Research Article

Machine Learning for Violence Risk Assessment Using Dutch Clinical Notes

Pablo Mosteiro[1,*], Emil Rijcken[2,1], Kalliopi Zervanou[2], Uzay Kaymak[2], Floortje Scheepers[3], Marco Spruit[1,4,5]

[1]*Utrecht University, Utrecht, The Netherlands*

[2]*Eindhoven University of Technology, Eindhoven, The Netherlands*

[3]*University Medical Center Utrecht, Utrecht, The Netherlands*

[4]*Leiden University Medical Center, Leiden, The Netherlands*

[5]*Leiden Institute of Advanced Computer Science, Leiden, The Netherlands*

ARTICLE INFO

ABSTRACT

Violence risk assessment in psychiatric institutions enables interventions to avoid violence incidents. Clinical notes written by practitioners and available in electronic health records are valuable resources capturing unique information, but are seldom used to their full potential. We explore conventional and deep machine learning methods to assess violence risk in psychiatric patients using practitioner notes. The performance of our best models is comparable to the currently used questionnaire-based method, with an area under the Receiver Operating Characteristic curve of approximately 0.8. We find that the deep-learning model BERTje performs worse than conventional machine learning methods. We also evaluate our data and our classifiers to understand the performance of our models better. This is particularly important for the applicability of evaluated classifiers to new data, and is also of great interest to practitioners, due to the increased availability of new data in electronic format.

1. INTRODUCTION

Two-thirds of mental health professionals working in Dutch clinical psychiatry institutions report having been a victim of at least one incident of physical violence in their careers [1]. These incidents can have a strong psychological effect on practitioners [2], as well as economic consequences for health institutions [3]. Multiple Violence Risk Assessment (VRA) approaches have been proposed to predict and avoid violence incidents, with some adoption in practice [4]. Traditionally, a common approach is the Brøset Violence Checklist (BVC) [5], a questionnaire used by nurses and psychiatrists to evaluate the likelihood for a patient to become involved in a violence incident. Filling out the form is a time-consuming and highly subjective process.

Machine learning methods might help improve this process by saving time and making predictions more accurate. Electronic Health Records (EHR) are a rich source of information containing both structured fields as well as written notes. EHR notes coupled with violence incident reports can be used to train machine learning models to classify notes as describing potentially violent patients.

Indeed, machine learning approaches trained on English-language psychiatric notes have shown promising results with values of the Receiver Operating Characteristic (ROC) Area Under the Curve (AUC) of 0.85 and higher [6–10].

Nevertheless, violence prediction based on Dutch written notes appears to be a challenging endeavor as no efforts have shown satisfying results, with the AUC stagnating below 0.8 [11–13]. In those papers, various machine learning methods were applied, including bag-of-words, document embeddings, and topic modeling to generate numerical representations of texts; and support vector machines (SVM) and random forests for classification. In this work, we explore a new approach to VRA with Dutch clinical notes using the deep-learning language model Bidirectional Encoder Representations from Transformers (BERT) [14], which has proven to deliver very good results in multiple languages and domains [15,16]. To the best of our knowledge, BERT has not been used for VRA in Dutch before.

The paper is organized as follows. Section 2 is a review of the related work in text analysis for VRA. Sections 3 and 4 describe our dataset and methodology, respectively. Afterwards, we present our results in Section 5, and we discuss our findings in Section 6. Finally, we present some conclusions in Section 7.

*Corresponding author. Email: p.mosteiro@uu.nl

2. TEXT ANALYSIS FOR VRA

The analysis of free text in EHR combined with structured data using machine learning approaches is gaining interest as anonymized EHR data become available for research [7,17,18]. However, the analysis of clinical free-text data presents numerous challenges due to (i) *highly imbalanced data* with respect to the class of interest [19]; (ii) *lack of publicly available datasets*, limiting research on private institutional data [20] 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 to predict admissions [21,22]. In combination with structured EHR variables, free text has been used in suicide prediction [23]; depression diagnosis [24]; and harm risk in an inpatient forensic psychiatry setting [25].

To our knowledge, there are few research approaches focusing on predicting violence in mental healthcare based on Dutch-language text from EHRs. Menger *et al.* [12] use Dutch clinical text to predict violence incidents from patients in treatment facilities. In Mosteiro *et al.* [13] we compared several classical machine learning methods for VRA of EHR notes, including Latent Dirichlet Allocation (LDA) for topic modeling, and we discussed the agreement between some of those classifiers. In this work we extend on that approach, introducing the BVC as a baseline and employing BERT for document classification.

The general pipeline for violence prediction using Natural Language Processing (NLP) is similar in all approaches. Firstly, the notes are represented in a numerical way. Secondly, the numerical representation is the input to a machine learning algorithm that is trained to perform predictions. Van Le *et al.* [25] use the presence/absence of predefined words as features to feed to several machine learning algorithms. Menger *et al.* [11] experiment with different representations such as bag-of-words, word2vec [26], and paragraph2vec [27]. Bag-of-words represents documents as vectors with a size equal to the dataset vocabulary length, encoding the presence/absence of a word. The bag-of-words representation is a sparse matrix disregarding word order and meaning. The word2vec algorithm learns word embeddings as dense vectors in high-dimensional space and takes the contexts of words into consideration during training, thus mitigating some of the limitations of the bag-of-words representations. Yet word vectors in word2vec are context-independent, meaning that there is only one vector representation for each word, not taking into account homonyms. The paragraph2vec algorithm is an extension of word2vec that produces a vector representation for an entire paragraph.

In contrast to word2vec and paragraph2vec, BERT produces a vector representation for each word, taking context into account with higher granularity. By doing so, BERT has improved the state-of-the-art for many NLP tasks [14].

