Making Sense of Violence Risk Predictions Using Clinical Notes

Pablo Mosteiro[1](✉), Emil Rijcken[1,2], Kalliopi Zervanou[2], Uzay Kaymak[2], Floortje Scheepers[3], and Marco Spruit[1]

[1] Utrecht University, Utrecht, The Netherlands
{p.mosteiro,m.r.spruit}@uu.nl

[2] Eindhoven University of Technology, Eindhoven, The Netherlands
{e.f.g.rijcken,k.zervanou,u.kaymak}@tue.nl

[3] University Medical Center Utrecht, Utrecht, The Netherlands
f.e.scheepers-2@umcutrecht.nl

Abstract. Violence risk assessment in psychiatric institutions enables interventions to avoid violence incidents. Clinical notes written by practitioners and available in electronic health records (EHR) are valuable resources that are seldom used to their full potential. Previous studies have attempted to assess violence risk in psychiatric patients using such notes, with acceptable performance. However, they do not explain *why* classification works and how it can be improved. We explore two methods to better understand the quality of a classifier in the context of clinical note analysis: random forests using topic models, and choice of evaluation metric. These methods allow us to understand both our data and our methodology more profoundly, setting up the groundwork for improved models that build upon this understanding. This is particularly important when it comes to the generalizability of evaluated classifiers to new data, a trustworthiness problem that is of great interest due to the increased availability of new data in electronic format.

Keywords: Natural Language Processing · Topic modeling · Electronic Health Records · Interpretability · Document classification · LDA · Random forests

1 Introduction

Two thirds of mental health professionals working in Dutch clinical psychiatry institutions report having been a victim of at least one physical violence incident in their careers [14]. These incidents can have a strong psychological effect on nurses [11], as well as economical consequences [20]. Multiple Violence Risk Assessment (VRA) approaches have been proposed to predict and avoid violence incidents, with some adoption in practice [27]. One common approach is the Brøset Violence Checklist (BVC) [2], a questionnaire used by nurses and psychiatrists to evaluate the likelihood for a patient to become involved in a

Z. Huang et al. (Eds.): HIS 2020, LNCS 12435, pp. 3–14, 2020.
https://doi.org/10.1007/978-3-030-61951-0_1

violence incident. This is a time-consuming and highly subjective process, and automation would be beneficial to the field.

Machine learning methods have been successfully applied to psychiatric Electronic Health Records (EHR) to predict readmission [25]. Most current applications of text processing in psychiatric EHR's are for the English language [12]. Building up on promising first attempts to systematically analyse EHR's in Dutch [17–19], the COVIDA project (COmputing VIsits DAta) aims to create a publicly available self-service facility for Natural Language Processing (NLP) of Dutch medical texts.

In order to build a self-service tool, it is essential to dig deep into the machine learning methods employed and build confidence and trust in practitioners. In this work, we investigate VRA using clinical notes in Dutch and attempt to provide better understanding and interpretable results. For this purpose, we re-implement an SVM document classification approach suggested in [17,19] on a 35% bigger data set; we expand on the text features used by combining text with existing structured data, such as the patient's age; and we experiment with alternative document representation techniques, such as LDA and word embeddings, using random forest classification in order to also gain better insights in potentially significant features. We find promising results, though much work remains to be done to achieve acceptable performance for the clinical practice.

2 Related Work

The analysis of free text in EHR's and the combination of these to structured data using machine learning approaches is gaining an increasing interest as anonymised EHR's become available for research. However, the analysis of clinical free-text data presents numerous challenges due to (i) *highly imbalanced data* with respect to the class of interest [23]; (ii) *lack of publicly available data sets*, limiting research on private institutional data [30]; and (iii) relatively *small data sizes* compared to the amounts of data currently used in text processing research.

In the psychiatric domain, structured data such as symptom codes and medication history have been used for the prediction of admissions [9,15]. Studies using structured information in EHR's to predict suicide risk [6] indicate that information from the unstructured data in clinical texts may provide better insights on risk factors and result in better predictions. Free text in combination with structured EHR variables has been used in suicide [7] and depression diagnosis [10] among healthy and unhealthy individuals. In such approaches, structured variables such as medication history, questionnaires, and demographics are expected to provide enough discriminatory power for the required analysis. Research approaches focusing on text from EHR's in mental healthcare are to our knowledge very few; Poulin *et al.* [21] attempted to predict suicide risk among veterans and more recently Menger *et al.* [19] used Dutch clinical text to predict violent incidents from patients in treatment facilities.

