De-Identification of Clinical Free Text in Dutch with Limited Training Data: A Case Study

Elyne Scheurwegs
Artesis Hogeschool Antwerpen
elynescheurwegs@hotmail.com

Kim Luyckx
biomina - biomedical informatics group
Antwerp University Hospital
& University of Antwerp
kim.luyckx@uza.be

Filip Van der Schueren
Artesis Hogeschool Antwerpen
filip.vanderschueren@artesis.be

Tim Van den Bulcke
biomina - biomedical informatics group
Antwerp University Hospital
& University of Antwerp
tim.vandenbulcke@uza.be

Abstract

In order to analyse the information present in medical records while maintaining patient privacy, there is a basic need for techniques to automatically de-identify the free text information in these records. This paper presents a machine learning de-identification system for clinical free text in Dutch, relying on best practices from the state of the art in de-identification of English-language texts. We combine string and pattern matching features with machine learning algorithms and compare performance of three different experimental setups using Support Vector Machines and Random Forests on a limited data set of one hundred manually obfuscated texts provided by Antwerp University Hospital (UZA). The setup with the best balance in precision and recall during development was tested on an unseen set of raw clinical texts and evaluated manually at the hospital site.

1 Introduction

In Electronic Health Records (EHRs), medical information about the treatment of patients is stored on a daily basis, both in structured (e.g. lab results, medication,) and unstructured (e.g. clinical notes) forms. EHRs are unique sources of information that need be further analyzed to improve diagnosis and treatment of future patients. However, these information sources cannot be freely explored due to privacy regulations (Privacy Rule, 2002; European Data Protection Directive, 1995; Belgian Data Protection Act, 1993). Automated de-identification is crucial to remove personal health information (PHI), while keeping all medical and contextual information as intact as possible. In the US, this is regulated under the Health Insurance Portability and Accountability Act (HIPAA, 1996).

Approaches to de-identification can be categorised into two main types, with rule-based and pattern matching approaches on the one hand and machine learning approaches on the other, as suggested in Meystre et al. (2010). Rule-based and pattern-matching approaches often rely on dictionaries and manually constructed regular expressions. While this type of approach does not require any annotation effort and can easily be customised to increase performance, it offers only limited scalability and is often highly language dependent. Machine learning approaches in general are better scalable and more robust to noise, but especially supervised learning algorithms require substantial amounts of annotated training data, a very time-consuming and expensive undertaking. The selection of meaningful features is a crucial aspect in the machine learning approach, especially when only limited data is available (Ferrández et al., 2012a). Hybrid approaches to de-identification such as that presented in Ferrández et al. (2012b) have been developed to combine the advantages of the machine learning approach with those of dictionaries and regular expressions. Below, we highlight a number of interesting studies from the state of the art in automated de-identification.

One of the first systems for de-identification, the Scrub system, was proposed in Sweeney et al. (1996). Scrub takes a dictionary rule-based approach and has been shown to be able to effectively model the human approach to locating PHI

entities. This study included well-formatted letters with a header block as well as shorthand notes, but does not provide details on recall and precision.

Stat De-Id (Uzuner et al., 2008; Sibanda, 2006) takes a machine learning approach using Support Vector Machines (SVM) as the learning algorithm. Features cover aspects of the target word as well as of the immediate context of the target. Conditional Random Fields (CRF) (Lafferty et al., 2001) are being used increasingly in de-identification research. Two examples are Health Information DE-identification (HIDE) (Gardner and Xiong, 2008) and the Mitre Identification Scrubber Toolkit (MIST) (Aberdeen et al., 2010; Deleger et al., 2013). Several of these de-identification systems (see also Douglass et al. (2004) and Neamatullah et al. (2008)) show excellent results rivaling manual de-identification. While most de-identification systems score well in terms of recall, they do produce quite a large amount of false positives (see Ferrández et al. (2012a)). This compromises the usability of the de-identified documents, as medically relevant data may have been removed.

