Anonymization Through Substitution: Words vs Sentences

Vasco Alves[1], Vitor Rolla[1], João Alveira[1], David Pissarra[1], Duarte Pereira[1], Isabel Curioso[1], André V. Carreiro[1], Henrique Lopes Cardoso[2]
[1]Fraunhofer AICOS, Portugal
[2]FEUP, Portugal
{andre.carreiro,vitor.rolla}@fraunhofer.pt

Abstract

Anonymization of clinical text is crucial to allow the sharing and disclosure of health records while safeguarding patient privacy. However, automated anonymization processes are still highly limited in healthcare practice, as these systems cannot assure the anonymization of all private information. This paper explores the application of a novel technique that guarantees the removal of all sensitive information through the usage of text embeddings obtained from a de-identified dataset, replacing every word or sentence of a clinical note. We analyze the performance of different embedding techniques and models by evaluating them using recently proposed evaluation metrics. The results demonstrate that sentence replacement is better at keeping relevant medical information untouched, while the word replacement strategy performs better in terms of anonymization sensitivity.

1 Introduction

With the increasing adoption of Electronic Health Record (EHR) systems, clinical data has become available in large amounts to be used by healthcare practitioners (Meystre et al., 2010). However, it often contains sensitive information about patients and healthcare professionals that needs to remain private when being shared in order to comply with data protection regulations such as the Health Insurance Portability and Accountability Act (HIPAA) (U.S. Department of Health & Human Services, 2013) in the United States of America and the General Data Protection Regulation (GDPR) (GDPR, 2018) in the European Union.

Many systems for the anonymization of clinical text have been developed throughout the years, ranging from solutions relying on hand-crafted rules and patterns (Sweeney, 1996; Beckwith et al., 2006; Friedlin and McDonald, 2008) to more complex systems based on machine and deep learning (Wellner et al., 2007; Aramaki et al., 2006; Yang and Garibaldi, 2015; Liu et al., 2017; Dernoncourt et al., 2016; Yang et al., 2019; Alsentzer et al., 2019). Although some of these systems show impressive results, their lack of adoption in real-world scenarios remains a barrier to sharing clinical data and its usage for secondary purposes. One should also consider whether perfect anonymization, i.e., removing all the sensitive information while keeping the non-sensitive information intact, is an achievable goal (Stubbs et al., 2015).

While traditional Named Entity Recognition (NER) based methods have shown impressive performance in anonymization tasks, achieving recall rates of over 90%, they still have limitations. Abdalla et al. (Abdalla et al., 2020) emphasized this issue, noting that relying solely on precision and recall for evaluating de-identification algorithms carries the risk of missing sensitive information. To tackle this challenge, they introduced an innovative solution. Instead of solely relying on NER, they proposed a method that utilizes proximity measures between word embeddings. This approach replaces each token in a clinical note with a semantically similar one, ensuring the removal of all sensitive information. However, this method raises concerns about potential information loss and readability issues. Ribeiro et al. (Ribeiro et al., 2023) have implemented this strategy on the INCOGNITUS toolbox, naming it K-Nearest Embeddings Obfuscation (KNEO). This work follows their approach and aims to compare two different strategies for the replacement - using word or sentence embeddings - by evaluating them on new and adapted metrics for anonymization sensitivity and clinical information loss.

The remainder of this paper is structured as follows: Section 2 provides an overview of word and sentence embeddings. Section 3 outlines the evaluation metrics to compare the proposed strategies, and Section 4 describes the used methodology. Additionally, Section 5 provides a discussion and anal-

Proceedings of the Fifth Workshop on Privacy in Natural Language Processing, pages 85–90

ysis of the obtained results, and Section 6 lays out the conclusions. Lastly, Section 7 provides insights into some limitations of the analyzed solutions.

2 Embeddings

Finding representations of text is a necessary step in most Natural Language Processing (NLP) tasks (Almeida and Xexéo, 2023). Word embeddings are commonly used by representing each word as a fixed-length vector of real numbers that captures useful syntactic and semantic properties (Turian et al., 2010). These representations allow the words to be the subject of mathematical operations that wouldn't otherwise be possible (Almeida and Xexéo, 2023), aiding in finding similarities between text pieces.

