Journal of the American Medical Informatics Association, 27(6), 2020, 901–907
doi: 10.1093/jamia/ocaa038
Advance Access Publication Date: 10 May 2020
Research and Applications

Research and Applications

Using word embeddings to improve the privacy of clinical notes

Mohamed Abdalla[1,2,3], Moustafa Abdalla[4,5,6], Frank Rudzicz[2,3,7], and Graeme Hirst[2,3]

[1]ICES, Toronto, Canada, [2]The Vector Institute for Artificial Intelligence, Toronto, Canada, [3]Department of Computer Science, University of Toronto, Toronto, Canada, [4]Computational Statistics & Machine Learning Group, Department of Statistics, University of Oxford, Oxford, UK, [5]Wellcome Centre for Human Genetics, Nuffield Department of Medicine, University of Oxford, Oxford, UK, [6]Harvard Medical School, Harvard University, Boston, USA, and [7]International Centre for Surgical Safety, Li Ka Shing Knowledge Institute, St Michael's Hospital, Toronto, Canada

***Corresponding Author:** Mohamed Abdalla, Department of Computer Science, University of Toronto, Vector Institute, 661 University Ave Suite 710, M5G 1M1 Toronto, Canada (mohamed.abdalla@mail.utoronto.ca)

Received 24 January 2020; Revised 10 March 2020; Editorial Decision 18 March 2020; Accepted 23 March 2020

ABSTRACT

Objective: In this work, we introduce a privacy technique for anonymizing clinical notes that guarantees all private health information is secured (including sensitive data, such as family history, that are not adequately covered by current techniques).

Materials and Methods: We employ a new "random replacement" paradigm (replacing each token in clinical notes with neighboring word vectors from the embedding space) to achieve 100% recall on the removal of sensitive information, unachievable with current "search-and-secure" paradigms. We demonstrate the utility of this paradigm on multiple corpora in a diverse set of classification tasks.

Results: We empirically evaluate the effect of our anonymization technique both on upstream and downstream natural language processing tasks to show that our perturbations, while increasing security (ie, achieving 100% recall on any dataset), do not greatly impact the results of end-to-end machine learning approaches.

Discussion: As long as current approaches utilize precision and recall to evaluate deidentification algorithms, there will remain a risk of overlooking sensitive information. Inspired by differential privacy, we sought to make it statistically infeasible to recreate the original data, although at the cost of readability. We hope that the work will serve as a catalyst to further research into alternative deidentification methods that can address current weaknesses.

Conclusion: Our proposed technique can secure clinical texts at a low cost and extremely high recall with a readability trade-off while remaining useful for natural language processing classification tasks. We hope that our work can be used by risk-averse data holders to release clinical texts to researchers.

Key words: privacy, data anonymization, natural language processing, personal health records

INTRODUCTION

Natural language processing (NLP) is being increasingly applied to free-text clinical notes, both to improve quality-of-care and to understand the pathophysiologies and natural histories of disease. Clinical notes are rich in information and carry observations that cannot be easily conveyed using structured data, but their use in NLP research is hampered by stringent security requirements, which in turn affects researcher access to data.

Healthcare providers are often hesitant to about (or prohibited from) sharing unstructured data with external researchers because of the difficulties associated with securing sensitive information. The

obligation to secure private patient information is enforced by legislation such as the United States Health Insurance Portability and Accountability Act (HIPAA), which defines a list of personal health information (PHI) that must be protected, including names, dates, and other unique identifying characteristics.

The need for automated intelligent tools to detect and secure sensitive personal information is growing as many NLP tasks with clinical notes involve training end-to-end systems requiring a large amount of data. Deidentifying notes automatically with NLP will increase the availability of these data for research. Many studies use NLP techniques to assist with the anonymization of clinical notes.[1–3] Existing techniques in the clinical privacy literature purport to remove sensitive data (such as names) with accuracies between 90%[3] and 99%,[2] although this is often measured on small test sets and the techniques are less effective on larger, noisier datasets.