In this paper, we present a de-identification case study following best practices from the state of the art. A machine learning approach is taken, using features based on dictionaries and string and pattern matching techniques. The objective of this study is to develop a de-identification system for clinical notes in Dutch, a language for which de-identification training data are not available. We evaluate three machine learning setups on a training set of 100 manually annotated medical notes and test the best performing setup on 100 previously unseen medical notes, the performance of which is manually evaluated at the hospital site.

2 Methods

2.1 Data set

The training set consists of 100 documents randomly selected from the Antwerp University Hospital (UZA) EHR system. This data set consists of (discharge) letters, comprising 52,829 words in total. These words have been annotated manually according to the following Personal Health Information (PHI) classes: *Name*, *Date*, *Address*, *ID* (indicating a personal identification code such as a social security number), and *Hospital*. 2,968 words were manually marked as containing PHI. Their occurrence rates are shown in Figure 1.

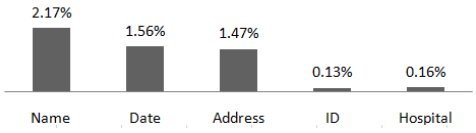

Figure 1: Average frequency per PHI class over total number of words (n=52.829)

For privacy reasons, all PHI words in these documents have been obfuscated in the hospital before we obtained them. The PHI-labelled documents were then reconstructed with fictitious names, addresses, etc. to enable their use as a training set. A test set with 100 randomly selected documents was held internally at the Antwerp University Hospital (UZA) and was manually annotated to be used for later testing (see Section 3.3). However, training and test set differ in quality since the former was manually obfuscated after manual de-identification to protect patient privacy and the latter was unaltered.

2.2 Experimental setup

For the development of our de-identification system using the training set described above, we apply the following experimental setup. First of all, the texts in the dataset are tokenized (i.e. splitting the text in individual words and removing punctuation). Next, features are derived and calculated. In a third step, the resulting set of derived features with associated PHI class per word is used for training. In a set of experiments, we (1) assess the performance of the classifiers for the individual PHI classes, (2) evaluate how adding more training data affects performance, and (3) validate the performance on 100 previously unseen documents. Due to the high cost of manual annotation, our training set is rather small. As a result, the performance scores can only be interpreted as indicative of performance in a realistic environment.

All results in Experiments 1 and 2 are averaged over fifty independent runs, each selecting different sets of training and test sets from the original training set. In each run, ten random documents are withheld as test set. In Experiment 1, the remaining ninety are used for training, while in Experiment 2, a learning curve is constructed, showing the effect of a stepwise (step size=10) increase in training set size.

2.3 Feature engineering

Since the choice of features affects algorithm behavior and performance (Kim et al., 2011; Sibanda, 2006), selecting features that discriminate PHI from non-PHI and are able to indicate the differences between the various PHI classes (Gardner and Xiong, 2008) is crucial. Because of the limited data available for training, external dictionaries are indispensable.

We use four types of features: (i) direct target word characteristics, (ii) pattern matching features, (iii) dictionary features, and (iv) contextual word features. Direct target word characteristics indicate the presence of capitalisation, punctuation, and numbers and includes word length information. Pattern matching features are linked to regular expressions that refer to social security numbers or date patterns. Dictionary features indicate whether the target word is present in a PHI dictionary (i.c. dictionaries of first and last names, streets, cities, hospital names, healthcare institutions, salutations) or whether it is part of a word group that is present in a PHI dictionary. For word groups, we take into account a context of three words to the right of the target word (i.e. sliding window size=4). For computational efficiency, we use a suffix tree algorithm by Ukkonen (1995). Contextual word features indicate whether words in the immediate context (i.e. left context=3, right context=3; sliding window size=7) of the target word have characteristics that might influence the classification of the target word (e.g. punctuation, capitalisation).

2.4 Classification

We apply three classification setups, each offering their own advantages for different data sets, dependent on the data set size, the heterogeneity of the data set, and the total number of classes. We use Weka (Witten and Frank, 2005), a toolkit for machine learning, for classification with Random Forests. For Support Vector Machines (SVM), we use the libSVM (Chang and Lin, 2011) library. In future de-identification experiments, we will evaluate Conditional Random Fields as well.