Similarly to word embeddings, sentence embeddings are representations of entire sentences as fixed-size vectors in a continuous vector space. These embeddings capture the semantic meaning and context of the entire sentence, encoding information about word usage, syntax, and semantics. Sentence embeddings models are trained on large text corpora and learn to encode sentences into meaningful vector representations.

2.1 Word2Vec

Word2Vec (Mikolov et al., 2013) is an algorithm based on neural networks that produce continuous vector representations of words by learning relationships between them using large amounts of plain text. These words are embedded in a vector space where close vectors represent words with similar meanings, and distant vectors represent differing meanings.

2.2 Doc2Vec

Doc2Vec (Le and Mikolov, 2014) extends the concept of Word2Vec to complete sentences or documents. It enables, through unsupervised learning, the generation of fixed-length numerical representations, or vectors, for variable-length pieces of text, such as sentences, paragraphs, or documents.

2.3 Sentence Transformers

Sentence transformers are a cutting-edge approach in NLP that leverages pre-trained transformer models to encode sentences into dense vector representations. It originates from the work of Sentence-BERT (Reimers and Gurevych, 2019), a modification of the pre-trained BERT network in order to obtain semantically meaningful sentence embeddings that can be compared. This approach obtained state-of-the-art results on common Semantic Textual Similarity (STS) tasks, outperforming other sentence embedding methods.

3 Evaluation Metrics

We evaluate the performance of our strategies using the evaluation metrics proposed by (Pissarra et al., 2024). The authors divide the metrics into two categories: anonymization sensitivity metrics and clinical information retention metrics. The first category, whose focus is on the masking of sensitive entities, contains the following metrics: String Matching-based Recall (SMR), Average Levenshtein Index of Dissimilarity (ALID), Levenshtein Recall (LR), Levenshtein Recall for Direct Identifiers (LRDI) and Levenshtein Recall for Quasi Identifiers (LRQI). The clinical information retention metrics, Jaccard Similarity Coefficient (JSC) and Normalized Softmax Discounted Cumulative Gain (NSDCG), are based on the usage of a BioBERT (Lee et al., 2020) model, which has been pre-trained on a hierarchical classification task of ICD-10 code categories. These evaluation metrics and their formulas are described in detail in the previously mentioned paper.

4 Methodology

The following methodology allows the comparison between the proposed strategies and models. Two anonymization strategies, word and sentence substitution, were evaluated using one and four models, respectively.

4.1 Data

The MIMIC-III clinical database (Johnson et al., 2016) is a large, de-identified and freely available dataset comprised of health-related data. A subset of 33,321 discharge summary notes were used to generate the embedding space, and another of 19,989 notes was used to evaluate the different approaches. MIMIC-III contains different note types with varying proportions, and it was assured that both subsets have the same distribution.

4.2 Pre-Processing

In the MIMIC-III dataset, the sensitive information is replaced by category tags. To obtain a more realistic version of the notes, the Faker[1] li-

[1] https://faker.readthedocs.io/en/master/

brary for Python was used to create fake entities according to each category. Lowercasing, removal of consecutive white spaces, and removal of non-alphanumeric characters were performed on the text before the respective embeddings were calculated.

4.3 Word2Vec Anonymization

A word embedding model was trained on the 33,321 clinical notes using Gensim's implementation of Word2Vec[2], creating a de-identified embeddings space. To anonymize a new clinical note, for each token, we obtain its embedding using the trained model and replace it with a different one selected randomly from the top 5 most similar ones present in the embeddings space. The Word2Vec model was trained for 100 epochs with the following parameters: vector_size = 256, window = 15, min_count = 1, workers = 1.

4.4 Doc2Vec Anonymization

Similarly to the Word2Vec Anonymization, a document embeddings model was trained on the same clinical notes using Gensim's implementation of Doc2Vec[3], creating the de-identified embeddings space. To anonymize a clinical note, we obtain the embedding for each sentence using the trained model and replace each of them with a different one selected randomly from the top 5 most similar ones present in the embeddings space. The Doc2Vec model was trained for 100 epochs with the following parameters: vector_size = 256, dm = 0, window = 15, min_count = 1, workers = 1.