In this work, we present a novel technique for masking all sensitive information in text that uses the semantic properties of word embeddings (namely, the correlation between proximity in the embedding space and semantic relevance). This unique paradigm outperforms all existing methods with respect to recall without suffering a large decrease in performance for machine learning (ML) classification tasks trained end-to-end. In contrast to a prior method by Fernandes et al[4] that also used word embeddings, we guarantee that each name will be changed. This previous method added a static amount of noise, drawn from a Gaussian distribution, to each token's vector representation before replacing the original token with the token closest to the newly noised vector. However, this approach is not secure, as it assumes equal spacing within the embedding space, which is empirically false.[5,6] With a fixed ϵ, certain tokens in sparser areas would not get replaced as frequently as other tokens, which presents a security risk. By contrast, our approach, designed with the clinical use-case in mind, forces each token to be replaced with a semantically proximate token chosen at random. That is, our technique protects more data with a reduction of less than 5% in F_1 score on average for multiple clinical tasks on the most secure obfuscation setting.

Since no perfect search algorithm exists, past deidentification approaches necessarily overlook some sensitive information. As long as researchers use paradigms that can be evaluated using precision and recall, we are risking patient privacy. Novel methods to protect privacy, belonging to new paradigms, are needed to increase the amount of data available to researchers. We present 1 such possible method. By trading human readability for security, we are able to provide 100% coverage on all names and all other identifying sensitive textual information. We evaluate the effect of our technique on the most common upstream and downstream clinical NLP tasks (ie, word similarity and disease code prediction).

BACKGROUND

Detecting PHI

The large body of literature dedicated to securing PHI in clinical notes can be grouped under 3 high-level approaches: i) dictionary-based methods, ii) statistical methods, and iii) hybrid methods.

Dictionary-based methods use a compiled list (ie, a dictionary) or predefined regular expressions to identify all occurrences of private attributes present in the data.[7,8] Dictionary-based approaches often make use of misspelling detectors; for example, Thomas et al, using a list of 1.8 million names and a misspelling detector, reported correctly identifying 98.7% of names in a test dataset.[8]

Statistical approaches, which are more robust to unseen data than dictionary-based methods, can be further classified into classical and neural approaches.[1,2,9] Regardless of the specific classification, these techniques rely on rules, whether explicitly programmed (eg, RegEx, POS tagging) or implicitly learned from the text. Certain works[9] make use of extracted syntactic and lexical features achieving recall over 90%, while others make use of no hand-coded features and a recurrent neural network to achieve recall of 99% on the MIMIC dataset.[2]

The most common approach to removing private attributes from clinical notes is to use a mixture of dictionary-based and statistical approaches.[3,10] The best of these approaches claims to detect sensitive information with recall and precision in the high 90s. Combining dictionary-based and statistical approaches allows each technique to compensate for the weaknesses of the other, and often win shared deidentification task challenges. Methods, such as that of Liu et al,[10] combined an long short-term memory-based model with a rule-based model and achieved an F_1 score in the mid-90%s, ranking first in the 2016 CEGS N-GRID NLP challenge.[11] Combining a keyword dictionary, a rule-based system, and a statistical ML method, Yang et al[12] achieved an overall micro-averaged F-measure of 93.6%, winning the 2014 i2b2 challenge.[13]

Anonymization

Once PHI has been detected using any of the above methods, data holders can either secure the data through deletion (ie, PHI removal) or random replacement of other information of the same type (ie, PHI replacement).

PHI removal is simple as clinical notes remain readable and there is a minimal information loss, which is a critical concern to most people working with clinical notes.[14] However, since no perfect search algorithm exists, sensitive data missed by PHI removal can be found by combing through the data for names and other personal information that were not removed.

PHI replacement is a more secure approach to deidentifying clinical records, as it is no longer clear which names are true and which have been randomly replaced. Unfortunately, this approach remains susceptible to attack by malicious individuals depending on the specifics of the replacement.[15] To do this, a malicious party would look at instances of notes where there exist multiple differing names and leverage external world knowledge to deduce real names.[15]

MATERIALS AND METHODS

Past approaches framed deidentification as search followed by removal or replacement. Since no perfect search algorithm exists, these approaches necessarily overlook some sensitive information. Our approach, Algorithm 1, is to replace every token in the clinical note with a random related token. By relying on the semantic properties of word embeddings, we can refactor the text to have the same properties for downstream tasks as the original text.

Word embedding algorithms are techniques that represent word tokens as dense numeric vectors. Most existing word embedding techniques rely on the distributional hypothesis[16] and, as a result, tokens that appear in similar contexts become closer to each other in the numeric vector space. Our technique leverages this semantic property by randomly selecting a token from a subset of tokens that appear in the same context. Specifically, we replace every token in each clinical note (in place within the note, in a one-off manner) with a random token from the closest N neighboring tokens in the

embedding space (excluding itself). We refer to N as the *degree of obfuscation*; the larger it is, the more obfuscated the text becomes. As we are replacing every token, we are able to achieve perfect recall, although with a decrease in readability. In the discussion, we will propose a means of ensuring perfect recall while improving readability.