SVMs calculate an optimal decision boundary between two classes (Chang and Lin, 2011), are powerful with high-dimensional data and promote the use of local context features. For de-identification with several PHI classes, multi-class classification is required. We test (i) a one-versus-one learning scheme (cf. 'OOSVM'), where the binary classifiers distinguish between each pair of classes and (ii) a one-versus-all scheme (cf. 'OMSVM'), where each class is distinguished from the other classes simultaneously. Both schemes apply majority voting with equal weights assigned to each (PHI as well as non-PHI) class.

Random Forests is a machine learning technique that generates multiple random Decision Trees (Breiman, 2001). Each of these trees randomly selects features and assigns a particular class to each instance containing those features. A voting system decides which of these decisions is finally assigned, potentially leading to a more robust decision since it is supported by multiple trees. The total number of trees is customisable, but a high number of trees increases training time. We tested multiple numbers of trees, but selecting ten random trees (cf. RF10) yielded the best balance between precision, recall, and training time.

2.5 Evaluation measures

We present results in terms of precision, recall, and F-score. We consider recall to be the most important measure for de-identification as it shows the number of PHI-items actually retrieved by the algorithm divided by the number of PHI items present. Precision indicates how many of the PHI items identified are actually correct. F-score is calculated as the harmonic mean between precision and recall. Precision and recall are macro-averaged, in a way that all classes have an equal weight in the end result.

3 Results

We present results of three experiments: we (1) evaluate the performance of the proposed method for five PHI classes, (2) perform a learning curve experiment to investigate how performance is affected by increasing training set size, and (3) evaluate the best experimental setup on a previously unseen test set of 100 documents.

3.1 Performance on individual PHI classes

Recall and precision are very similar for most classes, as is shown in Table 1, except for the *Name* and *ID* classes. This can be explained by the wide variety in types of IDs and the larger ambiguity between names and non-PHI words (e.g. 'Vrijdag', the Dutch word for 'Friday', also represents a last name found in libraries with a relatively high

	Name	*Date*	*Address*	*ID*	*Hospital*
OOSVM					
Recall	91.2	95.9	95.6	79.9	95.0
Precision	88.6	98.0	98.2	95.3	98.6
F-score	90.1	96.9	96.8	86.9	96.8
OMSVM					
Recall	91.2	95.8	96.2	77.2	95.4
Precision	88.9	98.0	98.4	95.3	98.6
F-score	90.0	96.8	97.3	85.3	97.0
RF10					
Recall	87.4	95.0	92.5	75.8	75.3
Precision	95.1	98.4	98.5	99.4	97.8
F-score	91.1	96.6	95.4	86.0	85.1

Table 1: Results per PHI class and classification setup

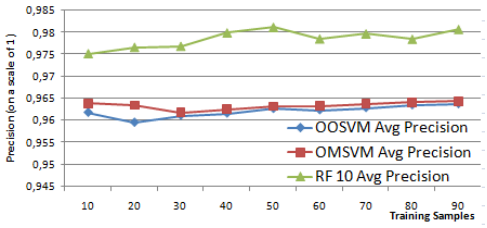

Figure 2: Average precision per setup with 90 training and 10 test documents

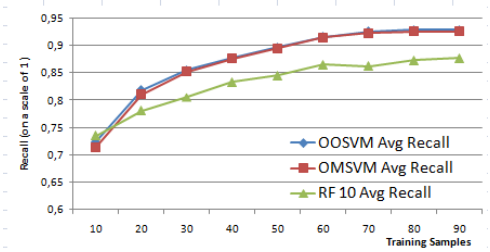

Figure 3: Average recall per setup with 90 training and 10 test documents

frequency). It should be noted that performance is calculated per word in a (potentially multi-word) name. If only part of the name gets classified as a *Name*, it is counted as a false negative, although the largest part of the name will be removed from the text.