Another method for representing documents is based on topic modeling. Topic modeling assumes that a corpus consists of a collection of topics and that each topic consists of a collection of words. After choosing the number of topics to retrieve, the algorithm finds words that contribute most to each topic. A document can then be expressed as a vector, indicating to which extent each topic is represented in that document. A study in the psychiatry domain [28] shows promising results when representing documents using topic models through LDA. A potential advantage of using topic modeling is that the different topics may help explain how the classification model makes its decisions. In this paper, we experiment with a model based on LDA topic modeling.

When comparing different models and ranking them, the selection of evaluation metric plays a significant role. One of the most commonly used scalars for ranking model performance is the area under the ROC curve, typically referred to as AUC [29]. Different points on the ROC correspond to different operating points of a classifier obtained by thresholding an underlying continuous output that the classifier computes, leading to different false-positive and true-positive rates. AUC, a scalar measure that takes multiple thresholds into account, is better than accuracy for evaluating the overall classifier performance and discriminating an optimal solution [30]. AUC is independent of class priors and it ignores misclassification costs. In our domain, misclassification costs can be asymmetric, as an unnecessary intervention might be less problematic than an unforeseen violence incident. Hence, it is useful to consider also other performance metrics.

In this paper, we evaluate two alternative metrics: Area Under the Precision-Recall Curve (AUPRC) and Area Under the Kappa curve (AUK). The AUPRC is similar to the AUC, but the axes are precision and recall, which are metrics used more commonly in text classification. AUPRC ignores true negatives, which is desirable when evaluating models on datasets where positives are the minority class and predicting positives correctly is the priority [31]. The AUK [32] is based upon Cohen's Kappa [33]. AUK corrects the accuracy of a model for chance agreements, and it is a nonlinear transformation of the difference between a model's AUC value and the AUC of a random model. The main difference between the AUC and AUK is that AUK accounts for class skewness, while AUC does not. Therefore, the AUK has desirable properties when there is considerable class skew.

3. DATA

The data used in this study consists of clinical notes written in Dutch by nurses and physicians during visits with patients from the psychiatry ward of the University Medical Center (UMC) Utrecht between 2012-08-01 and 2020-03-01. The 835k notes available are anonymized for patient privacy using DEDUCE [34]. The study was reviewed and approved by the UMC ethics committee.

A patient can be admitted to the psychiatry ward multiple times. Additionally, an admitted patient can spend time in various sub-wards of psychiatry. The time the patient spends in each of the sub-wards is called an *admission period*. In the present study, our data points are admission periods. All notes collected between 28 days before and 1 day after the beginning of the admission period are concatenated and considered a single *period note* for each admission period.[1] If a patient is involved in a violence incident between 1 and 28 days after the beginning of the admission period, the outcome is

---

[1]Notes are available from before the beginning of the admission period for patients that have spent time in other wards or sub-wards, or other institutions. For patients transferred within the past 5 days, only notes from 7 or fewer days before the beginning of the admission period are included.

recorded as *violent* (*positive*). Otherwise, it is recorded as *nonviolent* (*negative*). Admission periods having period notes with fewer than 100 words are discarded as in previous work [12,28].

In addition to the notes, we employ structured variables collected by the hospital and related to:

- admission periods (e.g., start date and time);
- notes (e.g., date and time of first and last notes in the period);
- patient (e.g., gender, age at the beginning of the admission period);
- medications (e.g., numbers prescribed and administered);
- diagnoses (e.g., presence or absence).

These variables are included to establish whether they correlate with violence incidents.

The resulting dataset consists of 4280 admission periods, corresponding to 2892 unique patients. The dataset is highly imbalanced, as a mere 425 admission periods have a violent outcome.

4. METHODOLOGY

In this work, we address the problem of VRA as a document classification task, where EHR document features are combined with additional structured data, as explained in Section 3. Section 4.1 describes our baseline, the BVC. In Sections 4.2 and 4.3 we outline our conventional machine learning and BERT approaches to VRA, respectively. We report all our results in three performance metrics: AUPRC, AUC, and AUK as explained in Section 2. We use AUPRC during development for hyperparameter tuning, while we calculate all three metrics to report the final results.

4.1. Brøset Violence Checklist

At the UMC Utrecht, where our data were collected, VRA is currently done using the BVC [5], a questionnaire that patients answer at the time of admission and approximately every 1-2 weeks during the remainder of the admission. The checklist provides a score from 0 to 6 that estimates the patient's propensity of becoming involved in a violence incident within the next 24 hours.

We use the performance of the BVC as a baseline to compare our machine learning models. The questionnaire was not filled out for all admission periods in our dataset. Therefore, we cross-referenced all the BVC scores in the psychiatry dataset with the violence dataset by patient ID, and took each BVC score as a data point. The resulting dataset has 11799 data points. The independent variable is the BVC score, and the target variable is the presence or absence of a violence incident within 27 days after the BVC was answered. This mimics the situation in our main analysis, where we consider violence incidents happening between 1 and 28 days after the beginning of the admission period. However, we do not know the time at which the BVC was filled out. Therefore, we run the analysis twice, once assuming it was filled at 0:00 hours, and another assuming it was filled at the end of the day, at 23:59.

Note that the BVC dataset is larger than the one we used in our machine learning VRA analyzes, and its class imbalance is different,

with about 1 positive for every 4 negatives (as compared to about 1/10 in the VRA dataset). The implications this has on the interpretation of our results is discussed in Section 5.1.

4.2. Machine Learning Analysis

Text preparation All notes are preprocessed by applying the following normalization steps:

- converting all period notes to lowercase;
- removing special characters (e.g., ë → e);
- removing non-alphanumeric characters;
- tokenizing the texts using the NLTK Dutch word tokenizer [35];
- removing stopwords using the default NLTK Dutch stopwords list;
- stemming using the NLTK Dutch Snowball stemmer;
- removing full stops (".").

Text representations The language used in clinical text is domain-specific, and the notes are rich in technical terms and spelling errors. Pretrained paragraph embedding models, trained on out-of-domain data, do not necessarily yield useful representations. For this reason, we use our dataset to produce numeric representations for our notes. We use the entire set of 835k anonymized clinical notes to train both the paragraph embedding and topic models. Only notes with at least 10 words are used in order to increase the likelihood that each note contains valuable information.