This 100% replacement gives data holders complete confidence that the anonymized notes that are produced are completely secure in the sense that they do not exist in the original dataset at all. Furthermore, as the tokens are selected randomly from neighbors that appear frequently in the same context, we maintain the approximate meaning of the text at the lexical level. It is important that the degree of obfuscation is not too small (as it would then be too easy to reconstruct the original note), nor too large (as the new tokens would be completely unrelated). We suggest that randomly varying the degree of obfuscation per token across a single clinical note will further strengthen the security of this approach (by increasing the difficulty of note reconstruction).

The step-by-step method is shown as Algorithm 1, and example outputs can be seen in Table 1.

Algorithm 1. Algorithm to replace each token in a clinical note with a semantically proximate token chosen at random.

Require: N (Degree of obfuscation)
Require: M (Word embedding model)
Require: D (Clinical note)
Ensure: $N > 1$
  $W \Leftarrow$ words in D as a list of tokens
  $W' \Leftarrow [\,]$
  for w_i in W **do**
    $w_v \Leftarrow$ vector of w_i from M
    $lst_w \Leftarrow$ nearest N tokens to w_v in $M - w_i$
    $rnd_w \Leftarrow$ random token from lst_w
    $W' \Leftarrow W'$ append rnd_w
  end for
  return W'

Table 1 shows a sample (artificial) clinical note along with deidentified versions of the note that result from traditional deidentification algorithms and from our algorithm with 3 different degrees of obfuscation (ie, values of N). In this example, we see that our replacement algorithm presents relevant medical terms such as *hcv* (hepatitis C) being replaced with *hbv* (hepatitis B, a common coinfection), *ebv* (a virus in the hepatitis family), or *hepc* (alternative shorthand for hepatitis C). *Alcoholic cirrhosis* (scarring of the liver due to alcohol abuse) is replaced by *alcohcclc cirhosis* (a misspelling of the same symptom), *abstainer steatohepatitis* (*abstainer* is close to *alcoholic*, and *steatohepatitis* is a type of fatty liver disease), and *exdrinker cirrhotic* (again relevant to *alcohol*, and the adjective form of the noun). The misspellings come from the corpus itself, as clinical texts are invariably filled with grammatical and spelling errors; correcting misspellings is still an unsolved research problem. We would stress that these terms are not truly interchangeable as they represent differing patient pathologies. Nonetheless, our empirical experiments show that both our upstream evaluations and our downstream classification tasks are not affected by these substitutions.

We also observe that all the names have been replaced with other names and not with misspellings of the original name. We hypothesize that this is because, considering all other tokens that occur in similar contexts, misspellings are less likely to occur than name tokens of other patients with the same ailment. This is the opposite to the situation for most other kinds of tokens (eg, grammatical and medical terms) where misspelling replacement is much more likely, because the context of misspelled tokens is likely to be extremely similar.

Table 1. An artificial clinical note, and the result of applying our technique with 3 different degrees of obfuscation. Our algorithm does not assume proper spelling or grammar from the input. The obfuscated notes have less readability but maintain important information for ML applications while covering PHI

Note Type	Text
Original note	arnold smith is a fifty year old male, with a history positive for alcholic cirrhosis, hcv, and variceal bleeds, presenting to the ed with syncope and an inner lip laceration after fall on face
PHI removal	**NAME* *NAME** is a **AGE** year old male, with a history positive for alcoholic cirrhosis, HCV, and variceal bleeds, presenting to the ED with syncope and an inner lip laceration after falling forward onto his face .
PHI replacement	*John Bobby* is a *sixty* year old male, with a history positive for alcoholic cirrhosis, HCV, and variceal bleeds, presenting to the ED with syncope and an inner lip laceration after falling forward onto his face .
$N = 3$	muller doug was another seventy ycar monthold man, wth an hislory equivocal ibr alcohcllc cirhosis, hbv, arid varlceal bdoands, chief restrainting this er wth palpitations however a outer lid lacerations afer falling onthe cheeks
$N = 5$	seth joe remains another sixty years olf female, wit another hx positivity forthe abstainer steatohepatitis, ebv, however varicies bleed, chief restrainting its ahc vith presyncopc but acardiogenic supralateral lid abrasion thereafter concussion onthe forehead
$N = 7$	*howard doug looks the thirteen decade monthoid man, wilh wiowill histoiy postive ofr exdrinker cirrhotic, hepc, similarly hemorroidal epistaxis, longstanding insalin ihe ahc wtih presyncopc similarly a posterior gingiva lacn before summer brewere scalp*

