Mining Free-Text Medical Notes for Suicide Risk Assessment

Marios Adamou
South West Yorkshire Partnership NHS Foundation Trust, UK
University of Huddersfield, UK
Marios.Adamou@swyt.nhs.uk

Grigoris Antoniou
University of Huddersfield, UK
G.Antoniou@hud.ac.uk

Elissavet Greasidou
Gnosis Data Analysis PC, Greece
greasidouelissavet@gmail.com

Vincenzo Lagani
Gnosis Data Analysis PC, Greece
University of Crete, Greece
vlagani@csd.uoc.gr

Paulos Charonyktakis
Gnosis Data Analysis PC, Greece
haronykt@gmail.com

Ioannis Tsamardinos
University of Crete, Greece
Gnosis Data Analysis PC, Greece
University of Huddersfield, UK
Institute of Applied and Computational Mathematics, FORTH, Greece
tsamard.it@gmail.com

ABSTRACT

Suicide has been considered as an important public health issue for a very long time, and is one of the main causes of death worldwide. Despite suicide prevention strategies being applied, the rate of suicide has not changed substantially over the past decades. Advances in machine learning make it possible to attempt to predict suicide based on the analysis of relevant data to inform clinical practice. This paper reports on findings from the analysis of data of patients who died by suicide in the period 2013-2016 and made use of both structured data and free-text medical notes. We focus on examining various text-mining approaches to support risk assessment. The results show that using advance machine learning and text-mining techniques, it is possible to predict within a specified period which people are most at risk of taking their own life at the time of referral to a mental health service.

CCS CONCEPTS

• **Computing methodologies → Machine learning**; • **Applied computing → Life and medical sciences**;

KEYWORDS

automated machine learning, suicide prevention, risk assessment tool, clinical data, text mining

1 INTRODUCTION

Suicide has been considered as an important public health issue for a very long time and presently, its scale in avoidable loss of life was described by the UK House of Commons as unacceptable. 4820 people died by suicide in England in 2015 with the true figure likely to be higher [20]; and it is the second leading cause of death in people aged between 15 and 34 years in the United States [10, 11]. In special populations such as active duty military personnel [14] or people with mental health problems [12], the rates of suicide are even higher. Despite increasing efforts to reduce suicides through improved assessment and treatment, awareness campaigns and support services, the rate of suicide has not changed substantially over the past decades, although medical knowledge and healthcare technologies developed rapidly and huge progress in combating other leading causes of death, from cancer to cardiovascular diseases to HIV, was achieved.

Indeed, suicide risk has proven extremely difficult to assess for medical specialists as several variables are involved in its pathway. As a result, clinical instruments already is use attempting to predict it were found to not be clinically useful when classifying high risk individuals [8]. It can be therefore claimed that the traditional methodologies deployed in assessing suicide have not lived up to promise. Recent technological advances in information technologies made available new tools that could help improve suicide prevention, and there is growing interest in this direction. Data analytics has specifically been identified as a possible solution to uncover yet unknown patterns contributing to suicide tendency and this approach was tested on a database sample with good effect [31].

A particularly interesting problem in this context is to automatically assess suicide risk by analyzing person-related data. This data could be clinical records, social care data, psychological assessments or social media entries, to name a few. There are a few recent works studying specific aspects of data-driven suicide prevention. Kessler and co-authors considered mental health related hospitalizations of over 40,000 active US soldiers in 2004-2009, and developed a suicide risk assessment model predicting the risk of suicide within 12 months from discharge, with good predictive power [15]. Poulin et al. worked on predicting risk of suicide for US army veterans by analyzing clinical text notes using a learning algorithm on a genetic programming framework [23]. Salini et al. carried out a thorough, data driven, retrospective analysis of suicides in the Northwest of England to compare suicide risk assessment in primary and secondary care [24].

SETN2018, July 2018, Patras, Greece

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM.
https://doi.org/10.1145/3200947.3201020

In this paper we study risk assessment in the context of the UK National Health Service (NHS) at the entry point to mental health services. The aims of this study were primarily to, (a) test out new methodology and research design, (b) explore how we can learn from previous suicides to structure and inform future clinical practice, and (c) consider possible application on practice. The entry point to mental health services is by a referral from another NHS unit. A patient is referred to the services, is treated, if and for as long as it is deemed necessary, and is then discharged from the services with or without follow-on care recommendations and arrangements. The specific problem being studied in this paper is the most critical one in this context: to determine which referrals are those requiring the highest attention. In particular, we sought to find a strong prediction of which referrals pose the highest risk, this being defined as referrals within 3 or 6 months of committing suicide.