Paragraph embeddings—We use Doc2Vec to convert texts to paragraph embeddings.[2] The Doc2Vec training parameters are set to the default Gensim 3.8.1 values [36], except for 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 to be considered—from 5 to 20 to avoid including repeated misspellings of words [12]; 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 1 to mitigate the effects of the lack of structure often present in EHR texts.

Topic modeling—We use the LdaMallet [36] implementation of LDA to train a topic model. To determine the optimal number of topics, we use the coherence model implemented in Gensim to compute the coherence metric [37]. 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 dataset, 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.

Classification methods We use two classification methods: SVM [38] and random forest classifiers [39]. For trained random forests, scikit-learn outputs a list of the most relevant classification features. This list can help us distinguish which features are most correlated with violence.

---

[2]Doc2Vec is the name of the Gensim implementation of paragraph2vec.

We use two loops of 5-fold cross-validation for the estimation of uncertainty and hyperparameter tuning, as shown on Figure 1. In each iteration of the outer loop, the admission periods corresponding to 20% of the patients are kept as test data. The admission periods corresponding to the remaining 80% are the development set. We then perform 5-fold cross-validation on the development set for hyperparameter tuning; in each iteration, 20% of patients become the validation set, with the remaining 80% as the training set. The combination of hyperparameters that maximizes the average AUPRC on the validation sets in the inner loop is kept. The model is then retrained on the entire development set using the hyperparameters found in the inner loop. Lastly, the classifier from each iteration of the outer loop is applied to the corresponding test set; we report the average and standard deviation across the outer loop for our three classification metrics.

For SVM, we employ the support-vector classifier provided by scikit-learn, with default parameters except for the following:

- "class_weight" is set to "balanced", which uses the truth labels to adjust weights inversely proportional to class frequencies, to account for our imbalanced dataset;

- "probability" is True to enable probability estimates of performance;

- the cost parameter "C" and the kernel coefficient "gamma" are determined by cross-validation.

The ranges of values used are "C" = $\{10^{-1}, 10^0, 10^1\}$ and "gamma" = $\{10^{-5}, 10^{-4}, 10^{-3}, 10^2, 10^{-1}, 10^0\}$, as motivated in a previous study [12].

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", which uses the truth labels to adjust weights inversely proportional to class frequencies, to account for our imbalanced dataset;

- "min_samples_leaf", "max_features" and "criterion" are determined by cross-validation.

The ranges of values used for the parameters determined by cross-validation are reported on Table 1. Values for "min_samples_leaf" are greater than the default value of 1, to prevent overfitting; we consider the default value of "auto" for "max_features", which sets the maximum number of features per split to the square root of the number of features, and two smaller values, again to prevent overfitting; both split criteria available in scikit-learn are considered.

4.3. BERT Analysis

We use the implementation of BERT for sequence classification included in Huggingface Transformers [40], which includes various choices of pretrained models. We explored Multilingual BERT [41], from Google Research, as well as BERTje [42].[3] During an exploratory analysis of the 110kDBRD Dutch book-review dataset [43], where we attempted to predict the sentiment of book reviews, BERTje outperformed Multilingual BERT. Therefore, we decided to continue using BERTje.

We tokenize all texts using the pretrained BERTje tokenizer provided by Huggingface Transformers. For classification, a linear layer is added to the BERTje language model. The optimizer used is AdamW, which is recommended by the BERT developers [14]. We use a scheduler that increases the learning rate linearly from 0 to the value set in the optimizer during a warm-up period, and then decreases linearly back to 0. The warm-up period consists of 10% of the training set. To choose this number, we ran an exploratory analysis on a dataset of Dutch-language tweets [44]. We experimented with warm-up periods of 5%, 10%, and 20%, and found the best results for 10%. Furthermore, we found that a good choice for the learning rate value set in the optimizer was 2×10^{-5}. We use the Transformers default dropout probability of 0.1. Finally, we set the "epsilon" parameter in the AdamW optimizer to 10^{-8}.

As BERTje can only handle up to 512 tokens in a sequence, we experiment with two shortening strategies to shorten our texts:

1. keeping the last 512 tokens in each period note (*truncate*);

2. summarizing texts using Gensim TextRank (*summarize*).

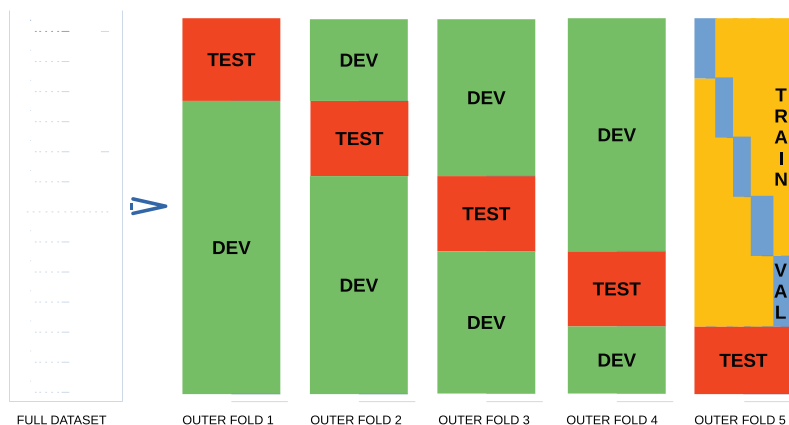[3]In Huggingface Transformers, BERTje is called BERT-base-Dutch-cased.

Figure 1 | Schematic representation of the data processing pipeline for hyperparameter tuning and statistical uncertainty estimation.

Because tokenization in BERTje splits words into its component morphemes, texts have more tokens than words. Thus, when summarizing texts with Gensim, we choose the summarized length to be shorter than 512 words, so that the tokenized texts have lengths shorter than 512 tokens.