4.5 Sentence-Transformer Anonymization

We experiment with different pre-trained sentence-transformer models available in the SentenceTransformers Python framework[4]. These models were used to encode the sentences contained in the 33,321 clinical notes into embeddings, generating the de-identified embeddings space. When anonymizing a clinical note, its sentences are encoded into embeddings using the same pre-trained model and replaced by a different one selected randomly from the top 5 most similar ones previously encoded. The following three models were used:

all-MiniLM-L6-v2 Baseline model that maps sentences into a 384-dimensional dense vector space.

avsolatorio/GIST-large-Embedding-v0 Model that has a good performance on the BIOSSES (biomedical sentence similarity estimation) benchmark. Generates embeddings with 1024 dimensions.

pritamdeka/S-PubMedBert-MS-MARCO Model trained on biomedical text from PubMed that maps sentences to a 768-dimensional dense vector space.

4.6 Evaluation

Each model's performance was tested on the 19,989 notes reserved for the evaluation. Anonymized versions of the clinical notes were produced using the previously described replacement strategies, which were then evaluated using the evaluation metrics mentioned in Section 3. The following distribution of MIMIC-III categories was used for the LRDI and LRQI metrics: NAME, CONTACT_NUMBER, ID, and EMAIL were considered direct-identifiers, and LOCATION, DATE, URL, AGE_ABOVE_89, INSTITUTION, and HOLIDAY were considered quasi-identifiers.

5 Results and Discussion

Figure 1 illustrates the performance obtained by each model on the different evaluation metrics by averaging the results obtained for all the test notes.

We can observe that word replacement obtains better results on all the anonymization metrics except for ALID but performs worse regarding clinical information retention. This is an expected outcome, as it is related to the way the anonymization is being performed. For example, when anonymizing a clinical note with the sentence "The patient's name is John Doe", the word replacement strategy will replace every word in the sentence. However, when using sentence replacement, it could be the case that it is replaced with a different sentence that contains common elements, such as "John" or "Doe," thus negatively impacting the performance of these metrics.

The same rationale explains the better performance of sentence replacement in the information retention metrics. For instance, if the name of a medical condition appears in the clinical note we want to anonymize, replacing every word will result in that medical condition no longer being there.

[2] https://radimrehurek.com/gensim/models/word2vec.html

[3] https://radimrehurek.com/gensim/models/doc2vec.html

[4] https://sbert.net/

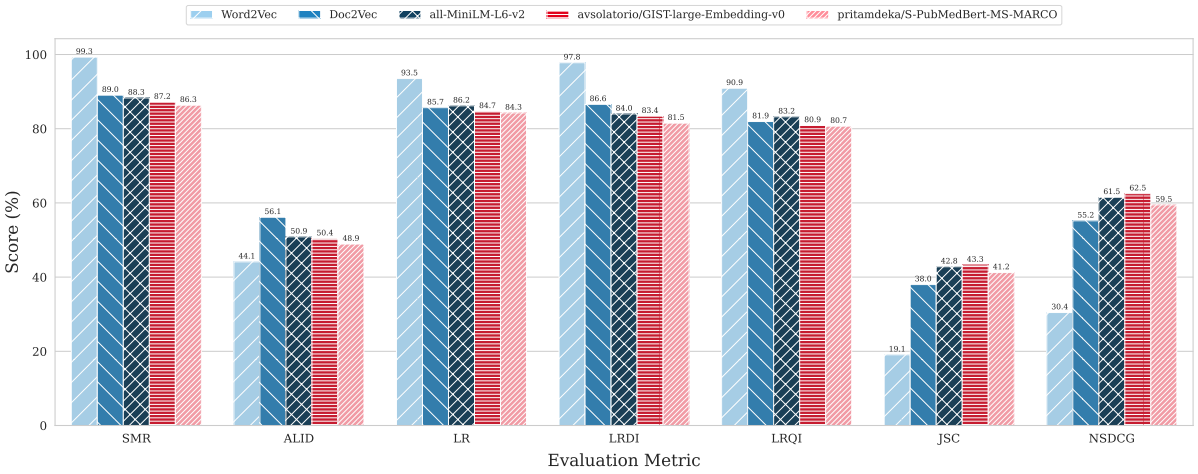

Figure 1: Performance results obtained by each model on the different evaluation metrics. The results are presented as the average of the metrics measured across 19,989 notes used for testing.