A companion paper [1] reports an analysis which used state of the art machine learning for structured data but a simple text-mining approach (bag-of-words). The results indicated that (a) it is indeed possible to identify the riskiest referrals in a fairly accurate way, and (b) the predictive accuracy is increased when free-text medical records are taken into account in conjunction with other structured database entries.

This paper explores whether employing more advanced text mining algorithms for constructing textual features (instead of bag-of-words), would help improve the performance of the resulting predictive model. In particular, we tested tf-idf [25, 27] features, n-grams (also known as bag-of-phrases), and features resulting from latent Dirichlet allocation [3] (LDA) models. We again coupled the newly constructed textual features with the structured ones from the analysis reported in [1].

2 DATA SAMPLE

A National Health Service specialist mental health provider (South West Yorkshire Partnership NHS Foundation Trust-SWYPFT) made available for analysis all the data it holds of mental health patients who died by suicide in the period 2013-2016. Overall there were 130 such patients. The data contain different type of information: demographics, referrals, appointments, progress notes, comprehensive assessments and Inpatient stays. All data types except demographics data contain several events defining the clinical trajectory of the subjects. The information is represented either as free text or as semi-structured fields. Here is a brief description of the most important data fields.

Demographics. For each patient is contains a subject ID, used to join with other entries related to the patient and information about date of birth, gender, marital status, ethnicity, religion, post code, date and age of death.

Referrals. There are 927 relevant referrals recorded in the data base. These are separate referrals made from primary care and other services to the mental health services of SWYPFT. Some patients had more than one distinct referral. Information included for each referral is referral sources and unit referred to (which organizational unit of SWYPFT), referral urgency, referral start and end date, discharge date and discharge reason (reasons ranging from intervention complete to return to primary healthcare to patient terminated or refused treatment). All this information is in (semi-) structured form.

Appointments. There are 12,167 relevant appointments recorded in the data base, providing the appointment history of each patient. Information included for each referral includes Team where appointment took place, appointment date and indication whether patient attended or not. All this information is in (semi-) structured form.

Progress Notes. There are a total of 40,268 relevant progress notes in the data base. Progress notes are added at various points of interaction with a patient by doctors, nurses, pharmacists, health visitors and other professional staff. Information included is the date of note entry, type of professional making the entry as free text.

Comprehensive Assessments. There are 273 such assessments that are much more informative than progress notes because they provide an abstraction. Information included for each assessment includes date, diagnosis, risk assessment, prognosis, client's view, communication quality, social and family circumstances, and ethnic, gender, cultural and spiritual issues. All fields other than subject ID and dates are provided in free text.

Inpatient Stays. There are 264 entries for relevant inpatient stays in a mental health hospital. Information about each stay includes hospital, dates of admission and discharge, ward and diagnosis.

3 SUICIDE RISK ASSESSMENT MODEL

In this section we summarize our work described in [1]. We devised a referral-centered analysis, with the goal of predicting whether a referred patient is close to attempting suicide. The objective is to realize a predictive model that assesses referrals according to the patient's risk for dying by suicide in the next t months (t was set to 3 and 6 months) and that can be made operational in a clinical environment. Referrals taken at most t months before suicide formed the positive class, and the rest formed the negative class.

The referrals carry their own structured information, such as urgency, length of the episode, reason for discharge, etc. We paired this data with the demographic characteristics of the patient of the corresponding referral as well as information from the last free-text medical note taken immediately before the referral. For the latter we converted the free-text information of the clinical data into structured data, suitable for a machine learning analysis with the use of the Natural Language Toolkit [2]: irrelevant information such as html tags, stop-words (i.e. most common words in a language), and human names was removed, all text was converted to lower case, and stemming was applied to the resulting words (separated by white spaces). The final textual features were constructed using the bag-of-words (BoW) model using the scikit-learn software [22]. With BoW text is represented as the multiset of its words, disregarding linguistic structure and structural markup, and only keeping word frequency.

Finally, we constructed new variables, e.g. for representing the number of clinical appointments that the patient had scheduled in the last X months before a referral, where $X = 1, \ldots, 12$. In total, there were 828 referrals (a.k.a. samples) and 7,711 variables, 7,686 of which derived from preprocessing textual information.

