph, and Bert. *Appl. Sci.* **2020**, *10*, 6429. [CrossRef]
27. Stubbs, A.; Christopher, K.; Uzuner, Ö. Automated Systems for the De-Identification of Longitudinal Clinical Narratives: Overview of 2014 I2b2/Uthealth Shared Task Track 1. *J. Biomed. Inform.* **2015**, *58*, S11–S19. [CrossRef]
28. Stubbs, A.; Uzuner, Ö. Annotating Longitudinal Clinical Narratives for De-Identification: The 2014 I2b2/Uthealth Corpus. *J. Biomed. Inform.* **2015**, *58*, S20–S29. [CrossRef]
29. Sang, E.F.; De Meulder, F. Introduction to the Conll-2003 Shared Task: Language-Independent Named Entity Recognition. *arXiv* **2003**, arXiv:cs/0306050.
30. Kumar, V.; Choudhary, A.; Cho, E. Data Augmentation Using Pre-Trained Transformer Models. *arXiv* **2020**, arXiv:2003.02245.
31. Alsentzer, E.; Murphy, J.R.; Boag, W.; Weng, W.H.; Jin, D.; Naumann, T.; McDermott, M. Publicly Available Clinical Bert Embeddings. *arXiv* **2019**, arXiv:1904.03323.
32. Zhang, Y.; Chen, Q.; Yang, Z.; Lin, H.; Lu, Z. Biowordvec, improving Biomedical Word Embeddings with Subword Information and Mesh. *Sci. Data* **2019**, *6*, 52. [CrossRef]
33. Wu, Y.; Schuster, M.; Chen, Z.; Le, Q.V.; Norouzi, M.; Macherey, W.; Krikun, M.; Cao, Y.; Gao, Q.; Dean, J. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *arXiv* **2016**, arXiv:1609.08144.
34. Kim, Y.-M.; Lee, T.-H. Korean Clinical Entity Recognition from Diagnosis Text Using Bert. *BMC Med Inform. Decis. Mak.* **2020**, *20*, 242. [CrossRef]
35. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. *arXiv* **2014**, arXiv:1412.6980.
36. Ting, K.M. Confusion Matrix. In *Encyclopedia of Machine Learning and Data Mining*; Claude, S., Webb, G.I., Eds.; Springer: Boston, MA, USA, 2017; p. 260.
37. Liu, Z.; Chen, Y.; Tang, B.; Wang, X.; Chen, Q.; Li, H.; Wang, J.; Deng, Q.; Zhu, S. Automatic De-Identification of Electronic Medical Records Using Token-Level and Character-Level Conditional Random Fields. *J. Biomed. Inform.* **2015**, *58*, S47–S52. [CrossRef]
38. Park, J.H.; Baek, J.H.; Sym, S.J.; Lee, K.Y.; Lee, Y. A Data-Driven Approach to a Chemotherapy Recommendation Model Based on Deep Learning for Patients with Colorectal Cancer in Korea. *BMC Med. Inform. Decis. Mak.* **2020**, *20*, 241. [CrossRef] [PubMed]