The most popular machine learning methods used for processing free text in psychiatric EHR data are support vector machines (SVM), logistic regression, naive Bayes, and decision trees [1]. Decision-tree classification is one of the most easily interpretable approaches, because it allows for inspection of the specific feature combination used for the classification. This line of classification approaches has also achieved significant improvements in classification accuracy by growing an ensemble of decision trees (a *random forest*) trained on subsets of the data set and letting them vote for the most popular class [4].

3 Data Set

The data used in this study consists of clinical notes written in Dutch by nurses and physicians about patients in the psychiatry ward of the University Medical Center (UMC) Utrecht between 2012-08-01 and 2020-03-01. The 834834 notes available are de-identified for patient privacy using DEDUCE [18].

Each patient can be admitted to the psychiatry ward multiple times. In addition, an admitted patient can spent time in various sub-departments of psychiatry. The time the patient spends in each of the sub-departments is called an *admission period*. In the present study, our data points are admission periods. For each admission period, all notes collected between 28 days before and 1 day after the start of the admission period are concatenated and considered as a single *period note*. If a patient is involved in a violence incident between 1 and 28 days after the start of the admission period, the outcome is recorded as *violent* (hereafter also referred to as *positive*). Otherwise, it is recorded as *non-violent* (*i.e.*, *negative*). Admission periods having period notes with fewer than or equal to 100 words are discarded as was done in previous work [19,25].

In addition to notes, we employ structured variables collected in various formats by the hospital. These include variables related to:

– Admission periods (*e.g.*, start date and time)
– Notes (*e.g.*, date and time of first & last notes in period)
– Patient (*e.g.*, gender, age at the start of the admission period)
– Medications (*e.g.*, numbers prescribed and administered)
– Diagnoses (*e.g.*, presence or absence)

These are included to establish whether some of these variables can be correlated with violence incidents.

The resulting data set consists of 4280 admission periods, corresponding to 2892 unique patients. The data set is highly imbalanced, as a mere 425 admission periods have a violent outcome. In further sections, we will discuss how the imbalanced nature of the data set affects the analysis.

4 Methodology

In this work, we address the problem of violence risk prediction as a document classification task, where EHR document features are combined with additional structured data, as explained in Sect. 3. For text normalization purposes, we perform a series of pre-processing steps outlined in Sect. 4.1. Then, for document representation purposes, we experiment with two alternative approaches—paragraph embeddings and LDA topic vectors—discussed in Sect. 4.2. For the classification task, we experiment with SVM [8] and random forest classification [4] (Sect. 4.3). Finally, we discuss our choice of evaluation metrics in Sect. 4.4.

4.1 Text Normalization

All notes are pre-processed by applying the following normalization steps:

- Converting all period notes to lowercase
- Removing special characters (*e.g.*, ë $\rightarrow$ e)
- Removing non-alphanumeric characters
- Tokenizing the texts using the NLTK Dutch word tokenizer [3]
- Removing stopwords using the default NLTK Dutch stopwords list
- Stemming using the NLTK Dutch snowball stemmer
- Removing periods

4.2 Text Representations

The language used in clinical text is domain-specific, and the notes are rich in technical terms and spelling errors. Pre-trained paragraph embedding models do not necessarily yield useful representations. For this reason, we use the entire available set of 834834 de-identified clinical notes to train both the paragraph embedding model and the topic model. Only notes with at least 10 words each are used, to remove notes that contain no valuable information.

Paragraph Embeddings. We use Doc2Vec [13] to convert texts to paragraph embeddings. The Doc2Vec training parameters are set to the default values in Gensim 3.8.1 [22], with the exception of four parameters: we increase `epochs` from 5 to 20 to improve the probability of convergence; we increase `min_count`—the minimum number of times a word has to appear in the corpus in order to be considered—from 5 to 20 to avoid including repeated mis-spellings of words [19]; we increase `vector_size` from 100 to 300 to enrich the vectors while keeping the training time acceptable; and we decrease `window`—the size of the context window— from 5 to 2 to mitigate the effects of the lack of structure often present in EHR texts.