On this data we defined sixteen predictive analysis binary classification tasks, by varying several characteristics of the analysis : (a)

Table 1: Results for the predictive analyses following the "clinical design" approach and for $t = 3$. N and M denote the number of samples and variables in the dataset, respectively. C is the best-performing configuration (combination of feature selection and classification algorithm) from which the final model is constructed. SES (Statistically Equivalent Signatures) is described in [17], RLR stands for Ridge Logistic Regression [13], and RFs for Random Forests [6].

clinical design analysis characteristics		N	M	C	#selected variables	AUC (CI 95%)
structured variables	complete	828	25	SES, RLR	7	0.652 (0.589, 0.709)
	complete	828	25	SES, RLR	7	0.662 (0.607, 0.719)
structured and textual variables	complete	828	7711	SES, RFs	262	**0.705 (0.646, 0.760)**
	complete	828	7711	SES, RLR	25	0.605 (0.545, 0.662)

the value of the time-point t used to distinguish between negative and positive samples/referrals (set to 3 and 6), (b) the analysis design ("clinical" or "mirror-image"), (c) the inclusion of textual variables or not, and (d) whether the analysis will return a humanly interpretable model or it will be unrestricted (i.e., all the available learning algorithms will be tested in order for the predictive performance to be optimized). The "mirror-image" approach resembles a "mirror-image" study in which one compares outcome prior and after some event, and thus, a within-patient analysis becomes feasible (i.e. each patient is their own control); this design is frequently used in psychiatric research [9].

All the classification analyses were performed using the Just-Add-Data (JAD) Bio tool [4, 21, 26]. JAD Bio is an automated multivariate statistical analysis pipeline comprising of a complete set / sequence of learning steps that lead to the production of the final predictive model. Figure 1 (adapted from [4]) shows a schematic overview of JAD Bio's pipeline. In particular, JAD Bio performs (a) preprocessing of the data, (b) feature selection, (c) training of predictive models, (d) automated selection of the best configuration (i.e. a combination of feature selection and learning algorithms as well as specific values for their hyper-parameters) with which to construct the final model otherwise known as tuning, (e) construction of the final predictive model using the best configuration and all available data, and finally, (f) accurate performance estimation of the final model as well as its 95% confidence intervals.

Preprocessing methods change the values of the input variables, where necessary. In this particular analysis, the preprocessing methods employed included imputation of missing values, binarization of categorical variables (i.e. the levels of a categorical variable are coded as a collection of binary variables), and standardization of continuous variables.

Feature selection, also known as variable selection, is the process of identifying the most salient features for learning. The feature selection algorithm used for this analysis is the Statistical Equivalent Signatures (SES) [18]. SES tries to identify as many as possible minimal sets of features that provide optimal classification accuracy, i.e., it reports multiple solutions to the feature selection problem.

JAD employs state-of-the-art supervised machine learning algorithms and trains a variety of multivariate advanced and basic predictive models. In particular, for binary classification problems, it uses the following learning algorithms: Support Vector Machines (SVMs) [5] with linear, polynomial, and Gaussian kernels, Random Forests (RFs) [6], Decision Trees (DT) [7], and Ridge Logistic Regression (RLR) [13]. The tool automatically determines the set of configurations (on the basis of the statistical properties of the dataset, such as the number of training samples and the number of variables) to try. The best configuration, from which the final model will be constructed, is then identified using stratified, repeated K-fold cross-validation (i.e. stratified cross-validation is repeated multiple times with different partitions of the data to folds). With K-fold cross-validation (in our case K = 10) the data is split into K mutually exclusive subsets (a.k.a. folds) of approximately equal size and each fold is considered in turn as a test case for the models trained on the rest of the folds. Stratification refers to a random partition of the data to folds in a way that maintains approximately the distribution of the class variable. Its use is recommended as a better option compared to unrestricted random partitioning, both for the bias and variance of the performance estimate [16, 29]. The "repeated" version of cross-validation repeats the whole procedure several times (five for this analysis) for different random splits to folds. Multiple repetitions with different random splits reduce the variance of the performance estimation due to the particular choice of folds, leading to a better choice on average for the configuration to produce the final model.

The final predictive model that is returned is trained with the best found configuration on all available data. Retraining on all data returns a different model than the ones employed for estimating the performance during cross-validation. However, under the assumption that the loss of a learning algorithm drops monotonically, on average, with increasing sample size, this is on average a more predictive model to use operationally. It has been shown, both theoretically as well as empirically [28–30] that the cross-validated estimation of performance of the best classifier is optimistic. This is due to trying numerous combinations of algorithms and hyper-parameter values. JAD removes this bias using a bootstrap-based method called BBC-CV [28] and returns both a point estimate of predictive performance as well as its 95% confidence interval.