Experiments

In this section, we demonstrate that our replacement technique has little to no effect on ML performance in clinical classification tasks that use the deidentified data. More specifically, we evaluate the effect of our technique on the tasks most common in the current literature. First, we perform intrinsic tests on word embeddings created from notes to which our algorithm has been applied. We do this to simulate researchers creating and using embeddings from data that have been provided to them after securing and using our method. We compare the performance of different degrees of anonymization alongside the performance of out-of-domain datasets to assess the relative decrease. Second, we evaluate downstream tasks—diagnostic code and International Classification of Diseases-9 (ICD-9) code prediction—on fully anonymized notes.

Intrinsic evaluation

To test the intrinsic quality of word embeddings generated from the anonymized clinical data, we employ the testing strategy of Wang

et al.[17] They compared word embeddings generated from a variety of sources against human-annotated values of word similarity for a list of clinically relevant terms.

To generate word embeddings for comparisons, we used consultation notes provided to ICES (previously known as the Institute for Clinical Evaluative Sciences) under data-sharing agreements with physicians for the purposes of evaluation and research. Consultation notes are written by physicians and healthcare providers after interacting with a patient. These notes describe history collected, results observed, tests performed, and other details that a physician thinks are important for the treatment of the patient. We used all patient consultation notes (9 051 707 notes), composed of 949 782 513 tokens (2 612 592 unique tokens), Table 2.

For data preprocessing, we lower-cased all tokens, removed special characters and numbers and split words on space and punctuation tokens. We used these notes to train word embeddings using the continuous bag-of-words (CBOW) algorithm with an embedding size of 100, a context window of 5, and a negative sampling rate of 5. These values were picked only once as standard values because they have been shown to work in the clinical setting.[17] We then applied our replacement algorithm, sampling randomly from the closest 3, 5, or 7 tokens. From the newly anonymized set of consultation notes, we created new embeddings with the exact same set of parameters. We compared the quality of embeddings created on the original consultation notes to the quality of embeddings created from anonymized consultation notes, defining quality as correlation to human judgments. We also included a comparison to the quality of embeddings trained on biomedical literature and news corpora to see whether the drop in quality from anonymization renders the specialized data useless in comparison to cheaper alternatives. The biomedical embeddings were trained on a snapshot of the Open Access Subset[18] of the PubMed Central in March 2016. PubMed Central is an online digital collection of freely available full-text biomedical literature containing more than 1.25 million biomedical articles, with 2 million distinct tokens in the vocabulary. The news corpus used was the Google News dataset.[19] This corpus is trained on approximately 100 billion tokens (composed of 3 million unique words or phrases).

Table 2. Description of the consultation notes dataset

	Counts
Number of patients	542 651
Number of notes	9 051 707
Number of tokens	949 782 513
Number of unique tokens	2 612 592

For this evaluation, we used 4 word-pair lists composed of pairs of biomedical words and the numeric degree of semantic similarity between the word-pairs. The semantic similarity is based on human judgments from medical coders and physicians that are provided in the datasets. Specifically, we analyze the performance of word embeddings on the following datasets: 1) Pedersen's[20] (30 medical term pairs), 2) Hliaoutakis'[21] (34 medical term pairs), 3) MayoSRS[22] (101 clinical term pairs), and 4) UMNSRS[23] (566 medical term pairs). Following Wang et al,[17] if a term is composed of multiple words, the overall vector for the term is the average of all the individual word vectors. FastText[24] was used to generate word embeddings for out-of-vocabulary words. For each of the paired terms, we measured the cosine distance and presented the Pearson correlation in Figure 1.

The results of this experiment show that our anonymization technique does not greatly impact the quality of the embeddings (except for $N = 3$ on the Hliaoutakis word-pair list). Believing that this poor performance was simply caused by chance during the shuffling of the data, we ran the models 5 times using the same settings and observed that this bad run was, in fact, caused by chance. The average Pearson correlation is over 10 points higher and within 2 points of the unanonymized model performance, shown in Table 3.