Topic Modeling. A previous study using Latent Dirichlet Allocation (LDA) for topic modeling in the psychiatry domain [25] suggests that numeric representations of texts obtained by topic modeling can be used alternatively or in addition to text embeddings in classification problems. We use the LdaMallet [22] implementation of LDA to train a topic model on our large set of 834834 clinical notes. In order to determine the optimal number of topics, we use the coherence model implemented in Gensim to compute the coherence metric [29].

We find that using 25 topics maximizes coherence. We use default values for the LdaMallet training parameters. Using the trained LDA topic model we compute, for each of the 4280 period notes in our data set, a 25-dimensional vector of weights, where each dimension represents a topic and the value represents the degree to which this topic is expressed in the note. These vectors are then used as input to the classifiers described below.

4.3 Classification Methods

Similarly to previous work [17,19], we use Support Vector Machines (SVM) [8]. Moreover, for interpretability purposes, we implement in this work random forest classification [4]. Random forests are ensemble models that are widely used in classification problems [31]. The `scikit-learn` implementation of random forests outputs after training a list of the most relevant features used for classification. This can help us determine whether some of the features are more important than others when it comes to classifying positive and negative samples.

We use two loops of 5-fold cross-validation for estimation of uncertainty and hyper-parameter tuning. In each iteration of the outer loop, the admission periods corresponding to 1/5 of the patients are kept as test data, and the remaining admission periods are used in the inner loop to perform a grid search for hyper-parameter tuning. The best classifier from the inner loop is applied to the test data, and the resulting classification metrics from each iteration of the outer loop are used to calculate a mean and a standard deviation for the metrics.

We employ the SVC support-vector classifier provided by `scikit-learn`, with default parameters except for the following: `class_weight` is set to 'balanced' to account for our imbalanced data set; `probability` is True to enable probability estimates for performance evaluation; the cost parameter `C` and the kernel coefficient `gamma` are determined by cross-validation. The ranges of values used are $\mathtt{C} = \{10^{-1}, 10^{0}, 10^{1}\}$ and $\mathtt{gamma} = \{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 10^{0}\}$. Both of these ranges were motivated in a previous study [19].

For the random forest classifier, we use the `scikit-learn` implementation, with default values for all the parameters except for the following: `n_estimators` is increased to 500 to prevent overfitting; `class_weight` is set to 'balanced' to account for the imbalanced data set; and `min_samples_leaf`, `max_features` and `criterion` are determined by cross-validation. Values for `min_samples_leaf` are greater than the default value of 1, to prevent overfitting. For `max_features`, we consider the default value of 'auto', which sets the maximum number of features per split to the square root of the number of features, and two smaller values, again in order to prevent overfitting. Finally, both split criteria available in `scikit-learn` were considered ('gini' and 'entropy'). These parameters are summarized in Table 1.

4.4 Evaluation Metrics

Binary classifiers predict probabilities for input samples to belong to the positive class. When employing a binary classifier in practice, a threshold is chosen, and all samples with positive probabilities above that threshold are considered positive *predictions*. While testing the performance of a classifier, then, we can compare the actual *conditions* with the predictions, and classify each sample as a *true positive* (TP), *true negative* (TN), *false positive* (FP) (condition negative, predicted positive) or *false negative* (FN) (condition positive, predicted negative).

Table 1. Random forest training parameters. Parameters with multiple values are optimized through cross-validation. Parameters not shown are set to default `scikit-learn` values.

Parameter	Value/s	Method
`min_samples_leaf`	{3, 5, 10}	Cross-validation
`max_features`	{5.2, 8.7, 'auto'}	Cross-validation
`criterion`	{'gini', 'entropy'}	Cross-validation
`n_estimators`	500	Fixed
`class_weight`	'balanced'	Fixed

Choosing an operating threshold in practice requires domain expert knowledge. *E.g.*, if violence incidents have very high costs (human or economic), avoiding false negatives would be a priority; if, on the other hand, interventions are costly and cannot be afforded for most patients, avoiding false positives would be more important. Because the decision of the operating threshold is usually made only when the classifier will be put into practice, we report the performance of classifiers using metrics that are agnostic to the operating threshold.