Overall, our SVM setups show a higher recall and F-measure than the Random Forest setup, while the latter has a higher precision. With 90 training documents, an average F-measure of 91.5% for the Random Forest method and an average F-measure of 94.5% for the one-against-one SVM setup is achieved.

3.2 Learning curve

To assess the amount of manually annotated data required, we increase the number of training documents in a learning curve experiment. Figures 2 and 3 represent precision and recall scores with varying training set size. The RF10 method has a generally higher precision than the SVM setups, but also a lower recall. Precision remains relatively constant for all methods and recall values seem to converge asymptotically.

3.3 Results on a previously unseen test set

In this experiment, we evaluate the algorithm in a more realistic setting where the algorithm - built from a limited set of manually obfuscated training data (cf. Section 2.1) - is tested on previously unseen test data and evaluated by a hospital staff member. The test data are qualitatively different since they were not subject to obfuscation. Experiments were conducted with OOSVM - the machine learning setup that yielded the best performance in the experiments described above - using 100 obfuscated documents for training and yielded a recall of 89.12% and a precision of 93%, which is lower than the performance on the *obfuscated* test data in Section 3.2.

Error analysis revealed that the use of all-caps (first and last) names and addresses is widespread in the test documents, whereas the training data were manually obfuscated and contained no all-caps names and addresses. Since capitalisation is a feature (cf. Section 2.3) in our de-identification system, the difference in quality between training and test data can explain the drop in performance.

3.4 Time measurements

Time measurements have been taken to check whether the de-identification algorithm is applicable to a larger set of documents. A de-identification speed of 109 ms/document (assuming an average length of 500 words) was achieved when de-identifying with the OOSVM method, while the Random Forest method only needed 42 ms/document. The OMSVM method requires a de-identification time of 205 ms/document.

If OOSVM, the best performing setup, would be used to de-identify documents from the hospitals EHR system on a daily basis, the processing time would be a matter of minutes. The larger amount

of time needed to use the one-against-one SVMs rather than the Random Forest method is worth it, since the performance of the former is significantly better.

4 Discussion

The results suggest that the de-identification algorithm we developed achieves reasonable performance considering the limited set of training data it is based on. However, to be of practical use without manual confirmation, de-identification recall should be as high as possible, making sure that no PHI remains in the text. High precision is of secondary importance, as long as the algorithm does not identify too many non-PHI words as containing Personal Health information, which can cause medically relevant information to be lost during de-identification.

The learning curve experiments show that recall scores start to converge asymptotically, which may indicate that relatively small amounts of training data already yield fair results, while the increase in precision with increasing training set size seems limited. However, we are aware that the data set is too limited to draw conclusions from these results.

The test on a non-obfuscated, previously unseen test set indicates that minor feature improvements and a more representative training set are needed. Although the current approach with a previously manually obfuscated training set is non-scalable, it allows us to automatically create a more representative training set from another dataset.

The results of the Random Forest method can be improved when increasing the amount of trees. However, this also increases training time linearly, with a minimal increase in performance. Recall scores of the current Random Forests setup are insufficient for most PHI classes.

5 Conclusion

In this paper, we presented a machine learning approach to de-identification based on a limited set of manually annotated Dutch-language clinical notes. We compared three types of classification approaches and found the one-versus-one SVM setup to be the method of choice for this particular case study. In terms of recall - which we consider the most crucial evaluation measure for practically usable de-identification - it is better than the Random Forest classifier, which in its turn scores better in terms of de-identification time and precision. Learning curve results seem to indicate that the amount of training data needed converges to an asymptote quite early in the curve.

We plan several extensions to the algorithm: adding syntactic (e.g. part-of-speech tags) and semantic features, investigating the use of semi-supervised learning to automatically increasing the set of training data, and testing Conditional Random Fields for classification. Another next step is the expansion to an ensemble method for two of our classifiers, taking advantage of properties of both classifiers.

6 Acknowledgments

The authors would like to thank Antwerp University Hospital (UZA) - in particular the innovative ICT programme - University of Antwerp, and Artesis Hogeschool Antwerpen for their support.