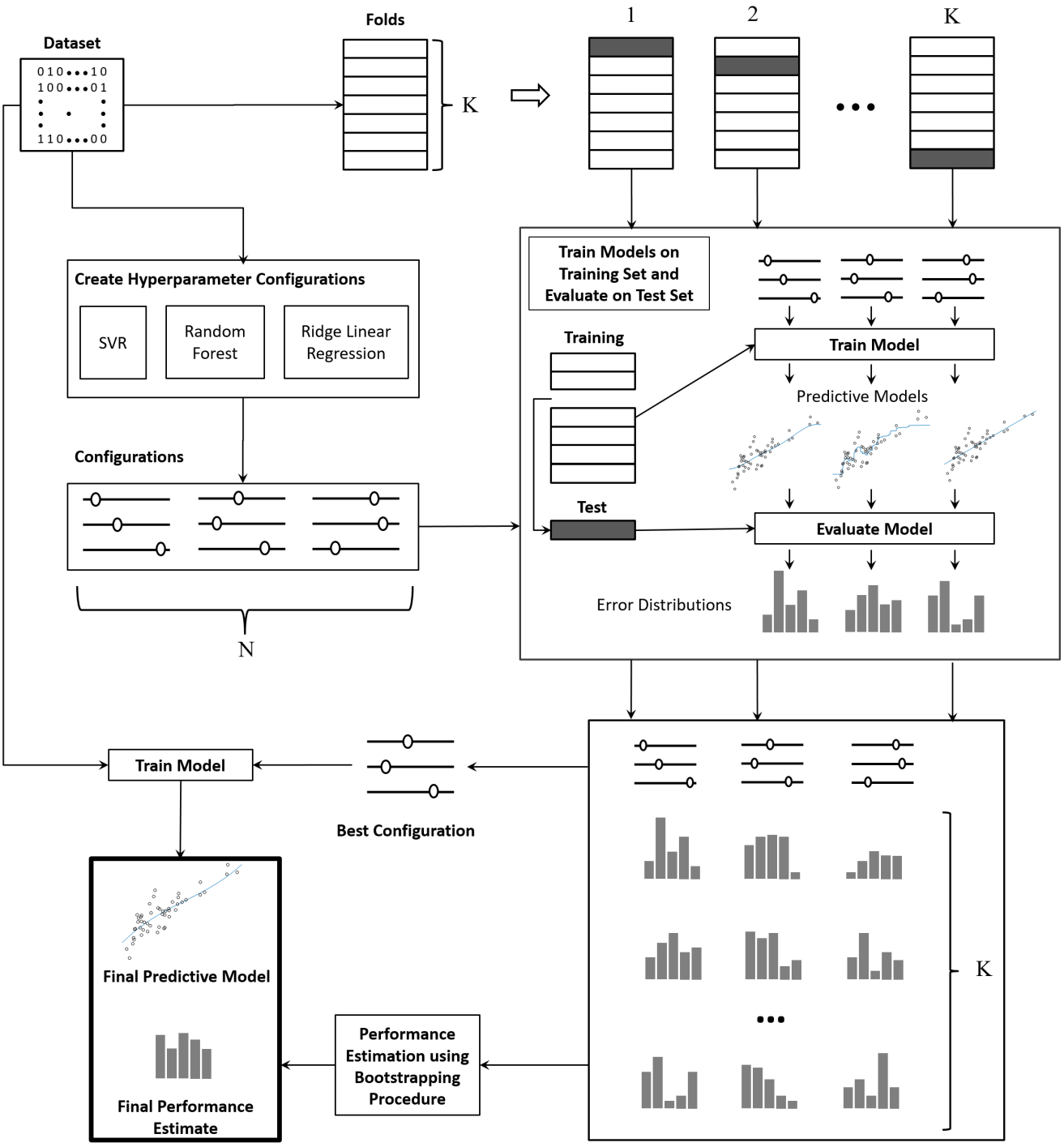

Figure 1: Schematic representation of JAD's data analysis pipeline: The tool determines the set of N configurations to try. Hyper-parameters are depicted as tuning sliders. The complete dataset is partitioned into K folds. Each fold is considered in turn as a test case for the models trained with every configuration on the union of the remaining folds. The best-performing configuration is selected on the basis of its average performance on the test folds. The final predictive model is trained with the best-performing configuration on the complete dataset. Finally, a bootstrap-based procedure is used to remove the optimism from the cross validated performance estimate (The figure is adapted from [4]).