The performance of classifiers is often reported in terms of the Receiver Operating Characteristic Area Under the Curve (ROC-AUC) [17,19]. The ROC is a plot of the true positive rate (TPR) as a function of the false positive rate (FPR), where TPR and FPR are defined as

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}; \ \mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{TN}+\mathrm{FP}} \tag{1}$$

In other words, the curve is constructed by choosing multiple classification thresholds, computing the TP, FP, TN, FN for each threshold, then computing FPR and TPR. As you vary the classification threshold, you allow more or fewer positive predictions, so FPR and TPR both vary in the same direction. In a random classifier, FPR and TPR vary at the same rate, so the baseline ROC is a straight line between (0, 0) and (1, 1), and the baseline ROC-AUC is 0.5. The maximum ROC-AUC is 1, which represents perfect discrimination between TP and FP.

It has been previously noted [26] that ROC-AUC is not a robust performance metric when dealing with imbalanced data sets such as ours. Because the data set is highly imbalanced, the FPR can be misleadingly small, simply because the denominator includes all negative samples, and this artificially increases the ROC-AUC. For this reason, we opt in this work to implement the area under the Precision-Recall curve (PR-AUC) evaluation measure [16]. The Precision-Recall curve is, as its name suggests, a plot of the precision of the classifier as a function of its recall, with precision and recall defined as:

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

Note that neither of these quantities are directly dependent on TN, which is a desirable feature because we have an imbalanced data set with a large number of negatives, and we are more interested in the few positives.

To determine the baseline value for PR-AUC, note that, no matter the recall, the best precision that can be achieved by guessing randomly is the real fraction of positive samples, f_P. Thus, the baseline PR-AUC is f_P, which in our case is $425/4280 = 0.10$.

Though we make no decision regarding the classification threshold, we believe that due to the nature of violence incidents it is more important to avoid FN than to avoid FP. Thus, a good metric to quantify the performance of a classifier in practice is F_2, given by:

$$F_\beta = (1 + \beta^2) \cdot \frac{\text{Precision} \cdot \text{Recall}}{\beta^2 \cdot \text{Precision} + \text{Recall}} \tag{3}$$

with $\beta = 2$ [24].

In this work, we report our classifier performance in both PR-AUC and ROC-AUC, for comparison with previous work on similar data sets [17,19,25,28]. We also report $F_2^{\max}$, *i.e.*, the value of F_2 at the classification threshold that maximizes F_2.

5 Experimental Results and Discussion

5.1 Classifier Performance

Table 2 reports the results of the analyses. All configurations gave results consistent with each other, as well as with previous work on a smaller data set [19]. Figure 1 shows the precision-recall curve for one of the folds of the outer uncertainty-estimation loop during the training of the SVM classifier. These metrics show modest performance, and they indicate that further work is needed to extract all the meaningful information contained in the clinical notes.

Table 2. Classification metrics for various training configurations.

LDA	Embeddings	Structured vars	Estimator	PR-AUC	ROC-AUC	$F_2^{\max}$
No	Yes	No	SVM	0.321 ± 0.067	0.792 ± 0.011	0.519
No	Yes	No	RF	0.293 ± 0.054	0.782 ± 0.011	0.514
No	Yes	Yes	RF	0.299 ± 0.056	0.782 ± 0.011	0.515
Yes	No	Yes	RF	0.309 ± 0.070	0.785 ± 0.011	0.503
Yes	Yes	Yes	RF	0.304 ± 0.058	0.792 ± 0.011	0.517

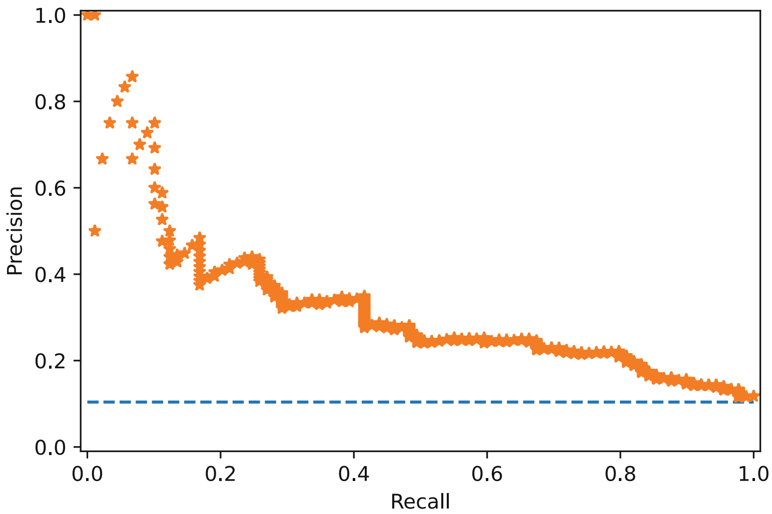

Fig. 1. Precision-Recall curve for one of the folds of the uncertainty-estimation loop during training of the SVM classifier. The PR-AUC is 0.33.