As for the sentence replacement, it is possible for the substitute sentence also to contain the name of the said medical condition.

Interestingly, replacing every word of the given clinical notes did not achieve a score of 100% in any of the metrics involving recall. This can be attributed to the fact that some sensitive entities can appear as subwords of other non-sensitive entities. Additionally, MIMIC-III also contains some labeling errors.

Regarding the anonymization sensitivity metrics, there is no discernable difference in performance between the Doc2Vec and the Sentence-Transformer models. It is interesting to notice that Doc2Vec and all-MiniLM-L6-v2, being the two models that produce vectors with the lowest number of dimensions, outperformed the two models that produce vectors with a much higher number of dimensions. This is because each dimension captures different semantic and syntactic attributes of the text, which may not be totally useful for the anonymization itself.

On the two information retention metrics, however, the Sentence-Transformer models perform better than the Doc2Vec model. In this case, the dimensionality of the produced vectors most likely influences the results, as these metrics rely on the similarity of the original and anonymized version of the note. The avsolatorio/GIST-large-Embedding-v0 model obtains the best performance in both metrics. It is an expected result, as it is the model that produces vectors with the highest number of dimensions, which results in a better capturing of similar sentences. Additionally, this pre-trained model is one of the best-performing models on the BIOSSES benchmark. As for the pritamdeka/S-PubMedBert-MS-MARCO model, its lower performance might indicate that the PubMed text it was trained on differs from the clinical text contained in the MIMIC-III database.

While no strategy was better across all metrics, our strategies are based on the premise that the replacement group contains no sensitive information, and therefore, neither will the anonymized version of a clinical note. The lower performance the sentence replacement strategy obtains on the information retention metrics can originate from the overlap of fake sensitive entities in the replacement group and the test set. For example, a fake entity appearing in a note we are anonymizing may have already appeared in a sentence for the embedding space generation, which influences the sentence replacement process. Although it is a fake entity, its presence in the anonymized version will have an influence on the results. Had we utilized a dataset with real sensitive information, this overlap would likely have decreased and boosted the anonymization sensitivity results. As such, we look at sentence replacement as the better approach.

6 Conclusions

This work presents a comparison between two different and novel techniques for the anonymization of clinical notes. Five different models were tested and evaluated across several evaluation metrics aimed at anonymization sensitivity and clinical information retention. The discussed results indicate that both replacement techniques have their

unique strengths and are viable alternatives to the traditional NER (Named-Entity Recognition) approaches when the removal of sensitive information is a priority over data usefulness, as the latter are never capable of detecting all the sensitive information.

7 Limitations

We present a strategy that assures the removal of all sensitive information by replacing every word/sentence with similar counterparts obtained from a de-identified dataset. However, it comes at the expense of readability and data usefulness, as there is no guarantee that the anonymized version of the note will be semantically or syntactically correct. Consequently, there is no guarantee that the agreement on gender, age group, and person will be maintained throughout the new clinical note.

One downside of the word replacement approach is that if a relevant medical term appears on the original version of the clinical notes, it is guaranteed that the same term will not appear on the anonymized version, as every word is being replaced. This is not the case with the sentence replacement approach, which is why there is better performance on the clinical information retention evaluation metrics. However, if we are trying to anonymize a clinical note that contains a sentence with a medical term not present in any sentence of the replacement group, it will result in that term also being permanently lost.

Finally, another possible limitation is the use of the same database for both the embeddings generation and anonymization evaluation. This has been a longstanding problem in the area of text anonymization, as many of the developed solutions are tailored to specific datasets or note types, and there is no guarantee that the performance will be maintained across different scenarios. Using the same type and structure of clinical notes across our whole process may facilitate the step of finding similar words/sentences and, as a result, inflate the clinical information retention results. The performance obtained in these experiments would probably be lower had we used a different dataset for evaluation, as finding similar words or sentences would be harder.

Acknowledgements

This work was supported by European funds through the Recovery and Resilience Plan, via project "Center for Responsible AI", with identification number C645008882-00000055.

References

Mohamed Abdalla, Moustafa Abdalla, Frank Rudzicz, and Graeme Hirst. 2020. Using word embeddings to improve the privacy of clinical notes. *Journal of the American Medical Informatics Association*, 27(6):901–907.