Table 2: Results of analyses with tf-idf textual features: C is the best performing configuration. P_{tf-idf} is the performance for the tf-idf analyses and P_{BoW} is the performance for the corresponding analyses where the textual features were constructed using the BoW model. The metric used to measure performance is the AUC. SES (Statistically Equivalent Signatures) is described in [17], and RFs for Random Forests [6].

clinical design analysis characteristics	t	C	#selected variables	P_{tf-idf} (CI 95%)	P_{BoW} (CI 95%)
complete, all variables	3	SES, RFs	198	0.649 (0.588, 0.707)	0.705 (0.646, 0.760)
	6	SES, RFs	274	0.675 (0.621, 0.723)	0.697 (0.646, 0.744)

Table 3: Results of analysis with 2-grams textual features: C is the best performing configuration. $P_{2-grams}$ is the performance for the analysis with the 2-grams and P_{BoW} is the performance for the corresponding analysis using the BoW textual features. The metric of performance is the AUC. SES (Statistically Equivalent Signatures) is described in [17], and RFs for Random Forests [6].

clinical design analysis characteristics	t	C	#selected variables	$P_{2-grams}$ (CI 95%)	P_{BoW} (CI 95%)
complete, all variables	3	SES, RFs	35	0.716 (0.663, 0.770)	0.705 (0.646, 0.760)

Table 1 summarizes the predictive analysis for the "clinical design". The best overall performing model was obtained with the "clinical design" analysis where (a) both structured and textual variable are included, (b) there is no restriction on the tested configurations, and (c) the time-point t is equal to three months. The AUC achieved in this case is 0.705 with 95% CI equal to [0.646, 0.760]. The 95% CI of the AUC does not include the value 0.5, thus the results are deemed statistically significant at the standard significance level of 5%. These results show that using advance machine learning, it is possible to predict within a specified period which people are most at risk of taking their own life at the time of referral to a mental health service. It is worth noting that the best prediction is achieved including both structured and free-text medical information, which demonstrates the value included in free-text medical notes.

4 EXPERIMENTS REGARDING TEXT MINING

4.1 Tf-idf

In information retrieval, tf-idf [25, 27], short for term frequency-inverse document frequency, is a numerical statistic that is intended to reflect how important a term (word) is to a document in a collection or corpus. The simplest way to calculate the term frequency $tf(t,d)$ is to take the raw count of a term t in a document d (i.e., the number of times that term occurs in document). The inverse document frequency $idf(t)$ is a measure of how unique/important a word is, that is, how infrequently the word occurs across all documents: $idf(t) = log(1 + D)/(1 + df(d,t))$, where D is the total number of documents and $df(d,t)$ is the number of documents that contain the term t [22]. The tf-idf value increases proportionally to the number of times a word appears in the document, but is often offset by the frequency of the word in the corpus, which helps to adjust for the fact that some words appear more frequently in general: $tf - idf(t,d) = tf(t,d) \times idf(t)$.

The set of tf-idf textual features involved in the following analyses are the exact same as in the BoW model: the set of unique terms (words) that occur in the corpus. The difference lies in the way the value of each feature is calculated. We used the scikit-learn software [22] to calculate the tf-idf values.

We performed analyses following the "clinical design" approach where: (a) JAD bio tested the entire search grid (complete analysis), (b) the datasets include both structured and textual variables that were coded with their tf-idf value, and (c) the time-point t used to distinguish between positive and negative samples was to set to 3 and 6 months. A summary of the results is shown in Table 2. C is the best performing configuration, P_{tf-idf} is the performance (in AUC) of the final model constructed from C, and P_{BoW} is the performance of the corresponding analyses in which the textual features were constructed using the BoW model.

We notice that the performance of the final predictive model slightly drops for both values of t. This could be due to the fact that the texts are quite short (268 words on average) for any frequencies to be estimated accurately and thus tf-idf may not perform as well as expected.

4.2 N-grams

An n-gram is a contiguous sequence of n terms/words from a given sequence of text. In this analysis we constructed features from n-grams otherwise known as a bag-of-phrases model. We constructed features from 2-grams using the scikit-learn software [22].

The originally created 2-grams dataset consisted of 105,238 textual features (2-word phrases) and was extremely sparse. This was expected to a degree, since the clinical notes are mostly short in length. We eliminated the sparsest features by setting a threshold on the minimal number of occurrences of each one. The threshold was chosen to be 5 and it resulted in 7,181 features being included in the final dataset to be analysed.

Since the analysis was computationally intensive we chose to run the most promising scenario (the scenario that performed better in previous analyses). We followed the "clinical design" approach where: (a) JAD bio tested the entire search grid (complete analysis),

Table 4: Features selected by SES in the analysis described in this section (see also Table 3). In italic are the structured features while all other are 2-grams (2-word phrases). In parentheses are the equivalent features (where applicable). Recall that stemming has been applied to all words as a preprocessing step.

nhs number	psycholog therapi	pend court (punch window)	activ plan	make awar
Ethnicity	identifi bed	alcohol team	thought get	accept medic
Length of Episode	actavi uk	stay sister	plan monitor	yesterday report
lose job	live alon	found difficult	requir intervent	cider per
tell take	histori depress	depress disord	today deni	harm plan
acut admiss	rapid access	want help	veri mood	sincer practition
becam distress	*Discharge Reason*	express ongo	night time	mg diazepam