5.2 Feature Importance

When using the random forest estimator, at each step in the outer cross-validation loop we stored the 10 most important features according to the best fit in the inner cross-validation loop. Gathering all the most important features together, we then studied both the 10 most repeated features and the 10 features with the highest total feature importance. These lists were reassuringly similar. The most repeated features were 5 of the text embedding features, plus the age at the beginning of the admission period (`age_admission`) and the number of words in the period note (`num_words`). The frequency distributions of these variables are shown, for both positive and negative samples, on Fig. 2. As can be seen in the figures, the average violent patient is younger than the average non-violent patient, and the average period note about a violent patient is longer than the average period note about a non-violent patient.

The fact that only two of the structured variables included in our study resulted in significant discrimination between the positive and negative classes further stresses that novel sophisticated methods are required.

5.3 Inter-classifier Agreement

To better understand why paragraph embeddings and topic models gave similar classification metrics, we studied the inter-classifier agreement using Cohen's kappa [5]. This metric quantifies how much two classifiers agree, taking into consideration the probability that they agree by chance. A value of Cohen's kappa equal to 0 means the agreement between the two classifiers is random, while a value of 1 means the agreement is perfect and non-random. We placed classification thresholds at multiple points between 0 and 1 (same threshold for both classifiers), and computed classification labels for each classifier for each threshold. Using those classification labels, we computed Cohen's kappa. The result is shown in Fig. 3.

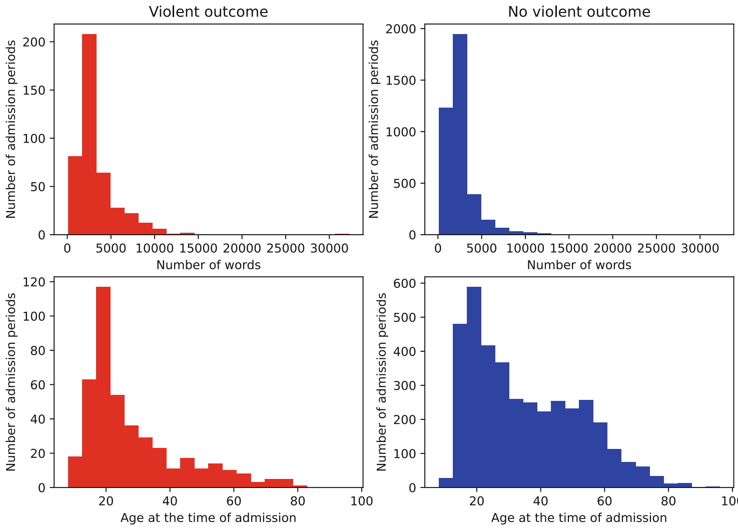

Fig. 2. Histograms of number of words per period note (top) and age (bottom) for violent (left/red) and non-violent (right/blue) patients. (Color figure online)

Next, for each classifier, we used the threshold that maximized the F_2 metric (see Sect. 4.4), and calculated classification labels for each classifier using those thresholds. We then computed Cohen's kappa using those classification labels, and obtained a value of:

$$\kappa = 0.633 \pm 0.012, \tag{4}$$

which is close to the maximum value reported in Fig. 3.