We use two loops of 5-fold cross-validation, as explained in Section 4.2 and Figure 1, for uncertainty estimation and to tune the batch size; we explore batch sizes of 16, 32, 64, 128, and 256. The optimal number of epochs is chosen as the number of epochs after which the validation loss increases while the training loss decreases; we run all training for a maximum of 4 epochs.

Finally, because Transformers uses the cross-entropy loss function, which is symmetric to positive and negative samples, we rebalance the training and validation sets before training the model and computing the loss. The performance metrics are computed based on the original (imbalanced) dataset.

5. EXPERIMENTAL RESULTS

5.1. Brøset Violence Checklist

The predictive power of the BVC for VRA within 27 days of collecting the questionnaire is shown on Figure 2 and Table 2. Because the

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

BVC returns integer predictions from 0 to 6, the ROC, the precision-recall curve, and the kappa curve all have a very small number of points. Thus, the areas under the curves depend heavily on the integration method. We report our central values with integration following the trapezoidal rule. We used the left- and right-hand rules to estimate the uncertainty due to the choice of the integration method.

Note that among the three metrics used, only the AUC can be used for comparison between the BVC and the machine learning models, as the class imbalance of the BVC dataset differs from the distribution of the data based on which the machine learning models are trained (see Section 6 for discussion).

5.2. Classifier Performance

Table 3 reports the results of the machine learning models not using BERTje. All configurations gave results consistent with each other, as well as with previous work on a smaller dataset [12], and with the BVC. 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 results of the BERTje analysis are shown on Table 4. The optimal number of epochs was 1 for all configurations. After that, the validation loss increased quickly, while the training loss continued to decrease. This is a sign of overfitting, and we will discuss it further in Section 6. For comparison, we also trained the models for a second epoch and computed the test metrics again.

Table 2 | Predictive power of the Brøset Violence Checklist (BVC). The uncertainty is given by the choice of the integration method.

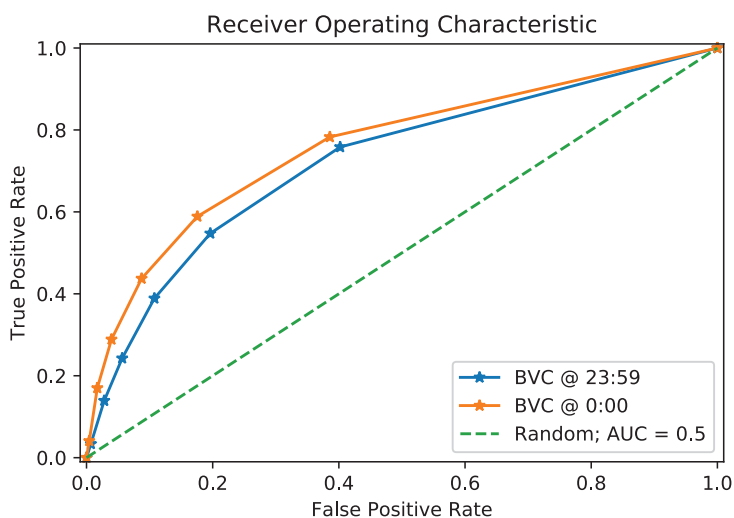| Assumed Filled Hour	AUC	AUPRC	AUK
00:00	0.761 ± 0.100	0.510 ± 0.050	0.207 ± 0.090
23:59	0.725 ± 0.107	0.391 ± 0.041	0.166 ± 0.069

Figure 2 | Receiver Operating Characteristic (ROC) of the Brøset Violence Checklist (BVC) to predict violence incidents within 27 days after filling out the BVC questionnaire. The orange line assumes the BVC was filled out at the beginning of the day; the blue line assumes the BVC was filled out at the end of the day. The dotted green line corresponds to random guessing.

Table 3 | Classification metrics for various training configurations. The statistical uncertainties were estimated via 5-fold cross-validation.

Text Representation	Classifier	AUC	AUPRC	AUK
doc2vec	SVM	0.792 ± 0.034	0.315 ± 0.075	0.136 ± 0.023
doc2vec	RF	0.782 ± 0.030	0.287 ± 0.061	0.130 ± 0.020
doc2vec + struct	RF	0.777 ± 0.026	0.292 ± 0.053	0.129 ± 0.017
LDA + struct	RF	0.785 ± 0.038	0.303 ± 0.079	0.133 ± 0.025
doc2vec + LDA + struct	RF	0.792 ± 0.035	0.298 ± 0.066	0.136 ± 0.022

Table 4 | Results of the Violence Risk Assessment (VRA) analysis using BERTje. The statistical uncertainties were estimated via 5-fold cross-validation.

Shortening Strategy	Trained Epochs	AUC	AUPRC	AUK
Summarize	1	0.664 ± 0.029	0.182 ± 0.049	0.071 ± 0.019
Summarize	2	0.658 ± 0.033	0.184 ± 0.051	0.069 ± 0.019
Truncate	1	0.667 ± 0.027	0.195 ± 0.041	0.074 ± 0.017
Truncate	2	0.657 ± 0.039	0.187 ± 0.045	0.070 ± 0.023

6. DISCUSSION

In this work we analyzed how conventional machine learning methods and BERTje performed on VRA. Although, typically, BERT-like models show high performance on similar tasks, we did not find this to be the case with our dataset. BERTje performed significantly worse than the other machine learning models. None of our models reached performance levels similar to the literature based on English notes. This is a surprising finding, given that we have used more advanced techniques.

In this section we investigate our results further. Firstly, we compare results from the different performance metrics. Secondly, we dive deeper into the dataset by studying the type-token ratio (TTR). Next, we examine some problems with our dataset. Finally, we study the feature importance in the random forest classifiers. We end this section with a discussion of the limitations of our work.