Table 5: Results of analysis with the topic composition features: C is the best performing configuration. P_{LDA} is the performance for the analysis with the topic composition features from the LDA model and P_{BoW} is the performance for the corresponding analyses where the textual features were constructed using the BoW model. The metric used to measure performance is the AUC. SES (Statistically Equivalent Signatures) is described in [17], RLR stands for Ridge Logistic Regression [13], and RFs for Random Forests [6].

clinical design analysis characteristics	t	T	C	#selected variables	P_{LDA} (95% CI)	P_{BoW} (95% CI)
complete, all variables	3	10	SES, RFs	28	0.671 (0.604, 0.726)	0.705 (0.646, 0.760)
		20	SES, RLR	7	0.669 (0.609, 0.726)	
	6	10	SES, SVM	10	0.632 (0.584, 0.676)	0.697 (0.646, 0.744)
		20	SES, SVM	12	0.647 (0.600, 0.691)	

(b) the dataset includes both structured and 2-grams textual variables, and (c) the time-point t was to set to 3. The results are shown in Table 3. We see that the performance of the resulting model is slightly better than that of the corresponding analysis using BoW textual features. It is interesting that also Poulin et al. in [23] noticed an increase in performance when they used "word-pairs" instead of "single-word terms". The selected features are shown in Table 4.

4.3 Latent Dirichlet Allocation

Topic models represent a family of computer programs that extract topics from texts. A topic to the computer is a list of words that occur in statistically meaningful ways. Latent Dirichlet allocation (LDA) is a special case of topic modeling introduced by discovery by David Blei, Andrew Ng, and Michael I. Jordan in 2002 [3].

Topic modeling, in this case, was performed with the use of the MAchine Learning for LanguagE Toolkit [19] (MALLET) for natural language processing which implements the LDA technique. The input to the toolkit was the set of progress and comprehensive notes (text files) that had first been processed, that is, (a) redundant information was removed (html tags, human names, punctuation, symbols, numbers), (b) stemming of words was applied, (c) the text was converted to lowercase and accents were removed, (d) token extraction was performed, and (e) stop words were filtered.

MALLET was trained to find $T = 10$ and 20 topics. For all T, the parameter for the number of iterations for the Gibbs sampling was set to 2000, and the parameter for the number of iterations between re-estimating the Dirichlet parameters was set to 20. The number of top key words to find for each topic used was 20 (the default value).

We used the topic proportions for the original text files (clinical notes) as features, that is, 10 features for $T = 10$ and 20 features when $T = 20$, and we constructed datasets by pairing them with the structured features that were used in all the other analyses. We conducted analyses following the "clinical design" approach where: (a) JAD bio tested the entire search grid (complete analysis), and (c) the time-point t was to set to 3. The results are presented in Table 5.

We notice that the predictive models perform slightly worse than the one resulting from the corresponding analysis in which the textual features were constructed using the BoW model. In future analyses it would be interesting to investigate whether increasing the number of topics would result in better performing predictive models.

5 DISCUSSION AND CONCLUSION

The first crucial step of this research was to define a problem that is most relevant in the clinical setting within the NHS. The problem chosen was to study which referrals are close to a suicide event, to aid the initial assessment of patients; this assessment is crucial for deciding which referrals to prioritize and where resources should focus. In addition, we are interested in results that are applicable in a clinical setting, meaning that no patients should a priori be omitted. To the best of our knowledge, this is the first work attempting to derive automated risk assessment results under these conditions. [31] reports on a USA-based project, carried out independently in parallel to our project that shares some of our objectives but has a

different setting. Among others, that work excludes actual suicides and focuses on (unsuccessful) suicide attempts, while our work focuses specifically to actual suicides; moreover, such a distinction between attempts and suicides is difficult to implement in a clinical setting. In our work we concentrate on the clinically most significant problem: identify at referral point which referrals are most risky of being close to suicide.

Another important decision underlying our research was to derive results tailored to the particular group of patients, namely patients in mental health services. As we stated in the Introduction, given that existing generic scales fail to provide accurate risk assessment, a strategy seeking to adapt to individuals or groups of individuals appears most promising. [8] This focus on risk assessment models tailored to particular settings distinguishes this research from other works, e.g. [31], that aim at developing generic suicide prediction algorithms. We argue that locally adapted algorithms and models are better suited to take into account local characteristics, leading to potentially better automatic prediction models and deeper insights made available to clinicians.