Felipe Almeida and Geraldo Xexéo. 2023. Word embeddings: A survey.

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Eiji Aramaki, Takeshi Imai, Kengo Miyo, and Kazuhiko Ohe. 2006. Automatic deidentification by using sentence features and label consistency. In *i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data*.

Bruce A Beckwith, Rajeshwarri Mahaadevan, Ulysses J Balis, and Frank Kuo. 2006. Development and evaluation of an open source software tool for deidentification of pathology reports. *BMC Medical Informatics and Decision Making*, 6(1):12.

Franck Dernoncourt, Ji Young Lee, Ozlem Uzuner, and Peter Szolovits. 2016. De-identification of patient notes with recurrent neural networks. *Journal of the American Medical Informatics Association*, 24(3):596–606.

F Jeff Friedlin and Clement J McDonald. 2008. A software tool for removing patient identifying information from clinical documents. *Journal of the American Medical Informatics Association*, 15(5):601–610.

GDPR. 2018. General data protection regulation. Official website of the European Union.

Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Li wei H. Lehman, Mengling Feng, Mohammad Mahdi Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific Data*, 3.

Quoc V. Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234 – 1240.

Zengjian Liu, Buzhou Tang, Xiaolong Wang, and Qingcai Chen. 2017. De-identification of clinical notes via recurrent neural network and conditional random field. *Journal of Biomedical Informatics*, 75:S34–S42. Supplement: A Natural Language Processing Challenge for Clinical Records: Research Domains Criteria (RDoC) for Psychiatry.

Stephane Meystre, F Friedlin, Brett South, Shuying Shen, and Matthew Samore. 2010. Automatic de-identification of textual documents in the electronic health record: A review of recent research. *BMC Medical Research Methodology*, 10:70.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space.

David Pissarra, Isabel Curioso, João Alveira, Duarte Pereira, Bruno Ribeiro, Tomás Souper, Vasco Gomes, André V. Carreiro, and Vitor Rolla. 2024. Unlocking the potential of large language models for clinical text anonymization: A comparative study.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks.

Bruno Ribeiro, Ricardo Santos, and Vitor Rolla. 2023. Incognitus: A toolbox for automated clinical notes anonymization. In *Proceedings of the 17th Meeting of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics.

Amber Stubbs, Christopher Kotfila, and Özlem Uzuner. 2015. Automated systems for the de-identification of longitudinal clinical narratives: Overview of 2014 i2b2/uthealth shared task track 1. *Journal of Biomedical Informatics*, 58:S11–S19. Supplement: Proceedings of the 2014 i2b2/UTHealth Shared-Tasks and Workshop on Challenges in Natural Language Processing for Clinical Data.

Latanya Sweeney. 1996. Replacing personally-identifying information in medical records, the scrub system. *Proceedings : a conference of the American Medical Informatics Association. AMIA Fall Symposium*, pages 333–7.

Joseph Turian, Lev-Arie Ratinov, and Yoshua Bengio. 2010. Word representations: A simple and general method for semi-supervised learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 384–394, Uppsala, Sweden. Association for Computational Linguistics.

U.S. Department of Health & Human Services. 2013. Summary of the HIPAA Privacy Rule. `https://www.hhs.gov/hipaa/for-professionals/privacy/laws-regulations/index.html`. [Online; accessed May 5, 2023].

Ben Wellner, Matt Huyck, Scott Mardis, John Aberdeen, Alex Morgan, Leonid Peshkin, Alex Yeh, Janet Hitzeman, and Lynette Hirschman. 2007. Rapidly Retargetable Approaches to De-identification in Medical Records. *Journal of the American Medical Informatics Association*, 14(5):564–573.

Hui Yang and Jonathan M. Garibaldi. 2015. Automatic detection of protected health information from clinic narratives. *Journal of Biomedical Informatics*, 58:S30–S38. Supplement: Proceedings of the 2014 i2b2/UTHealth Shared-Tasks and Workshop on Challenges in Natural Language Processing for Clinical Data.

Xi Yang, Tianchen Lyu, Qian Li, Chih-Yin Lee, Jiang Bian, William R. Hogan, and Yonghui Wu. 2019. A study of deep learning methods for de-identification of clinical notes in cross-institute settings. *BMC Medical Informatics and Decision Making*, 19(S5):232.