6.1. Performance Metrics

In this work, we have used three different performance metrics: AUC, AUPRC, and AUK. The AUC is the most widely used metric; since it is invariant under a change in the class imbalance, it gives an idea of the classifier performance regardless of the precise dataset used. However, when choosing among multiple models applied on the same dataset, it is a poor metric because it does not account for the possibility that positive- and negative misclassification might have different costs. The AUPRC and AUK attempt to account for this; the former does so by ignoring true negatives, the latter by accounting for chance agreements between the classifier and the truth labels.

Looking at Tables 3 and 4, we can see that all models within each table performed consistently with each other in all metrics. The relative uncertainty is smallest for AUC.

An advantage of the AUK is that the kappa curve not only provides a performance metric but also suggests an optimal operating point. Figure 3 shows the kappa curve for one of the iterations of the SVM classifier. Because the curve is concave down, we can choose the operating threshold that maximizes Cohen's kappa—and thus the performance after accounting for chance agreements—versus

the false-positive rate. In the figure we can see that there is a wide domain of false-positive rates that give values of Cohen's kappa close to the maximum. In practice, psychiatrists would also inform the choice of what the operating threshold should be, based on the costs of misclassifying positives and negatives.

6.2. Type-Token Ratio

As mentioned in Section 5.2, there is evidence that the BERTje models overfit the training data after 2 or more training epochs. To explore why our data might be prone to overfitting, we looked at the TTR for this dataset and compared it to that of the 110kDBRD dataset [43], consisting of book reviews written in Dutch (see Section 4.3). The TTR is the ratio of the number of distinct words (types) in a corpus to the total number of words (tokens). The higher the TTR, the more varied the vocabulary. A varied vocabulary could be associated with overfitting, as the models can exploit individual rare words to differentiate period notes from each other. We computed the TTR both on the words and on the tokens. Tokenization was done by the pretrained BERTje tokenizer used in Section 4.3.

The results are shown on Table 5. The raw texts from our VRA dataset seem to have a less rich vocabulary than the 110kDBRD dataset; however, after summarization the vocabulary is more varied than before. This could mean that the summarization algorithm picks the most distinctive words and removes very repetitive words. More interestingly, after tokenization both the full texts *and* the summarized texts from the VRA dataset have similar TTRs, while the 110kDBRD data has a much lower TTR. It is expected that tokenization will reduce the TTR, because the tokenizer merges several forms of the same word into its root morphemes. But the effect on the VRA dataset is smaller than on the 110kDBRD dataset. The BERTje tokenizer often splits unknown words into component letters that can have morphemic meaning under some contexts. If our dataset contains more idiosyncratic words that the tokenizer does not know about, it might split them into multiple morphemes that are not really what the word is made up of. This would then lead to classification errors, as the text is interpreted to mean something it does not.
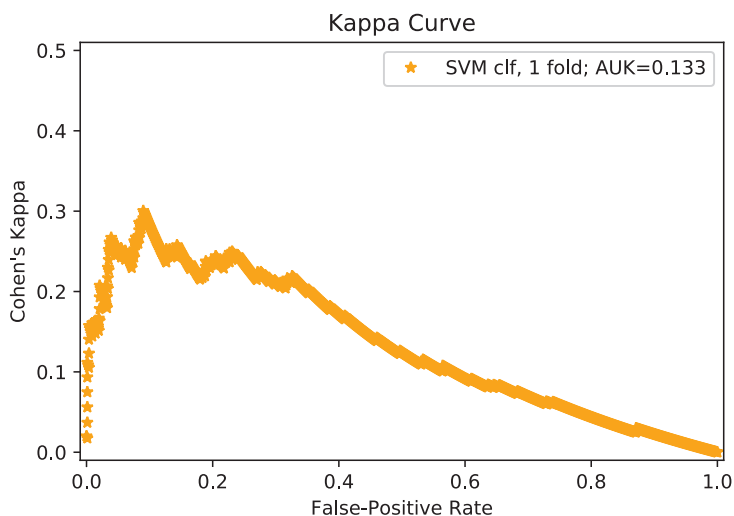
Figure 3 │ Kappa curve for one of the iterations of the outer cross-validation loop of our support vector machines (SVM) classifier. The maximum value in this graph can be interpreted as the optimal operating point of the classifier. Random guessing results in Cohen's kappa equal to 0 for all false-positive rates.

Table 5 │ The Type-Token Ratio (TTR) for our dataset, compared to the book-review dataset 110kDBRD. When tokenizing, only the first 512 tokens were considered. The five most common tokens are also reported.

Dataset	TTR (Raw)	TTR (Tokenized)	Most Common Tokens
VRA (Summarized)	0.045	0.0069	[PAD], ', ', ',', [UNK], *en*
VRA (Full)	0.025	0.0076	', [UNK], ',', '-', *en*
110kDBRD	0.052	0.0033	[PAD], ', *de*, ',', *het*

On Table 5 we see that a large number of raw words in our VRA dataset are interpreted as the [UNK] (unknown) token. Some such words were <, >, *cq*, # and +. This could mean that a lot of words are either missed by the tokenizer (incorrectly classified as unknown, with their meaning getting lost), or included unnecessarily (< and > are used at the beginning and end of anonymized institution and person names but carry no meaning). This could contribute to the bad performance of the classifier.

Upon inspecting the cross-validation loss, we found that, after two epochs of training, overfitting was apparent and yet the performance metrics did not significantly worsen. Thus, although overfitting can be a problem with our dataset, it cannot be the reason why our BERTje results are worse than those obtained using simpler machine learning methods.

6.3. Data Problems

While exploring our data further, we found that some violence incidents described in practitioner notes were not reported in the violence incidents dataset. To assess the significance of this, we examined all practitioner notes collected between 1 and 28 days after the beginning of every admission period. We curated a list of words related to violence incidents.[4] We visually examined some sentences from negative admission periods containing at least one mention of a word from the list and found that other important

---
[4]An expert psychiatrist curated this list.

Table 6 │ Numbers of negative admission periods for which at least one practitioner note collected between 1 and 27 days after the first day of admission contains a keyword associated with violence incidents.