From a technical perspective, a challenge in the problem we set ourselves is that we decided to analyse both structured and free-text data. While the few previous works on automatically predicting suicide risk considered only unstructured medical notes [15] or only structured data [23], we include in our analysis structured information, e.g. about demographics, appointments, hospitalizations and treatments; as well as medical notes written in free text. Our intuition suggested that both sources of information can provide important clues for assessing suicide risk. Indeed, our experimental results show that considering medical notes in addition to structured information allows for more accurate prediction results. In addition, the results reported in this paper demonstrate that more sophisticated text mining algorithms, in particular n-grams, are capable of getting better results that a simpler approach based on bags of words.

The best prediction models in our analysis had an AUC value of just over 0.7. While this result would be considered fair for engineering and some biomedical applications, in the context of mental health diagnoses the picture is very different. Many of the best-performing behaviour checklists and interventions in psychology and psychiatry currently available deliver AUC estimates in the 0.7 - 0.8 range under clinically realistic conditions [33]. In fact, Youngstrom et al. argue that AUCs greater than 0.90 more likely to indicate design flaws rather than exceptional discriminative validity [32]. The case of suicide risk assessment is even more difficult as it seeks to predict future human behavior. According to [8], clinical instruments already is use attempting to predict suicide were found to not be clinically useful when classifying “high risk” individuals, suggesting an AUC value of around 0.5. Thus, the results reported here would be a major step towards a more accurate assessment of suicide risk.

This work has a number of limitations. One limitation is related to the fact that only data of patients who did die by suicide was available to us. Thus, negative referrals originate from a population of patients who eventually died by suicide. The major assumption for our model to be clinically valid is that referrals collected prior of t months from patients who committed suicide and referrals collected from patients who do not commit suicide have similar distributions. Clearly, it is important to incorporate data of living patients in future studies, both because more machine learning techniques can be employed and because the data will contain more information. The combination of the two make us expect better predictive power once control group data is included. In addition, this study was based on a fairly small data sample. Doing a similar analysis with more data, e.g. from other NHS Trusts, would provide a firmer foundation for the findings.

In summary, the findings are very promising, but this research should be seen as a first step towards improving suicide assessment in clinical settings through the use of machine learning. Further studies are needed before deployment in a clinical setting can be considered.

ACKNOWLEDGMENTS

This work was supported by the NHS South West Yorkshire Foundation Trust.

REFERENCES

[1] Marios Adamou, Grigoris Antoniou, Elissavet Greasidou, Vincenzo Lagani, Paulos Charonyktakis, Ioannis Tsamardinos, and Michael Doyle. [n. d.]. Towards Automatic Risk Assessment to Support Suicide Prevention. ([n. d.]). (*to appear*).
[2] Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.".
[3] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3, Jan (2003), 993–1022.
[4] Giorgos Borboudakis, Taxiarchis Stergiannakos, Maria Frysali, Emmanuel Klontzas, Ioannis Tsamardinos, and George E Froudakis. 2017. Chemically intuited, large-scale screening of MOFs by machine learning techniques. *npj Computational Materials* 3, 1 (2017), 40.
[5] Bernhard E Boser, Isabelle M Guyon, and Vladimir N Vapnik. 1992. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*. ACM, 144–152.
[6] Leo Breiman. 2001. Random forests. *Machine learning* 45, 1 (2001), 5–32.
[7] Leo Breiman, JH Friedman, Richard A Olshen, and Charles J Stone. 1984. Classification and Regression Trees. *Wadsworth* (1984).
[8] Gregory Carter, Allison Milner, Katie McGill, Jane Pirkis, Navneet Kapur, and Matthew J Spittal. 2017. Predicting suicidal behaviours using clinical instruments: systematic review and meta-analysis of positive predictive values for risk scales. *The British Journal of Psychiatry* (2017), bjp–bp.
[9] Andrea Fagiolini, Paola Rocca, Serafino De Giorgi, Edoardo Spina, Giovanni Amodeo, and Mario Amore. 2017. Clinical trial methodology to assess the efficacy/effectiveness of long-acting antipsychotics: Randomized controlled trials vs naturalistic studies. *Psychiatry research* 247 (2017), 257–264.
[10] National Center for Health Statistics (US et al. 2017. Health, United States, 2016: with chartbook on long-term trends in health. (2017).
[11] American Foundation for Suicide Prevention. 2017. Suicide Statistics. https://afsp.org/about-suicide/suicide-statistics/. (2017).
[12] Beth Han, Wilson M Compton, Joseph Gfroerer, and Richard McKeon. 2014. Mental health treatment patterns among adults with recent suicide attempts in the United States. *American journal of public health* 104, 12 (2014), 2359–2368.
[13] Arthur E Hoerl and Robert W Kennard. 1970. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* 12, 1 (1970), 55–67.
[14] Jeffrey Hyman, Robert Ireland, Lucinda Frost, and Linda Cottrell. 2012. Suicide incidence and risk factors in an active duty US military population. *American journal of public health* 102, S1 (2012), S138–S146.
[15] Ronald C Kessler, LTC Christopher H Warner, LTC Christopher Ivany, Maria V Petukhova, Sherri Rose, Evelyn J Bromet, LTC Millard Brown III, Tianxi Cai, Lisa J Colpe, Kenneth L Cox, et al. 2015. Predicting US Army suicides after hospitalizations with psychiatric diagnoses in the Army Study to Assess Risk and Resilience in Servicemembers (Army STARRS). *JAMA psychiatry* 72, 1 (2015), 49.
[16] Ron Kohavi et al. 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai*, Vol. 14. Montreal, Canada, 1137–1145.
[17] Vincenzo Lagani, Giorgos Athineou, Alessio Farcomeni, Michail Tsagris, and Ioannis Tsamardinos. 2016. Feature selection with the r package mxm: Discovering statistically-equivalent feature subsets. *arXiv preprint arXiv:1611.03227*