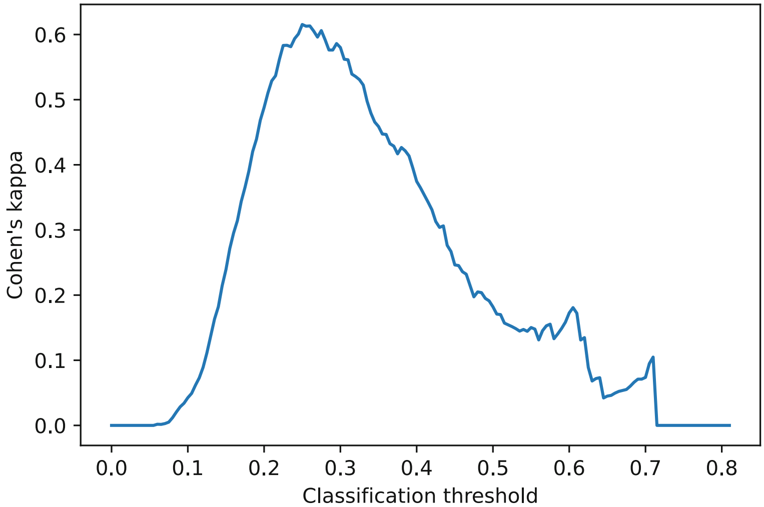

Fig. 3. Cohen's κ score for inter-classifier reliability, comparing the LDA-based classifier and the embeddings-based classifier, as a function of the classification threshold. The same 200 equidistant thresholds between 0 and 1 were used for both classifiers.

According to the standard interpretation of Cohen's kappa, this value implies that the agreement between the two classifiers is better than random, though not close to perfect agreement. This might indicate that the paragraph embeddings and the representations based on LDA are capturing similar information from the notes.

6 Conclusions

We applied machine learning methods to Dutch clinical notes from the psychiatry department of the University Medical Center (UMC) in Utrecht, the Netherlands. We trained a model that predicts, based on the content of the notes, which patients are likely to be involved in a violence incident within their first 4 weeks of admission. The performance of our classifiers is assessed using the area under the precision-recall curve (PR-AUC), and its value is approximately 0.3, well above the baseline value of around 0.1; the ROC-AUC is approximately 0.8. The maximum F_2 score, which puts twice as much importance on recall as on precision, is approximately 0.5. Our results are competitive with a study based on structured variables that obtained ROC-AUC = 0.7801 [28]. These metrics show modest performance, and they indicate that further work is needed to extract all the meaningful information contained in the clinical notes.

The fact that only two of the structured variables included in our study—number of words and patient age—resulted in significant differentiation between the positive and negative classes further stresses that novel sophisticated methods are required. In particular, deep learning is a promising approach.

We have also, for the first time as far as we are aware, applied topic modeling to clinical notes in Dutch language for Violence Risk Assessment. We found that the performance of classifiers on numerical representations produced by topic models is comparable to the performance of similar classifiers on document embeddings. Since topic models are easier to interpret than paragraph embeddings, this is a promising avenue for implementation in practice. However, this is a preliminary study, and more extensive research on larger data sets should be performed to confirm our findings. This is the direction we will follow for future research.

References

1. Abbe, A., Grouin, C., Zweigenbaum, P., Falissard, B.: Text mining applications in psychiatry: a systematic literature review. MPR **25**(2), 86–100 (2016)
2. Almvik, R., Woods, P., Rasmussen, K.: The Brøset violence checklist: sensitivity, specificity, and interrater reliability. J. Interpers. Violence **15**(12), 1284–1296 (2000)
3. Bird, S., Klein, E., Loper, E.: Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit. https://www.nltk.org/book/
4. Breiman, L.: Random forests. Mach. Learn. **45**, 5–32 (2001)
5. Cohen, J.: A coefficient of agreement for nominal scales. Educ. Psychol. Meas. **20**(1), 37–46 (1960)
6. Conner, K.R., et al.: Mental disorder comorbidity and suicide among 2.96 million men receiving care in the veterans health administration health system. J. Abnorm. Psychol. **122**(1), 256–263 (2012)
7. Cook, B.L., Progovac, A.M., Chen, P., Mullin, B., Hou, S., Baca-Garcia, E.: Novel use of natural language processing (NLP) to predict suicidal ideation and psychiatric symptoms in a text-based mental health intervention in Madrid. Comput. Math. Methods Med. **2016**, 8708434 (2016)
8. Cortes, C., Vapnik, V.: Support-vector networks. Mach. Learn. **20**(3), 273–297 (1995)
9. Friedman, S., Margolis, R., David, O.J., Kesselman, M.: Predicting psychiatric admission from an emergency room. J. Nerv. Ment. Dis. **171**(3), 155–158 (1983)
10. Huang, S.H., LePendu, P., Iyer, S.V., Tai-Seale, M., Carrell, D., Shah, N.H.: Toward personalizing treatment for depression: predicting diagnosis and severity. JAMIA **21**(6), 1069–1075 (2014)
11. Inoue, M., Tsukano, K., Muraoka, M., Kaneko, F., Okamura, H.: Psychological impact of verbal abuse and violence by patients on nurses working in psychiatric departments. Psychiatry Clin. Neurosci. **60**, 29–36 (2006)
12. Kim, Y.K. (ed.): Frontiers in Psychiatry: Artificial Intelligence, Precision Medicine, and Other Paradigm Shifts. Springer, Singapore (2019). https://doi.org/10.1007/978-981-32-9721-0
13. Le, Q., Mikolov, T.: Distributed representations of sentences and documents. In: ICML 2014, Beijing, China, pp. 1188–1196. PMLR (2014)
14. van Leeuwen, M., Harte, J.: Violence against mental health care professionals: prevalence, nature and consequences. J. Forensic Psychiatry Psychol. **28**(5), 581–598 (2017)