Word	English	Exact	Admission Periods
trap	kick	no	472
geslagen	hit (participle)	yes	415
slaat	hits (3rd person)	yes	415
slaan	hit (infinitive)	yes	414
sloeg	hit (past)	yes	408
duw	push	no	343
schop	kick	no	299
bijt	bite	no	240

words had to be added to the list. The final list of keywords chosen is shown in Table 6.

Simply searching for a keyword can lead to too many false positives, because some words have multiple meanings. This is especially true in Dutch, where certain phrasal verbs have completely different meanings than their root verbs (e.g., *slaan* (hit) vs *omslaan* (change)). Hence, we created a list of exceptions. If the word was part of a larger expression in the list of exceptions, we removed it from the list of candidate unreported incidents. We also removed unrelated longer words that contain our keywords (e.g., *weduwe* (widow) contains *duw*, but is unrelated to the word *duw* (push)).

The keywords "slaat," "slaan," "geslagen,"and "sloeg" are all variants of the verb "slaan" (to hit). They are kept separate because the common string "sl" is too short and using it as a keyword would lead to too many false positives. Since we already include them in our search, these inflections are required to match full words in the data (not parts of words).

Table 6 shows, for each keyword, how many negative admission periods in our dataset correspond to patients for whom a practitioner note containing that keyword was collected within that same 27-day frame (after exclusion of exceptions). These are candidate admission periods that were labeled as negative but are positive. The total number of such admission periods for all keywords is 1663 (the numbers on the table do not add up to 1663 because some notes contain multiple keywords). This represents 43% of our 3855 negative admission periods (see Section 3).

Upon visual inspection of some of the notes, we found that the large majority of them did not report violence incidents. A lot of them described hitting objects without causing any harm[5]; some of them described past events; others talked about hypothetical incidents. However, some cases were very clearly unreported violence incidents. Part of the challenge in selecting relevant notes is that what constitutes a violence incident is not well defined, as different practitioners have different standards. Therefore, an objective definition of what behavior constitutes violence is needed.

In this paper, we have assumed that the number of unreported violence incidents is small. Researchers and practitioners at UMC Utrecht make use of the violence incident reports frequently, and if the number of unreported incidents were large, they would have encountered problems before. However, further work needs to be done to mitigate this problem carefully.

---

[5]Incidentally, this might explain the high correlation between violence incidents and the word *deur* (door) found in [12].

6.4. Feature Importance

When using the random forest classifier, we stored the ten most important features according to the best fit in the inner cross-validation loop for each iteration in the outer cross-validation loop. Gathering the most important features together, we then studied both the ten most repeated features and the ten features with the highest total feature importance; these lists were reassuringly similar. The most repeated features were the age at the beginning of the admission period and the number of words in the period note. The frequency distributions of these variables are shown, for both positive and negative samples, in Figure 4.

As can be seen in the figures, the admission periods with a violent outcome correspond to younger patients on the average, and the average period note with a violent outcome is longer than the average period note with a nonviolent outcome. 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, where more careful feature engineering could lead to better discrimination.

6.5. Limitations

In this work, we used the results from the BVC as a benchmark to the predictions of our machine learning models, because the BVC is currently used for VRA at the UMC Utrecht. We found that the performance of some of our models was consistent with that of the BVC, but it should be noted that we computed the AUC for the BVC using violence incidents from the 27 days after the first day of admission. Therefore, our analysis does not compare with the standard use of the BVC, which is meant to predict violence within 24 hours. Additionally, both datasets seem to be drawn from different distributions, as the class imbalances differ from each other significantly. Efforts to align both datasets were out of the scope of this work. Furthermore, in our analysis we found indications that the dataset used for the machine learning models has some cases that
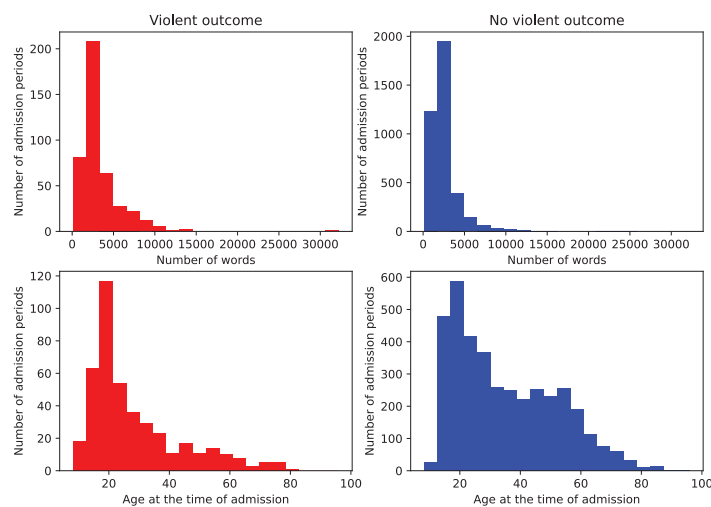

Figure 4 | Histograms of number of words per period note (top) and age at the beginning of the admission period (bottom) for positive/violent (left/red) and negative/nonviolent (right/blue) admission periods.

were not reported as violent, while the associated texts seem to indicate that a violence incident did occur. We do not know the rate of unreported cases, and future work should explore this problem further. Lastly, in this work we used BERTje to perform classification on the clinical notes. BERTje is not suited to perform classification on long texts, like clinical notes. Therefore, we experimented with two shortening strategies to fit the requirements of the model: summarization and truncation of texts. Each shortening strategy imposes information loss as not all the original text is used to train the BERTje model. This loss is likely to contribute to the relatively low performance of the model. Recently, a new BERT-like mode, called SMITH [45], has been released that is designed to analyze longer texts. Therefore, an implementation of SMITH for VRA can potentially improve the predictions.