(2016).
[18] Vincenzo Lagani, Giorgos Athineou, Alessio Farcomeni, Michail Tsagris, Ioannis Tsamardinos, et al. 2017. Feature Selection with the R Package MXM: Discovering Statistically Equivalent Feature Subsets. *Journal of Statistical Software* 80, i07 (2017).
[19] Andrew Kachites McCallum. 2002. Mallet: A machine learning for language toolkit. (2002).
[20] House of Commons Health Committee. 2017. Suicide prevention: Sixth Report. (2017).
[21] Georgia Orfanoudaki, Maria Markaki, Katerina Chatzi, Ioannis Tsamardinos, and Anastassios Economou. 2017. MatureP: prediction of secreted proteins with exclusive information from their mature regions. *Scientific reports* 7, 1 (2017), 3263.
[22] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in Python. *Journal of machine learning research* 12, Oct (2011), 2825–2830.
[23] Chris Poulin, Brian Shiner, Paul Thompson, Linas Vepstas, Yinong Young-Xu, Benjamin Goertzel, Bradley Watts, Laura Flashman, and Thomas McAllister. 2014. Predicting the risk of suicide by analyzing the text of clinical notes. *PloS one* 9, 1 (2014), e85733.
[24] Pooja Saini, David While, Khatidja Chantler, Kirsten Windfuhr, and Navneet Kapur. 2014. Assessment and management of suicide risk in primary care. *Crisis: The Journal of Crisis Intervention and Suicide Prevention* 35, 6 (2014), 415.
[25] G Salton and MJ McGill. 1983. Introduction to modern information Philadelphia, PA. American Association for Artificial Intelligence retrieval. (1983).
[26] Olympia Simantiraki, Paulos Charonyktakis, Anastasia Pampouchidou, Manolis Tsiknakis, and Martin Cooke. 2017. Glottal Source Features for Automatic Speech-based Depression Assessment. *Proc. Interspeech 2017* (2017), 2700–2704.
[27] Karen Sparck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation* 28, 1 (1972), 11–21.
[28] Ioannis Tsamardinos, Elissavet Greasidou, and Giorgos Borboudakis. [n. d.]. Bootstrapping the Out-of-sample Predictions for Efficient and Accurate Cross-Validation. *Machine Learning* ([n. d.]). *to appear.*
[29] Ioannis Tsamardinos, Amin Rakhshani, and Vincenzo Lagani. 2015. Performance-estimation properties of cross-validation-based protocols with simultaneous hyper-parameter optimization. *International Journal on Artificial Intelligence Tools* 24, 05 (2015), 1540023.
[30] Sudhir Varma and Richard Simon. 2006. Bias in error estimation when using cross-validation for model selection. *BMC bioinformatics* 7, 1 (2006), 91.
[31] Colin G Walsh, Jessica D Ribeiro, and Joseph C Franklin. 2017. Predicting risk of suicide attempts over time through machine learning. *Clinical Psychological Science* 5, 3 (2017), 457–469.
[32] Eric Youngstrom, Oren Meyers, Jennifer Kogos Youngstrom, Joseph R Calabrese, and Robert L Findling. 2006. Comparing the effects of sampling designs on the diagnostic accuracy of eight promising screening algorithms for pediatric bipolar disorder. *Biological Psychiatry* 60, 9 (2006), 1013–1019.
[33] Eric A Youngstrom. 2013. A primer on receiver operating characteristic analysis and diagnostic efficiency statistics for pediatric psychology: we are ready to ROC. *Journal of pediatric psychology* 39, 2 (2013), 204–221.