15. Lyons, J.S., Stutesman, J., Neme, J., Vessey, J.T., O'Mahoney, M.T., Camper, H.J.: Predicting psychiatric emergency admissions and hospital outcome. Medical care **35**(8), 792–800 (1997)
16. Manning, C., Raghavan, P., Schütze, H.: Introduction to Information Retrieval. Cambridge University Press, Cambridge (2008)
17. Menger, V., Scheepers, F., Spruit, M.: Comparing deep learning and classical machine learning approaches for predicting inpatient violence incidents from clinical text. Appl. Sci. **8**(6), 981 (2018)
18. Menger, V., Scheepers, F., van Wijk, L., Spruit, M.: DEDUCE: a pattern matching method for automatic de-identification of Dutch medical text. Telemat. Inform. **35**(4), 727–736 (2018)
19. Menger, V., Spruit, M., van Est, R., Nap, E., Scheepers, F.: Machine learning approach to inpatient violence risk assessment using routinely collected clinical notes in electronic health records. JAMA Netw. Open **2**(7), e196709 (2019)
20. Nijman, H., Bowers, L., Oud, N., Jansen, G.: Psychiatric nurses' experiences with inpatient aggression. Aggress. Behav. **31**(3), 217–227 (2005)
21. Poulin, C., et al.: Predicting the risk of suicide by analyzing the text of clinical notes. PLoS ONE **9**(3), e91602 (2014)
22. Řehůřek, R., Sojka, P.: Software framework for topic modelling with large corpora. In: Proceedings of LREC 2010 Workshop on New Challenges for NLP Frameworks, pp. 45–50. University of Malta, Valletta, Malta (2010)
23. Rijo, R., Martinho, R., Pereira, L., Silva, C.: Text mining applied to electronic medical records. Int. J. E-Health Med. Commun. **6**(3), 1–18 (2015)
24. van Rijsbergen, C.J.: Information Retrieval, 2nd edn. Butterworth-Heinemann, Oxford (1979)
25. Rumshisky, A., et al.: Predicting early psychiatric readmission with natural language processing of narrative discharge summaries. Transl. Psychiatry **6**, e921 (2016)
26. Saito, T., Rehmsmeier, M.: The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS ONE **10**(3), e0118432 (2015)
27. Singh, J.P., et al.: International perspectives on the practical application of violence risk assessment: a global survey of 44 countries. Int. J. Forensic Ment. Health **13**(3), 193–206 (2014)
28. Suchting, R., Green, C.E., Glazier, S.M., Lane, S.D.: A data science approach to predicting patient aggressive events in a psychiatric hospital. Psychiatry Res. **268**, 217–222 (2018)
29. Syed, S., Spruit, M.R.: Full-text or abstract? Examining topic coherence scores using latent Dirichlet allocation. In: DSAA2017, pp. 165–174 (2017)
30. Wang, Y., Wang, L., Rastegar-Mojarad, M., Moon, S., Shen, F., Afzal, N., Liu, S., Zeng, Y., Mehrabi, S., Sohn, S., Liu, H.: Clinical information extraction applications: a literature review. J. Biomed. Inform. **77**, 34–49 (2018)
31. Yin, H., Camacho, D., Tino, P., Tallón-Ballesteros, A.J., Menezes, R., Allmendinger, R. (eds.): IDEAL 2019. LNCS, vol. 11872. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-33617-2