7. CONCLUSIONS

We applied conventional and deep machine learning methods to the problem of VRA, using Dutch-language clinical notes from the psychiatry ward of the UMC in Utrecht, the Netherlands. Our results, reported on Tables 3 and 4, were competitive with a study based on structured variables that obtained AUC = 0.7801 [46], and with the BVC, currently used at the UMC, which gave AUC=0.761±0.100 when predicting violence incidents within 27 days after a questionnaire was filled out. These metrics show modest performance, and they indicate that further work is needed to extract all the meaningful information in the clinical notes.

We have applied a BERT-like model to VRA using clinical notes in Dutch. We found no improvement from using BERTje as opposed to conventional machine learning methods; indeed, the results from BERTje were worse, with AUC≈0.66. Yet we know that our dataset is small and may not be sufficient to fine-tune the large number of parameters present in BERTje. A larger dataset would be highly beneficial but is not trivial to accomplish.

To enlarge the dataset, data from multiple institutions can be aggregated via federated learning, whereby different institutions train the same central model without sharing the data with each other; this is very important, considering privacy restrictions. Additionally, it would be very beneficial to pre-train a "medical-BERTje" model, using Dutch clinical notes from various medical domains.

Finally, a key assumption in our work was that the number of unreported violence incidents was small. The validity of this assumption needs to be scrutinized. Further work is needed to refine the process of selecting unreported incidents from practitioner notes. If the number is small, the data points could be removed from the dataset. This could also inform a re-evaluation of the violence incident reporting in practice. We have seen that practitioners can be very subjective in reporting violence incidents. A unified strategy for incident reporting informed by machine learning could significantly improve data quality.

CONFLICTS OF INTEREST

We are not aware of any conflicts of interest from any of the authors.

AUTHORS' CONTRIBUTIONS

Mosteiro co-wrote the software for data preparation, model training and analysis, did the analysis and co-wrote the manuscript. Rijcken did literature review, co-wrote the software for calculation of performance metrics and co-wrote the manuscript. Zervanou did literature review, study design and supervision, and co-wrote and reviewed the manuscript. Kaymak did study design and supervision, and co-wrote and reviewed the manuscript. Scheepers provided the list of violence keywords and procured authorization from the UMC ethics committee and expert evaluation of the results. Spruit did study design and supervision and co-wrote and reveiwed the manuscript.

REFERENCES

[1] M. van Leeuwen, J. Harte, Violence against mental health care professionals: prevalence, nature and consequences, J. Forensic Psychiatry Psychol. 28 (2017), 581–598.

[2] M. Inoue, K. Tsukano, M. Muraoka, F. Kaneko, H. Okamura, Psychological impact of verbal abuse and violence by patients on nurses working in psychiatric departments, Psychiatry Clin. Neurosci. 60 (2006), 29–36.

[3] H. Nijman, L. Bowers, N. Oud, G. Jansen, Psychiatric nurses' experiences with inpatient aggression, Aggress. Behav. 31 (2005), 217–227.

[4] J.P. Singh, S.L. Desmarais, C. Hurducas, K. Arbach-Lucioni, C. Condemarin, K. Dean, *et al.*, International perspectives on the practical application of violence risk assessment: a global survey of 44 countries, Int. J. Forensic Ment. Health. 13 (2014), 193–206.

[5] R. Almvik, P. Woods, K. Rasmussen, The Brøset violence checklist: sensitivity, specificity, and interrater reliability, J. Interpers. Violence. 15 (2000), 1284–1296.

[6] C.Y. Chen, P.H. Lee, V.M. Castro, J. Minnier, A.W. Charney, E.A. Stahl, *et al.*, Genetic validation of bipolar disorder identified by automated phenotyping using electronic health records, Transl. Psychiatry. 8 (2018), 1–8.

[7] C. Colling, M. Khondoker, R. Patel, M. Fok, R. Harland, M. Broadbent, *et al.*, Predicting high-cost care in a mental health setting, BJPsych Open. 6 (2020), E10.

[8] R. Perlis, D. Iosifescu, V. Castro, S. Murphy, V. Gainer, J. Minnier, *et al.*, Using electronic medical records to enable large-scale studies in psychiatry: treatment resistant depression as a model, Psychol. Med. 42 (2012), 41–50.

[9] G. Gorrell, S. Oduola, A. Roberts, T. Craig, C. Morgan, R. Stewart, Identifying first episodes of psychosis in psychiatric

patient records using machine learning, in Proceedings of the 15th Workshop on Biomedical Natural Language Processing, ACL, Berlin, Germany, 2016, pp. 196–205.

[10] K.J. Moon, Y. Jin, T. Jin, S.M. Lee, Development and validation of an automated delirium risk assessment system (auto-delras) implemented in the electronic health record system, Int. J. Nurs. Stud. 77 (2018), 46–53.

[11] V. Menger, F. Scheepers, M. Spruit, Comparing deep learning and classical machine learning approaches for predicting inpatient violence incidents from clinical text, Appl. Sci. 8 (2018), 981.

[12] V. Menger, M. Spruit, R. van Est, E. Nap, F. Scheepers, Machine learning approach to inpatient violence risk assessment using routinely collected clinical notes in electronic health records, JAMA Netw. Open. 2 (2019), e196709.

[13] P. Mosteiro, E. Rijcken, K. Zervanou, U. Kaymak, F. Scheepers, M. Spruit, Making sense of violence risk predictions using clinical notes, in: Z. Huang, S. Siuly, H. Wang, R. Zhou, Y. Zhang (Eds.), Health Information Science, Springer International Publishing, Cham, Switzerland, 2020, pp. 3–14.

[14] J. Devlin, M.W. Chang, K. Lee, K. Toutanova, 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, Minneapolis, MN, USA, 2019, vol. 1, pp. 4171–4186.

[15] A.J. Quijano, S. Nguyen, J. Ordonez, Grid search hyperparameter benchmarking of BERT, ALBERT, and LongFormer on DuoRC, arXiv:2101.06326 [Preprint], 2021, p. 9. https://arxiv.org/abs/2101.06326

[16] Y. Cui, W. Che, T. Liu, B. Qin, S. Wang, G. Hu, Revisiting pretrained models for Chinese natural language processing, in EMNLP 2020: Findings of the Association for Computational Linguistics, ACL, 2020, pp. 657–668.