As observed, the quality of the anonymized word embeddings, as measured by these tests, is still higher than that of embeddings trained on out-of-domain corpora, informing us that: i) the noise added to the corpora by replacing each token with a random neighbor generally maintains the overall co-occurrence statistics (hence no significant change in the positive or negative direction), and ii) the embeddings created from anonymized data remain more informative (insofar as they correlate better with human annotations) than embeddings trained on out-of-domain corpora, demonstrating that the anonymized data remains useful.

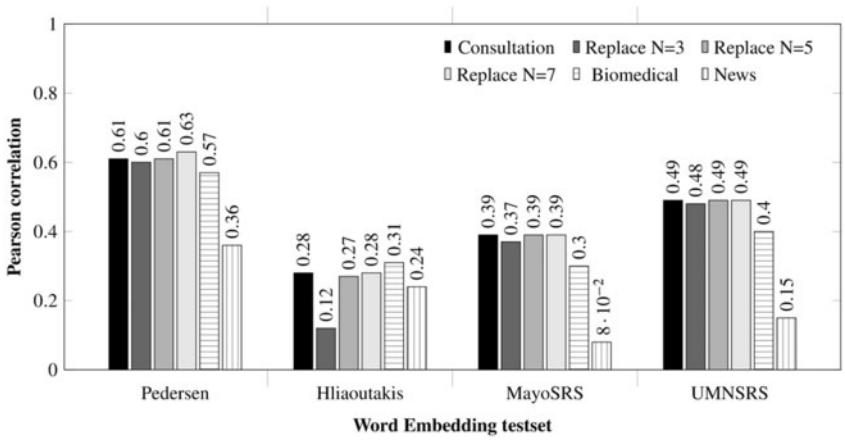

Figure 1. Pearson correlations of the intrinsic word embedding test. The baseline is in solid black, outputs from our technique are in shades of grey, and nonclinical sources are in horizontal and vertical grey lines. As shown, increasing the degree of obfuscation with which we randomly sample does not greatly impact the quality of the word embeddings.

Table 3. Pearson correlations (with 90% confidence interval bracketed beneath) of the intrinsic word embedding test done 5 times for each setting of N = 3, 5, and 7 to measure the effect of randomly shuffling. As can be seen, conclusions drawn regarding comparable performance can still be observed. This also demonstrates that the bad result shown in the body was a result of bad luck/randomization

	Consultation	$N = 3$	$N = 5$	$N = 7$
Pedersen	0.61	0.54 (0.51, 0.56)	0.64 (0.62, 0.65)	0.62 (0.61, 0.63)
Hliaoutakis	0.28	0.26 (0.19, 0.32)	0.26 (0.24, 0.27)	0.24 (0.19, 0.29)
MayoSRS	0.39	0.38 (0.37, 0.39)	0.39 (0.39, 0.40)	0.39 (0.38, 0.40)
UMNSRS	0.49	0.49 (0.48, 0.49)	0.49 (0.49, 0.49)	0.48 (0.48, 0.49)

Table 4. Summary of all experiments. The list of models is organized column-wise by task. In brackets, we present the word embedding algorithm used to randomly replace each token (CBOW or Skipgram). We also present the size of the nearest neighboring set of obfuscating tokens from which we randomly sample. For obfuscation settings, $N = 0$ is the evaluation on the original unprotected dataset, and for $N = 3$–14, we varied the size of the nearest neighbor set for each word between 3 and 14 instead of holding it constant for each token

Obfuscation (N)	Models for ICES diagnostic code classification	Models for MIMIC ICD-9 classification	Models for sentiment analysis
$N = 0$	Logistic regression (CBOW)	CNN (CBOW)	Logistic regression (CBOW)
$N = 3$	SVM (CBOW)	CNN with attention (CBOW)	SVM (CBOW)
$N = 5$	CNN (CBOW)	LSTM (CBOW)	CNN (CBOW)
$N = 7$	Logistic regression (Skipgram)	LSTM with attention (CBOW)	Logistic regression (Skipgram)
$N = 9$	SVM (Skipgram)		SVM (Skipgram)
$N = 3$–14	CNN (Skipgram)		CNN (Skipgram)