References

J. Aberdeen, S. Bayer, R. Yeniterzi, B. Wellner, C. Clark, D. Hanauer, B. Malin, and L. Hirschman. 2010. The mitre identification scrubber toolkit: design, training, and assessment. *International journal of medical informatics*, 79(12):849–859.

Belgian Data Protection Act. 1993. Consolidated text of the Belgian law of December 8, 1992 on Privacy Protection in relation to the Processing of Personal Data as modified by the law of December 11, 1998 implementing Directive 95/46/EC.

L. Breiman. 2001. Random forests. *Machine learning*, 45(1):5–32.

C. Chang and C. Lin. 2011. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27.

L. Deleger, K. Molnar, F.i Savova, G.and Xia, T. Lingren, Q. Li, K. Marsolo, A. Jegga, M. Kaiser, and L. Stoutenborough. 2013. Large-scale evaluation of automated clinical note de-identification and its impact on information extraction. *Journal of the American Medical Informatics Association*, 20(1):84–94.

M. Douglass, G.D. Clifford, A. Reisner, G.B. Moody, and R.G. Mark. 2004. Computer-assisted de-identification of free text in the mimic ii database. *Computers in Cardiology*, 29:641–644.

European Data Protection Directive. 1995. Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data.

O. Ferrández, B.R. South, S. Shen, F.J. Friedlin, M.H. Samore, and S.M. Meystre. 2012a. Generalizability and comparison of automatic clinical text de-identification methods and resources. In *Proceedings of the AMIA Annual Symposium*, pages 199–208.

O. Ferrández, B.R. South, S. Shen, and S. Meystre. 2012b. A hybrid stepwise approach for de-identifying person names in clinical documents. In *Proceedings of the 2012 Workshop on Biomedical Natural Language Processing*, pages 65–72. Association for Computational Linguistics.

J. Gardner and L. Xiong. 2008. HIDE: An Integrated System for Health Information DE-identification. In *Proceedings of the 21st IEEE International Symposium on Computer-Based Medical Systems*, pages 254–259.

HIPAA. 1996. Health Insurance Portability and Accountability Act of 1996.

Y. Kim, E. Riloff, and S.M. Meystre. 2011. Improving Classification of Medical Assertions in Clinical Notes. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers*, volume 2, pages 311–316.

J. Lafferty, A. McCallum, and F. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *Proceedings of the Eighteenth International Conference on Machine Learning*, pages 282–289.

S. Meystre, F. Friedlin, B. South, S. Shen, and M. Samore. 2010. Automatic de-identification of textual documents in the electronic health record: a review of recent research. *BMC medical research methodology*, 10:70:1–70:16.

I. Neamatullah, M.M. Douglass, L.H. Lehman, A. Reisner, M. Villarroel, W.J. Long, P. Szolovits, G.B. Moody, R.G. Mark, and G.D. Clifford. 2008. Automated De-Identification of Free-Text Medical Records. *BMC Medical Infomatics and Decision Making*, 8(32):1–17.

Privacy Rule. 2002. Standards for Privacy of Individually Identifiable Health Information: Final Rule. *Federal Register 53181*, 67(157):53181–53273. (codified at 45 CFR 160 and 164).

T.C. Sibanda. 2006. Was the Patient Cured? Understanding Semantic Categories and Their Relationships in Patient Records. Master's thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology.

L. Sweeney. 1996. Replacing Personally-Identifying Information in Medical Records, the Scrub System. In *Proceedings of the AMIA Annual Fall Symposium*, pages 333–337.

E. Ukkonen. 1995. On-line construction of suffix trees. *Algorithmica*, 14(3):249–260.

Ö. Uzuner, T.C. Sibanda, Y. Luo, and P. Szolovits. 2008. A de-identifier for medical discharge summaries. *Artificial intelligence in medicine*, 42(1):13–35.

I.H. Witten and E. Frank. 2005. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, San Francisco, 2 edition.