[17] N. Vaci, Q. Liu, A. Kormilitzin, F. De Crescenzo, A. Kurtulmus, J. Harvey, *et al.*, Natural language processing for structuring clinical text data on depression using UK-CRIS, Evid. Based Ment. Health. 23 (2020), 21–26.

[18] M. Senior, M. Burghart, R. Yu, A. Kormilitzin, Q. Liu, N. Vaci, *et al.*, Identifying predictors of suicide in severe mental illness: a feasibility study of a clinical prediction rule, Front. Psychiatry. 11 (2020), 268.

[19] R. Rijo, R. Martinho, L. Pereira, C. Silva, Text mining applied to electronic medical records, Int. J. E-Health Med. Commun. 6 (2015), 1–18.

[20] Y. Wang, L. Wang, M. Rastegar-Mojarad, S. Moon, F. Shen, N. Afzal, *et al.*, Clinical information extraction applications: a literature review, J. Biomed. Inform. 77 (2018), 34–49.

[21] S. Friedman, R. Margolis, O.J. David, M. Kesselman, Predicting psychiatric admission from an emergency room, J. Nerv. Ment. Dis. 171 (1983), 155–158.

[22] J.S. Lyons, J. Stutesman, J. Neme, J.T. Vessey, M.T. O'Mahoney, H.J. Camper, Predicting psychiatric emergency admissions and hospital outcome, Med. Care. 35 (1997), 792–800.

[23] B.L. Cook, A.M. Progovac, P. Chen, B. Mullin, S. Hou, E. Baca-Garcia, 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 (2016), 8708434.

[24] S.H. Huang, P. Le Pendu, S. VIyer, M. Tai-Seale, D. Carrell, N.H. Shah, Toward person-alizing treatment for depression: predicting diagnosis and severity, JAMIA. 21 (2014), 1069–1075.

[25] D. Van Le, J. Montgomery, K.C. Kirkby, J. Scanlan, Risk prediction using natural language processing of electronic mental health records in an inpatient forensic psychiatry setting, J. Biomed. Inform. 86 (2018), 49–58.

[26] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, in 1st International Conference on Learning Representations, Workshop Track Proceedings (ICLR 2013), Scottsdale, AZ, USA, 2013.

[27] Q. Le, T. Mikolov, Distributed representations of sentences and documents, PMLR. 32 (2014), 1188–1196.

[28] A. Rumshisky, M. Ghassemi, T, Naumann, P. Szolovits, V.M. Castro, T.H. McCoy, R.H. Perlis, Predicting early psychiatric readmission with natural language processing of narrative discharge summaries, Transl. Psychiatry. 6 (2016), e921.

[29] T. Fawcett, An introduction to ROC analysis, Pattern Recognit. Lett. 27 (2006), 861–874.

[30] J. Huang, C.X. Ling, Using AUC and accuracy in evaluating learning algorithms, IEEE Trans. Knowl. Data Eng. 17 (2005), 299–310.

[31] M. Sokolova, G. Lapalme, A systematic analysis of performance measures for classification tasks, Inf. Process. Manag. 45 (2009), 427–437.

[32] U. Kaymak, A. Ben-David, R. Potharst, The AUK: a simple alternative to the AUC, Eng. Appl. Artif. Intell. 25 (2012), 1082–1089.

[33] J. Cohen, A coefficient of agreement for nominal scales, Educ. Psychol. Meas. 20 (1960), 37–46.

[34] V. Menger, F. Scheepers, L. van Wijk, M. Spruit, DEDUCE: a pattern matching method for automatic de-identification of dutch medical text, Telemat. Inform. 35 (2018), 727–736.

[35] S. Bird, E. Klein, E. Loper, Natural language processing with python: analyzing text with the natural language toolkit. 2019. https://www.nltk.org/book/

[36] R. Řehůřek, P. Sojka, Software framework for topic modelling with large corpora, in Proceedings of LREC 2010 Workshop on New Challenges for NLP Frameworks, University of Malta, Valletta, Malta, pp. 45–50.

[37] S. Syed, M.R. Spruit, Full-text or abstract? Examining topic coherence scores using latent dirichlet allocation, in 2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA), Tokyo, Japan, pp. 165–174.

[38] C. Cortes, V. Vapnik, Support-vector networks, Mach. Learn. 20 (1995), 273–297.

[39] L. Breiman, Random forests, Mach. Learn. 45 (2001), 5–32.

[40] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, *et al.*, Transformers: state-of-the-art natural language processing, in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, ACL, 2020, pp. 38–45.

[41] Google Research, Multilingual BERT [repository], 2019. https://github.com/google-research/bert/blob/master/multilingual.md

[42] W. de Vries, A. van Cranenburgh, A. Bisazza, T. Caselli, G. van Noord, M. Nissim, BERTje: a Dutch BERT model. 2019. https://arxiv.org/abs/1912.09582

[43] B. van der Burgh, S. Verberne, The merits of universal language model fine-tuning for small datasets - a case with Dutch book reviews, arXiv:1910.00896 [Preprint], 2019. https://arxiv.org/abs/1910.00896

[44] L. Joosten, Sentiment Analysis of Dutch Tweets: a Comparison of Automatic and Manual Sentiment Analysis, Bachelor's Dissertation, Utrecht University, Utrecht, The Netherlands, 2015.

[45] L. Yang, M. Zhang, C. Li, M. Bendersky, M. Najork, Beyond 512 tokens: Siamese multi-depth transformer-based hierarchical encoder for long-form document matching, in Proceedings of the 29th ACM International Conference on Information & Knowledge Management, 2020, pp. 1725–1734.

[46] R. Suchting, C.E. Green, S.M. Glazier, S.D. Lane, A data science approach to predicting patient aggressive events in a psychiatric hospital, Psychiatry Res. 268 (2018), 217–222.