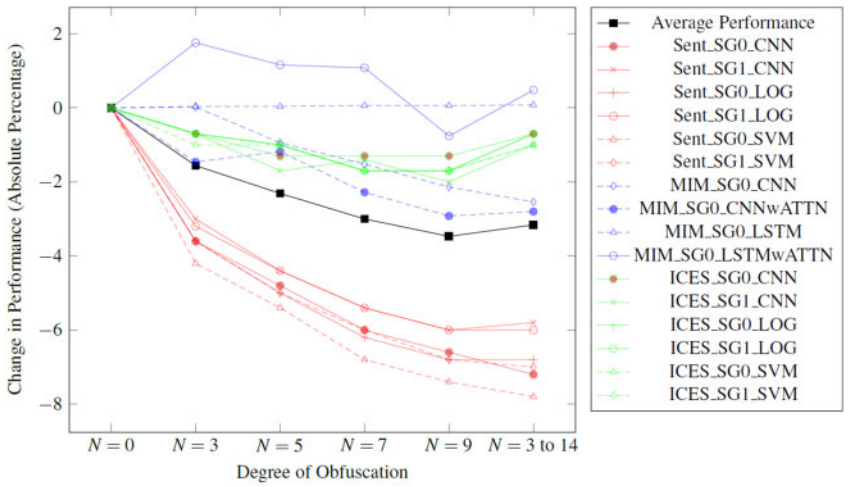

Figure 2. Absolute percentage change of performance (F_1 score) as a function of different obfuscation settings for various tasks, settings, and models. Each model name is broken into 3 parts: 1) The task performed, of which there are 3 (*Sent* for Sentiment Classification, *MIM* for MIMIC III ICD-9 code classification, or *ICES* for ICES diagnostic code classification); 2) the word embedding representation used to learn randomly replace the tokens (either SG0 or SG1 for CBOW or Skipgram); and 3) the type of model used to classify the texts. More details regarding each of these settings and models can be found in the Supplementary Material.

Extrinsic evaluation

In this section, we test the effect of our anonymization technique on multiple prediction tasks.

Our first task is diagnostic code prediction from clinical notes. We perform this task using 2 different datasets (ICES and MIMIC III[25]) to demonstrate that our approach works with differing types of clinical notes (eg, progress notes and discharge notes). We work with embeddings created from 2 of the most popular word embedding algorithms (CBOW and Skipgram) to demonstrate that our results do not hinge on any single algorithm. We also test a variety of ML models to demonstrate that our technique preserves enough signal to remain useful for many different classifiers. To further demonstrate the generalizability of our algorithm, we also perform binary sentiment classification on movie reviews. We purposefully choose sentiment analysis, because tokens of opposing sentiments tend to appear in similar contexts (eg, "This movie was good" and "This movie was bad") and are therefore mutual candidates for replacement. By showing that our algorithm does not negate the signal in sentiment analysis classification, we thereby highlight its ability to preserve information.

Table 4 presents the complete set of experiments conducted. Figure 2 presents the average decrease in performance for different

degrees of obfuscation for each task for each classifier. The complete details of all the data used and the setup of each experiment is given in the Supplemental Material.

From Figure 2, we observe that increased obfuscation generally results in decreased classification performance (measured using F_1 score). However, the observed decreases are small in magnitude, no more than approximately 5% for clinical tasks, thereby demonstrating the utility of data protected using our method.

DISCUSSION

The obfuscated data created by our technique remains useful for many ML classification tasks. By replacing tokens with other tokens that occur frequently in the same context, we are not changing the underlying distribution greatly. Thus, the performance of ML and NLP classification methods is not greatly impacted and may be used for pilot research projects. In the Supplementary Appendix, we validate the practicality of our technique by performing additional classification tasks, which demonstrate that performance is preserved when applying secured embeddings to nonanonymized text and for using pretrained embeddings on anonymized text. Of course, there are tasks for which our technique may not be the optimal approach for anonymizing data—for example, clinical named-entity recognition and tasks requiring human interaction or interpretability. More research is required to evaluate the impact of our method on other tasks.

Quantitative assessment of the security of notes itself is a challenge, since existing measures of security are insufficient, and therefore so are existing shared tasks.[11,13] Using precision, recall, or Carrell et al's[15] approach (The issue of using other information in notes to deduce the patient name is no longer an issue here either, because: i) the original patient name is no longer in the note, and ii) all other personal details will have been obfuscated as well.) to evaluate our technique would not be appropriate, as our technique purposely alters each token, thereby achieving perfect recall with each PHI replaced at the cost of low precision.

This perfect recall also comes at the cost of agrammatical, and sometimes even unreadable, transformations. Using a dictionary-based search method to preclude certain words from being replaced would increase readability; however, choosing to keep a predetermined list of informative words, for example, stop-words or medical names (some of which might also be human names, such as Parkinson's) would increase readability as well as risk. This reintroduction of risk should only be done for specific use-cases and under controlled access measures.

Theoretically, this approach could also use contextual word embeddings[26,27]—embeddings that change depending on the context. Future work would have to show that the trends observed above hold and study whether model size and its contextual nature have any negative effects regarding random neighbor generation or privacy.

To increase security, we considered choosing differing values of N for different tokens. Instead of choosing a replacement from a fixed-sized set of neighboring tokens, allowing for the size of the set to randomly change per token will protect isolated clusters of low-frequency tokens (which may have been easier to isolate, but not fully detect). The results of such a setting are presented for all tasks under the setting $N = 3$–14. Regardless, our technique, if used alone, may still be susceptible to attacks. Should a malicious actor gain access to the original word embeddings, it may be possible to deduce patient identities, as it would be possible to reconstruct the neighbors for each token. This risk is also why our approach should not be used on small datasets. For maximum security, our approach to anonymization should be applied after a more traditional privacy-replacement technique in order to obfuscate instances where the more traditional approaches to privacy have failed.

We are not advocating that this method should be used on the input of a model deployed in real clinical settings. Rather, we propose that this method can be used in pilot classification tasks very quickly and at low cost. For example, to explore the possibility of automatically classifying text, data holders can share data that have been anonymized using our method at reduced risk. If any of the research groups involved were able to achieve acceptable performance, then that collaboration or development could be brought in-house to work on real data. Our approach allows data holders to outsource ML research and data analytics to outside research groups without the overhead of creating and maintaining a manually secured data repository.

CONCLUSION

In this work, we introduced a novel anonymization technique for clinical notes that can be applied to any body of text. The method:

- is generalizable across different types of text data, as demonstrated by our application to consultation notes, progress notes, and movie reviews;
- guarantees that all PHI will be randomly replaced with perfect recall, a claim that cannot be made of algorithms that currently exist in the literature; and
- does not result in a significant decrease in performance for classification tasks using either neural networks or more traditional machine learning.

The algorithm provides complete coverage on all sensitive information at the cost of introducing some noise that reduces human readability. However, we have shown through our intrinsic tests (ie, correlation scores with human annotated word-pair lists) and extrinsic tests (ie, 10-way diagnostic code classification and binary sentiment classification) that the amount of noise introduced does not negate the benefits of having a specialized corpus to create embeddings for certain ML and NLP classification tasks. Future work should further explore additional NLP tasks, improving the human readability aspect and possibly performing mixed-method experiments with physicians and clinicians.

Using our algorithm, in conjunction with other advanced NLP techniques to detect PHI, may both maintain readability and improve privacy, but this is left for future work. In clinical settings, we propose that our technique should be used to protect data granted to outside researchers in the broader research team, expanding the potential for future discoveries.

FUNDING

This work was supported by grants from the Natural Sciences and Engineering Research Council of Canada to GH and FR, a Vanier Canada Graduate Scholarship to $MohA_1$. FR is supported by a CIFAR Chair in Artificial Intelligence.

AUTHOR CONTRIBUTIONS

$MohA_1$ and $MouA_2$ designed the experiments. $MohA_1$ programmed the experiments. $MohA_1$, $MouA_2$, GH, and FR wrote the paper.

MohA$_1$ and GH formulated the original problem. FR provided direction and guidance.

SUPPLEMENTARY MATERIAL

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

ACKNOWLEDGMENTS

We thank the helpful staff at ICES (Dr. Liisa Jaakkimainen, Dr. Therese Stukel, Elisa Candido, Daniella Barron) for granting us access to data and for their insight on our manuscript. This study was supported by ICES, which is funded by an annual grant from the Ontario Ministry of Health and Long-Term Care (MOHLTC). The analyses, conclusions, opinions, and statements expressed herein are solely those of the authors and do not reflect those of the funding or data sources; no endorsement is intended or should be inferred.

CONFLICT OF INTEREST

None declared.

REFERENCES

1. Kajiyama K, Horiguchi H, Okumura T, Morita M, Kano Y. De-identifying free text of Japanese dummy electronic health records. In: *Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis*.
2. Dernoncourt F, Lee JY, Uzuner O, Szolovits P. De-identification of patient notes with recurrent neural networks. *J Am Med Inform Assoc* 2017; 24 (3): 596–606.
3. Neamatullah I, Douglass MM, Li-Wei HL, *et al*. Automated de-identification of free-text medical records. *BMC Med Inform Decis Mak* 2008; 8 (1): 32.
4. Fernandes N, Dras M, McIver A. Author obfuscation using generalised differential privacy. arXiv preprint arXiv: 1805.08866. 2018.
5. Schakel AM, Wilson BJ. Measuring word significance using distributed representations of words. arXiv preprint arXiv: 1508.02297. 2015.
6. Gong C, He D, Tan X, Qin T, Wang L, Liu TY. Frage: Frequency-agnostic word representation. In: *Advances in Neural Information Processing Systems*. Cambridge, MA: MIT Press; 2018: 1334–1345.
7. Miller R, Boitnott JK, Moore GW. Web-based free-text query system for surgical pathology reports with automatic case deidentification. *Arch Pathol Lab Med* 2001.
8. Thomas SM, Mamlin B, Schadow G, McDonald C. A successful technique for removing names in pathology reports using an augmented search and replace method. *AMIA Annu Symp Proc* 2002; 2002: 777–781.
9. Sibanda T, He T, Szolovits P, Uzuner O. Syntactically-informed semantic category recognizer for discharge summaries. *AMIA Annu Symp Proc* 2006; 2006: 714–8.
10. Liu Z, Tang B, Wang X, Chen Q. De-identification of clinical notes via recurrent neural network and conditional random field. *J Biomed Inform* 2017; 75: S34–S42.
11. Stubbs A, Filannino M, Uzuner Ö. De-identification of psychiatric intake records: Overview of 2016 CEGS N-GRID Shared Tasks Track 1. *J Biomed Inform* 2017; 75: S4–18.
12. Yang H, Garibaldi JM. Automatic detection of protected health information from clinic narratives. *J Biomed Inform* 2015; 58: S30–S38.
13. Stubbs A, Kotfila C, 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: S11–9.
14. Meystre SM, Ferrandez O, Friedlin FJ, South BR, Shen S, Samore MH. Text de-identification for privacy protection: a study of its impact on clinical text information content. *J Biomed Inform* 2014; 50: 142–50.
15. Carrell D, Malin B, Aberdeen J, *et al*. Hiding in plain sight: use of realistic surrogates to reduce exposure of protected health information in clinical text. *J Am Med Inform Assoc* 2013; 20 (2): 342–8.
16. Sahlgren M. The distributional hypothesis. *Italian J Linguist* 2008; 20 (1): 33–54.
17. Wang Y, Liu S, Afzal N, *et al*. A comparison of word embeddings for the biomedical natural language processing. *J Biomed Inform* 2018; 87: 12–20.
18. Website: http://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/Accessed December 13, 2019.
19. Website: https://code.google.com/archive/p/word2vec/Accessed December 13, 2019.
20. Pedersen T, Pakhomov SV, Patwardhan S, Chute CG. Measures of semantic similarity and relatedness in the biomedical domain. *J Biomed Inform* 2007; 40 (3): 288–99.
21. Hliaoutakis A. *Semantic similarity measures in MeSH ontology and their application to information retrieval on Medline* [Master's thesis]. 2005.
22. Pakhomov SV, Pedersen T, McInnes B, Melton GB, Ruggieri A, Chute CG. Towards a framework for developing semantic relatedness reference standards. *J Biomed Inform* 2011; 44 (2): 251–65.
23. Pakhomov S, McInnes B, Adam T, Liu Y, Pedersen T, Melton GB. Semantic similarity and relatedness between clinical terms: an experimental study. *AMIA Annu Symp Proc* 2010; 2010: 572.
24. Bojanowski P, Grave E, Joulin A, Mikolov T. Enriching word vectors with subword information. *Trans Assoc Comput Linguist* 2017; 5: 135–46.
25. Johnson AE, Pollard TJ, Shen L, *et al*. MIMIC-III, a freely accessible critical care database. *Sci Data* 2016; 3 (1).
26. Peters M, Neumann M, Iyyer M, *et al*. Deep contextualized word representations. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.
27. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.
28. Liendo Z, Roo GD, Karmakar A. Classifying medical notes into standard disease codes. https://github.com/zliendo/AI_MedicalNotes/blob/master/w266FinalReport_ICD_9_Classification.pdfAccessed December 13, 2019.
29. Maas AL, Daly RE, Pham PT, Huang D, Ng AY, Potts C. Learning word vectors for sentiment analysis. In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*.
30. Kim Y. Convolutional Neural Networks for Sentence Classification. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP).