Identifying First Episodes of Psychosis in Psychiatric Patient Records using Machine Learning

Genevieve Gorrell [1], Sherifat Oduola [2], Angus Roberts [1]
Thomas Craig [2], Craig Morgan [2] and Robert Stewart [2]

[1] Department of Computer Science, University of Sheffield, UK
g.gorrell,angus.roberts@sheffield.ac.uk

[2] Institute of Psychiatry, Kings College London, UK
sherifat.oduola,tom.craig,craig.morgan,robert.stewart@kcl.ac.uk

Abstract

Natural language processing is being pressed into use to facilitate the selection of cases for medical research in electronic health record databases, though study inclusion criteria may be complex, and the linguistic cues indicating eligibility may be subtle. Finding cases of first episode psychosis raised a number of problems for automated approaches, providing an opportunity to explore how machine learning technologies might be used to overcome them. A system was delivered that achieved an AUC of 0.85, enabling 95% of relevant cases to be identified whilst halving the work required in manually reviewing cases. The techniques that made this possible are presented.

1 Introduction

The epidemiology of first episode psychosis (FEP) is the central tenet on which psychiatric research builds an understanding of psychotic disorder, and accurate estimates of incidence rates of psychosis are important to measure the burden of the disease in the population (Baldwin et al., 2005; Hogerzeil et al., 2014). Yet challenges recruiting patients with FEP and variation in incidence rates are widely reported (Patel et al., 2003; Borschmann et al., 2014; Kirkbride et al., 2006). Sampling methods used for estimating incidence of psychosis may contribute to some of the reported challenges. For example, some previous studies have used a first contact sampling frame e.g. first hospital admission or 'first early intervention' (i.e. patients presenting to early-phase psychosis services). However these methods of identifying cases do not take into account individuals who may already be receiving treatment for non-psychotic disorder but who later manifest psychotic symptoms (Hogerzeil et al., 2014). Electronic health records can help alleviate these problems, whereby clinical information is screened using a diagnostic instrument to identify symptoms of psychosis within a defined period and consequently classify new FEP cases. An example of such an endeavour is work being carried out at the Institute of Psychiatry and South London & Maudsley (SLaM) NHS Trust using the Biomedical Research Centre Clinical Records Interactive Search (CRIS) to identify FEP cases in the CRIS-First Episode Psychosis study (Bourque, 2015). To summarize this work, psychiatric experts manually coded data in the free-text of clinical records between 1st May 2010 and 30th April 2012 for patients presenting to SLaM with compliance for psychotic disorder using a psychiatric diagnostic tool. Whilst the screening of clinical records sampling method comprehensively identifies cases and reduces risk of underestimation, this approach raises resource and efficiency challenges. For example, review of clinical records requires expert level resource (such as a psychiatrist or psychiatric nurse) for annotation, which can be very expensive. On average approximately 80-100 individual clinical records were screened per week by each annotator. It is clear that manual screening of electronic records is resource-intensive and time-consuming.

With these challenges in mind, advances in natural language processing technology have been drawn on in this work to apply techniques to identify and classify FEP cases based on the data generated from the manual screen (Bourque, 2015). An automated screening application has the potential to improve the efficiency of the FEP case identification task, reducing the burden of manual screening as well as saving time and money. Such

Proceedings of the 15th Workshop on Biomedical Natural Language Processing, pages 196–205,
Berlin, Germany, August 12, 2016.

an approach may also provide a methodological advantage in identifying FEP cohorts who may be followed up longitudinally to answer important questions about outcomes following their experience of psychosis. The use of natural language processing has potential implications for service planning and evaluation for patients with FEP.

CRIS contains both the structured information and the unstructured free text from the SLaM EHR. The free text consists of 20 million text field instances containing correspondence, patient histories and notes describing encounters with the patient. These free text fields contain much information of value to mental health epidemiologists. clinicians often record vital information in the textual portion of the record even when a structured field is designated for this information. For example, a query on the structured fields for Mini Mental State Examination scores (MMSE, a score of cognitive ability) in a recent search returned 5,700 instances, whereas a keyword search over the free text fields returned an additional 48,750 instances. Previous research has noted that free text is convenient, expressive, accurate and understandable (Meystre et al., 2008; Rosenbloom et al., 2011), making it appealing for clinical record data entry despite the greater research value of structured data. Powsner et al (1998) observe that structured data is more restrictive, whereas Greenhalgh (2009) comments that free text is tolerant of ambiguity, which supports the complexity of clinical practice; a particularly relevant factor, perhaps, in psychiatry. Medical language is often hedged with ambiguity and probability, which is difficult to represent as structured data (Scott et al., 2012). For these reasons, the diagnosis structured field in the patient record is of only minor utility in identifying FEP cases. On the other hand, the free text field may often not give a clear initial opinion of the diagnosis. The understanding of this episode as a first episode of psychosis may instead unfold over time; for example, an unclear episode may be more conclusively identified as psychotic in the light of subsequent episodes, or ruled out as psychosis through the finding of organic causes. Furthermore the records we are interested in are those that record the initial psychotic episode, rather than subsequent ones in a patient already diagnosed, though the language surrounding the event may be extremely similar. The task therefore presents challenges for NLP.

1.1 Previous Research

Previous work has attempted to identify relevant cases for research in patient records, and has tended to make use of keyword search and rule-based approaches, though a body of work exists on statistical case classification. Ford et al (2016) note that making use of the free text information consistently improves accuracy compared with structured fields only, but there is little to distinguish the success of rule-based and machine learning approaches to case classification. Of the 67 studies they reviewed, 23 used a data-driven approach to classification, with logistic regression being the most popular choice, but with all of the better known classification algorithms represented. Features on which the classification took place are often bespoke gazetteers, though various established biomedical information extraction systems are used as a preparatory step, most notably cTAKES (Savova et al., 2010), MetaMap (Aronson, 2001) and HITEx (Zeng et al., 2006). Bag-of-words representations and character n-grams are common. Systems often include some form of assertion and/or negation detection, such as NegEx (Chapman et al., 2001). The studies cover a variety of general medical conditions, and results vary, with recalls (sensitivities) and precisions (positive predictive values, or PPVs) typically between around 50% and the high 90s. A further study not included by Ford et al uses word trigrams to achieve a good result in detecting patients with acute lung injury (Yetisgen-Yildiz et al., 2013). Given the varied task conditions, it is difficult to generalize about what constitutes a good result.

Several studies are of more specific relevance to psychiatry. Castro et al (2014) report an AUC of 0.82 classifying patients based on their record according to bipolar status, and an AUC of 0.93 for classifying individual notes (subdocuments) within the patient record, a result they achieved using HITEx for feature generation, along with a bespoke gazetteer, and logistic regression (LASSO) for classification. Among previous work, theirs is perhaps the most comparable to the study presented here, in particular the classification of the entire case, rather than the individual note, since this is a closer parallel to this work, in which a portion of the patient record covering a window of many subdocuments is used to classify the whole case. Bellows et al (2014) focus on terms rather

than classifying the whole case to identify binge eating disorder diagnoses. They provide an accuracy figure with no kappa, and a sensitivity (recall) without a specificity or a PPV (precision) so it is hard to compare their outcome with other similar work. Perlis et al (2012) have had some success using bespoke text features and logistic regression to classify patients with major depressive disorder according to their current status, achieving AUCs in the range of 0.85 to 0.88. Huang et al (2014) classify depression patients according to disease severity, and predict 12 month outcomes. Seyfried et al (2009) provide technological support for manual depression case identification, but do not include automated classification.

The approaches used here are in keeping with previous work, whilst applying the techniques to a novel domain with new challenges. Linking to a medical ontology has not been done here since existing vocabularies do not provide a good coverage of terms relevant in this case, but a contextualizer was utilized (discussed in more detail below) to distinguish mentions being experienced by the patient, now, from, for example, those having been experienced by a family member or in the past. Sentence classification has not been used in preference to whole case classification because first episode psychosis diagnoses are so very rarely clearly stated.

2 Data

The manual case identification is described elsewhere (Bourque, 2015). In brief, a three stage screening of clinical records was conducted by three clinically trained researchers (a psychiatrist, a medical doctor and a psychiatric nurse), and a research assistant, overseen by a principal investigator.

Firstly, SQL commands were used to retrieve anonymised information for all persons presenting to all adult mental health services serving the population of interest. Search criteria were weekly search period, service location (i.e. all SLaM services in Lambeth and Southwark), age-range and symptom terms (e.g. psychos*; psychot*, delusion*, voices, hallucinat* paranoia). Once retrieved, individual patient records were screened and reviewed by the aforementioned researchers using a validated diagnostic screening tool, namely, the Item Checklist Group of the Schedule of Clinical Assessment of Neuropsychiatry SCAN (WHO, 1994), to identify first episode psychosis cases. Individuals were included as cases if they were: resident in the London boroughs of Lambeth or Southwark; aged 18-64 years (inclusive); experiencing psychotic symptoms of at least one day duration during the study periods and scored at least 2 or more for psychotic symptoms as assessed using the SCAN. This screening process described above enabled the assignment of population at risk into three categories i.e. FEP cases, no psychosis and excluded.

Secondly, two primary researchers (a psychiatric nurse and a psychiatrist) reviewed all the included cases from the first stage screen to ensure cases met all inclusion criteria. An inter-rater reliability test was carried out between the two experts and Cohen's Kappa coefficient of 0.77 ($p=<0.01$) was achieved. Finally, discrepant or ambiguous cases were resolved by consensus with the principal investigator.

In total, 9109 individual clinical records were screened, of whom 560 screened positive and were FEP cases, 5234 screened negative for psychosis (but remain at risk and allocated to re-evaluation) and 3315 were excluded (because of evidence of any of the following: previous psychosis, organic psychosis, not resident, too young or too old). In the work described below, these 9109 records were split into a tuning set (two thirds of the total) and a test set (the remainder).

3 Experiments

In order to facilitate further identification of relevant cases, a case classification application was created using GATE (Cunningham et al., 2013), since this technology provides a wide variety of different information extraction tools that can be used to create features for machine learning, as well as an integration of LibSVM's (Chang and Lin, 2011) support vector machine and various of the Mallet (McCallum, 2002) and Weka (Hall et al., 2009) algorithms, and has been in use at SLaM for several years.

Due to the challenging nature of the data, systematic exploration of available tools was required to produce a good result. This work focuses on three algorithms; support vector machines (SVMs, in particular LibSVM), and Weka's Random Forest and JRip. In the course of experimentation, many algorithms were tried, but these three have been chosen as the focus here because they formed

good practical propositions, both in terms of accuracy of classification and speed, and being diverse, provide insight into the ways that different techniques interact with algorithm choice.

The work is presented here in two parts. Feature selection and parameter tuning is discussed, showing how these can be used to improve the accuracy of the classifiers. Then the problem of bias against the minority class is explored. Since the cases to be identified are by far the minority, and the priority is finding as many of them as possible, optimizing overall accuracy was not sufficient. The second section, therefore, addresses this issue, and concludes with an assessment of the utility of confidence scores for providing fine-grained control over the level of recall achieved.

All experimental software is available to download[1]. A Docker file is provided that builds the experimental environment, with an entrypoint script running the complete experiment set presented in this paper, generating the results shown. The data is however highly confidential and therefore cannot be shared.

3.1 Feature Selection and Parameter Tuning

Early experiments used a feature set including word bigrams and unigrams (trigrams lead to an very high dimensionality of problem, and previous experience has shown that they are likely to overfit all but the largest of corpora, which our 9109 cases, whilst sizeable for an expert-annotated set of complex cases, is not) as well as presence of terms in a comprehensive gazetteer provided by a medical expert. This gazetteer covered symptoms relevant to psychosis, and was further supplemented with a speculative term set relevant to diagnosis and treatment, such as the phrase "first episode [of] psychosis" or phrases relevant to sectioning and hospital admission. ConText (Harkema et al., 2009) was applied to these gazetteer mentions to add information about whether it is the patient that is experiencing the observation or another individual, for example a family member; whether they are stated as experiencing or not experiencing it (e.g. "no evidence of psychosis"); and whether the finding is noted in the present or past (e.g. "had previously experienced auditory hallucinations"). Note that the phrase "first episode psychosis" or "first episode of psychosis", whilst telling, is extremely rare, occurring only a couple of times in the whole corpus. A typical case record progresses all the way from first presentation through to treatment and management with minimal discussion of the diagnosis.

In addition to these features, there are some structured data fields associated with the cases, including diagnosis, as well as demographic information such as gender, ethnicity and date of birth. A number of quite detailed diagnosis categories correspond to psychosis of the type we are interested in. Diagnosis fields (of which there are several) are utilized to differing extents by clinicians, and may be empty or out of date. Furthermore, diagnosis does not help us to identify that this record describes a first episode of psychosis. However, diagnosis fields are an obvious feature to include.

GATE was used to create feature representations of the tuning instances, which were then exported in ARFF format, in order to experiment with feature extraction techniques available in Weka but not in GATE. Weka's CfsSubsetEval was used with BestFirst feature selection, as this is a pragmatic option. However due to time constraints, this was impractical over the very large dimensionalities necessitated by the inclusion of unigram and bigram features. Instead feature selection was performed without including n-grams. Results are presented for the feature set including unigrams in order to contrast the overall performance, but the feature set across which feature selection was performed was limited to the feature set without n-grams.

Feature selection provides an insight into the data. Note that a feature not being selected does not imply it is of no utility in separating the cases, since it may be redundant in conjunction with a better feature. Note also that the feature selection methods employed may not be congruent with the algorithms we then go on to use, since some algorithms may be able to, for example, combine features differently to produce useful information. Nonetheless it is interesting to note what seems to help to separate the cases. Listed here are the features strongly selected, being found valid over three out of three folds of the data. Below that, the features presented were found valid in two out of three folds. All selected gazetteer features are positive mentions experienced by the patient in the present, as ascertained using ConText.

- Validated in 3/3 folds

[1] http://www.dcs.shef.ac.uk/~genevieve/bionlp-docker-fep.zip

  - Null or empty values in the following structured fields; borough, ethnicity, gender, postcode, first primary diagnosis
  - Age
  - First primary diagnosis:
    * bipolar, hypomanic (F31.0)
    * bipolar, unspecified (F31.9)
    * severe depressive w/psychotic symptoms (F32.3)
  - Text features, presence of gazetteer terms; “olanzapine”, “risperidone”, “auditory hallucinations”, “voices”, “paranoid”, “psychotic”, “psychosis”

- Validated in 2/3 folds
  - First primary diagnosis
    * bipolar (F31)
    * organic delusional schizophrenia-like disorder (F06.2)
    * organic mood disorder (F06.3)
  - Text features, presence of gazetteer terms; “aripiprazole”, “quetiapine”, “persecutory”, “schizophrenia”

Reflecting on these features, it is interesting that the absence of some structured information, for example an empty value for postcode, enables some separation of the cases. It may be that first episodes of psychosis, perhaps because they often present under troubled circumstances, tend to arrive in the system via a different route that has some systematic differences to more routine cases, resulting in these differences in the case record. It is unsurprising that diagnosis fields are of value, being likely to assist both in finding positive cases and ruling out negative ones (e.g. organic causes). Furthermore, antipsychotic drugs and the more telling of symptoms appear prominently, as do terms such as “psychosis”, that suggest a postulated diagnosis. It is also interesting that only a small number of features is selected, the majority being redundant.

Next, the impact feature selection has on accuracy with regards to the three algorithms is investigated. Firstly, the SVM is tuned. Then, feature scaling (normalization) and cost are considered in conjunction with feature selection. The cost parameter of an SVM refers to the importance attached to creating a classifier that correctly classifies the training instances. A high cost results in a better fit to the training data, though may potentially overfit. A low cost may result in a weak classifier that hasn't made the best use of the training data.

Feature normalization describes a process whereby numeric features are brought into a similar statistical distribution with each other, for example by scaling them all to have the same mean and variance. In this case, age is a numeric feature with a very different range than the nominal features that otherwise dominate. Nominal features are expanded out to one dimension per value and assigned counts, which for many fields such as diagnosis fields, amount to ones and zeroes for presence or absence. The greater magnitude of the age feature in no way reflects its greater importance, yet vector space algorithms may attach more importance to larger values. In this work, this is relevant to the SVM. The other two algorithms used here are unaffected by the magnitude of numeric features.

Table 1 shows a sample of the results obtained from evaluating using three-fold cross-validation on the tuning corpus with the large feature set including unigrams, and table 2 shows a sample of the results obtained from evaluating using threefold cross-validation on the tuning corpus with the reduced feature set. We can see that where the larger feature set is used, including unigrams, cost and feature normalization have an important role to play in getting a competitive result. At lower costs, feature normalization is detrimental, but once cost comes into the right range, it helps. However, on the reduced feature set, obtaining a good result is far easier. Feature normalization does not have much impact any more, and cost, whilst an important parameter to tune, is less critical. This result emphasizes the potential value of selectiveness with features to the SVM, whilst highlighting the role that cost tuning and feature normalization may play in working with a less optimal feature selection.

Having tuned the SVM, this was now compared to the other two algorithms with regards to feature selection. GATE was used to produce a new ARFF file of the tuning instances with the reduced feature set, in addition to the full set with and without unigrams, which were then evaluated in Weka using threefold cross-validation. Adapting

Cost	Feat Norm?	Accuracy	Kappa
1	No	66.3%	0.2496
10	No	70.4%	0.3809
1000	No	74.2%	0.4937
1	Yes	59.86%	0
10	Yes	60.0%	0.0068
1000	Yes	79.6%	0.5629

Table 1: Parameter tuning on the large feature tuning set including unigrams.

Cost	Feat Norm?	Accuracy	Kappa
1	No	81.8%	0.6244
10	No	82.2%	0.6392
1000	No	78.6%	0.5802
1	Yes	76.2%	0.4682
10	Yes	77.8%	0.5105
1000	Yes	81.9%	0.6368

Table 2: Parameter tuning on the reduced feature tuning set.

our GATE application to utilize the features identified as being more useful resulted in an approximation that captures the spirit of what was learned, rather than an exact match, for practical reasons. The GATE Learning Framework machine learning integration [2] makes it easier to simply include the diagnosis field, for example, having shown itself to be of value, rather than picking the diagnoses of interest.

Feature selection wasn't performed on the feature set that included unigrams, so therefore we are interested to see results on this set to get a heuristic feel for whether unigrams are of value, although one can't rule out that had feature selection been performed on the unigrams, some of them would have been found to be of utility. We proceed therefore with three datasets; the full feature set including unigrams (419531 features), the full feature set without unigrams (3256 features) and the reduced set of 2027 features. Note that the reason the reduced feature sets number thousands despite the list being short as above is that a nominal feature is expanded out to a number of numeric (count) features equivalent to one per unique value found in the training set. Table 3 shows the impact of feature set reduction on the results obtained with each algorithm.

In all cases, reducing features results in an improvement, marginal for SVM and JRip but substantial for Random Forest, indeed being required to bring the result obtained up to a competitive standard. The main improvement comes from the removal of unigrams. A further contribution of feature reduction lies in the speed gains obtained at training time. The SVM was trained using a cost of 1000 with feature scaling included. We can see that whilst the algorithms respond differently to feature reduction, using the smaller set there is no very clear winner among them.

[2] https://github.com/GenevieveGorrell/gateplugin-LearningFramework

Algorithm	Feature set	Acc.	Kappa
SVM	Full+uni	79.6%	0.5629
SVM	Full	81.4%	0.6254
SVM	Reduced	81.9%	0.6368
JRip	Full+uni	81.3%	0.6234
JRip	Full	81.5%	0.6296
JRip	Reduced	82.0%	0.6349
Rand. Forest	Full+uni	66.46%	0.2385
Rand. Forest	Full	81.5%	0.6136
Rand. Forest	Reduced	82.2%	0.6274

Table 3: Trying different feature sets with different algorithms.

3.2 Class Balancing

Having focused evaluation so far around classification accuracy, the question of how effective our classifiers are at obtaining a high sensitivity (recall) on first episode psychosis cases has not yet been considered. The goal of the work is to enable medical researchers to obtain a sample of positive cases with little cost in the way of missing any, whilst reducing the amount of time they spend rejecting negative cases. Finding as near as possible to all of the relevant cases is the main priority. Precision needs to be high enough to justify the exercise, but there is much more flexibility regarding how high is good enough. A classifier that is tuned to produce as high an overall accuracy as possible will tend to favour the dominant class, since in the case of uncertainty, assigning to the dominant class will tend to be right more often than it is wrong. Therefore some innovation must be introduced to counteract this.

Early experimentation focused on the weights parameter on the support vector machine. Figure 1 gives the ROC curve thus obtained, using threefold cross-validation on the tuning corpus. The

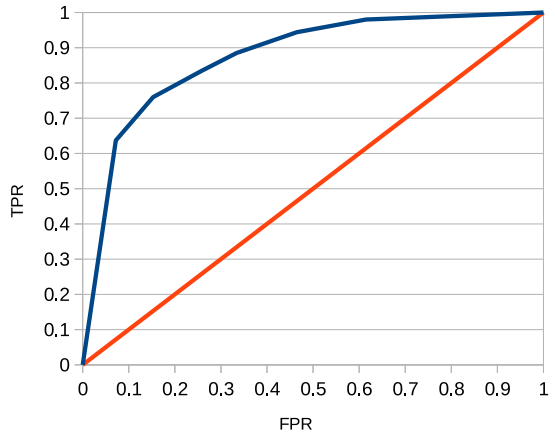

Figure 1: ROC curve for SVM negative class down-weighting

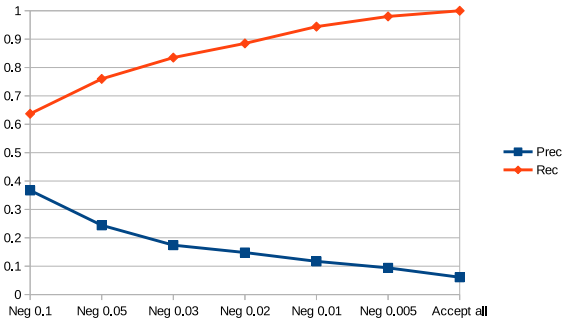

Figure 2: Precision and recall vary with SVM class weighting

AUC (area under the curve) is 0.87. For a recall of 0.944, this gives a specificity of 0.535, which equates roughly to halving the number of cases required to be viewed, at a cost of missing one in 20 cases. It is clear from the graph of precision and recall against weight in figure 2 that the weights parameter provides an effective option for increasing recall of the positive cases to the required level.

Unfortunately this parameter is not available or relevant to the other two algorithms, and also did not transfer easily to the larger training set used to prepare the final application. Further experimentation instead focused on creating a balanced training set that would not penalize the minority class. A balanced training set should lead to a fairer classifier for many algorithms, which aim to minimize the number of misclassified points. Table 4 shows results obtained using Weka to sample the tuning set fairly across classes, having first taken out one third for testing. No replacement of instances was opted for, and the dataset was reduced to 20%,

Algorithm	Conds	Prec	Rec	F1
SVM	0.05	0.244	0.760	0.370
SVM	No	0.544	0.358	0.432
SVM	Yes	0.286	0.675	0.402
JRip	No	0.508	0.258	0.343
JRip	Yes	0.226	0.783	0.351
Rand. Forest	No	0	0	0
Rand. Forest	Yes	0.306	0.725	0.431

Table 4: Class balancing interacts with algorithm choice.

this being large enough to ensure that all positive cases were included, thus thinning the negatives and creating an effect that could be broadly replicated back in GATE by removing some of the negative cases. Separating out a test set is necessary to ensure that the result obtained is indicative of what might be obtained on a naturalistic sample. Had cross-validation been used on an artificially balanced set, the result would have been misleading. The first line in the table gives the most comparable result for weight tuning in SVM ("conds" in this case gives the weight assigned to the negative classes), for comparison. Below that, "conds" indicates whether or not class balancing was used on the training data. We see that for SVM and JRip, class balancing allows recall of the positive class to be improved whilst retaining a broadly similar F1. For Random Forest, class balancing allows us to find the positive cases where previously they were not found at all, and produces a competitive model.

A further option for altering the precision/recall balance lies in making use of the confidence scores provided by the algorithms. However different algorithms are differently able to provide a sensitive and informative confidence score. Confidence scores are made use of in this work to provide the medical researchers with an *ordered* list of cases to review, leaving the power in their hands to progress as far down the list as provides them with the recall they require. This does not negate the need for a classifier tuned to the needs of the task. An appropriately tuned classifier can be expected to give a better F1 for a certain recall than one obtained simply by applying a confidence threshold to a mistuned one. A Random Forest model was trained in GATE using the full tuning set, but with the negative instances thinned to 1 in 13, roughly

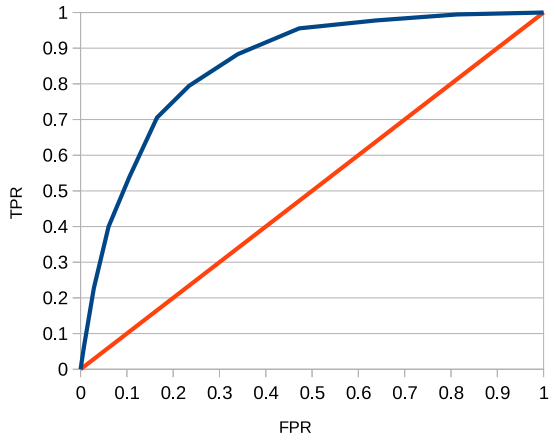

Figure 3: ROC curve based on Random Forest confidence scores in GATE.

balancing the classes. Figure 3 gives a ROC curve based on confidence scores assigned on the test set (AUC 0.85). In keeping with previous results, a recall is obtained of in excess of 95% for a specificity of 0.53, halving the number of cases required to be viewed, by setting the confidence threshold at 0.2. For a recall of almost 0.8, only around a quarter of cases would need to be viewed (specificity 0.77).

4 Conclusion

This paper presents work on a challenging psychiatry domain case classification application. The goal was to facilitate medical researchers' collection of a (further) sample of cases describing a first episode of psychosis, by learning a model from 9109 cases already manually classified. This classification problem, requiring the highest level of domain expertise to accomplish manually, proves challenging for natural language processing techniques. The problem is complicated by the subtlety of distinction between the positive and negative cases; for example, psychotic episodes that are not the first, and those with organic causes are negative instances, although the language surrounding their case is very similar to the positive cases. Furthermore the sample with which we are working is already selected on the basis of psychosis-related keyword search, meaning that the NLP work is required to offer value over and above that. Feature normalization proves essential to making the support vector machine competitive on the task. Feature selection is generally beneficial, in particular making Random Forest competitive, and allowing a much smaller feature space to be used. Since the task is to identify the minority class with a high recall, an important part of task success focused on tuning the algorithms in favour of the positive class. This was accomplished by thinning the negative instances in the training set. Attempts to use the weights parameter with the SVM were complicated by the apparent sensitivity of this parameter to variations in the task conditions. The final GATE application achieves an AUC of 0.85, a result that compares favourably with previous similar work despite the additional challenges, and allows medical researchers to select their own recall based on the confidence score of the Random Forest algorithm, for example halving the number of cases they are required to examine with a loss of only 5% of positive cases. No one machine learning algorithm notably excelled in this work; success might be attributed to an exceptional training set, both in terms of size and quality, and the freely available machine learning technologies that provided a solution to the problems that arose.

References

Alan R Aronson. 2001. Effective mapping of biomedical text to the umls metathesaurus: the metamap program. In *Proceedings of the AMIA Symposium*, page 17. American Medical Informatics Association.

P. Baldwin, D. Browne, P. J. Scully, J. F. Quinn, M. G. Morgan, A. Kinsella, J. M. Owens, V. Russell, E. O'Callaghan, and J. L. Waddington. 2005. Epidemiology of first-episode psychosis: illustrating the challenges across diagnostic boundaries through the cavan-monaghan study at 8 years. *Schizophr Bull*, 31:624–38.

Brandon K Bellows, Joanne LaFleur, Aaron WC Kamauu, Thomas Ginter, Tyler B Forbush, Stephen Agbor, Dylan Supina, Paul Hodgkins, and Scott L DuVall. 2014. Automated identification of patients with a diagnosis of binge eating disorder from narrative electronic health records. *Journal of the American Medical Informatics Association*, 21(e1):e163–e168.

R. Borschmann, S. Patterson, D. Poovendran, D. Wilson, and T. Weaver. 2014. Influences on recruitment to randomised controlled trials in mental health settings in england: a national cross-sectional survey of researchers working for the mental health research network. *BMC medical research methodology*, 14(1).

F. Bourque. 2015. *A mixed methods study of relation between migration, ethnicity and psychosis.* Ph.D. thesis, Kings College London.

Victor M Castro, Jessica Minnier, Shawn N Murphy, Isaac Kohane, Susanne E Churchill, Vivian Gainer, Tianxi Cai, Alison G Hoffnagle, Yael Dai, Stefanie Block, et al. 2014. Validation of electronic health record phenotyping of bipolar disorder cases and controls. *American Journal of Psychiatry*.

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

Wendy W Chapman, Will Bridewell, Paul Hanbury, Gregory F Cooper, and Bruce G Buchanan. 2001. A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of biomedical informatics*, 34(5):301–310.

Hamish Cunningham, Valentin Tablan, Angus Roberts, and Kalina Bontcheva. 2013. Getting more out of biomedical documents with gate's full lifecycle open source text analytics. *PLoS Comput Biol*, 9(2):e1002854.

Elizabeth Ford, John A Carroll, Helen E Smith, Donia Scott, and Jackie A Cassell. 2016. Extracting information from the text of electronic medical records to improve case detection: a systematic review. *Journal of the American Medical Informatics Association*, page ocv180.

T. Greenhalgh, H. W. Potts, G. Wong, P. Bark, and D. Swinglehurst. 2009. Tensions and paradoxes in electronic patient record research: a systematic literature review using the meta-narrative method. *Milbank Quarterly*, 87(4):729–788, Dec.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. 2009. The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18.

Henk Harkema, John N Dowling, Tyler Thornblade, and Wendy W Chapman. 2009. Context: an algorithm for determining negation, experiencer, and temporal status from clinical reports. *Journal of biomedical informatics*, 42(5):839–851.

S. Hogerzeil, A. Van Hemert, F. Rosendaal, E. Susser, and H. Hoek. 2014. Direct comparison of first-contact versus longitudinal register-based case finding in the same population: early evidence that the incidence of schizophrenia may be three times higher than commonly reported. *Psychological medicine*, 44:3481–3490.

Sandy H Huang, Paea LePendu, Srinivasan V Iyer, Ming Tai-Seale, David Carrell, and Nigam H Shah. 2014. Toward personalizing treatment for depression: predicting diagnosis and severity. *Journal of the American Medical Informatics Association*, 21(6):1069–1075.

J. B. Kirkbride, P. Fearon, C. Morgan, P. Dazzan, K. Morgan, J. Tarrant, T. Lloyd, J. Holloway, G. Hutchinson, J. P. Leff, R. M. Mallett, G. L. Harrison, R. M. Murray, and P. B. Jones. 2006. Heterogeneity in incidence rates of schizophrenia and other psychotic syndromes: Findings from the 3-center aesop study. *Archives of General Psychiatry*, 63:250–258.

Andrew Kachites McCallum. 2002. Mallet: A machine learning for language toolkit.

S.M. Meystre, G.K. Savova, K.C. Kipper-Schuler, and J.F. Hurdle. 2008. Extracting information from textual documents in the electronic health record: A review of recent research. *IMIA Yearbook of Medical Informatics*, pages 128–144.

Maxine X Patel, Victor Doku, and Lakshika Tennakoon. 2003. Challenges in recruitment of research participants. *Advances in Psychiatric Treatment*, 9(3):229–238.

RH Perlis, DV Iosifescu, VM Castro, SN Murphy, VS Gainer, Jessica Minnier, T Cai, S Goryachev, Q Zeng, PJ Gallagher, et al. 2012. Using electronic medical records to enable large-scale studies in psychiatry: treatment resistant depression as a model. *Psychological medicine*, 42(01):41–50.

S. M. Powsner, J. C. Wyatt, and P. Wright. 1998. Opportunities for and challenges of computerisation. *Lancet*, 352(9140):1617–1622, Nov.

S. T. Rosenbloom, J. C. Denny, H. Xu, N. Lorenzi, W. W. Stead, and K. B. Johnson. 2011. Data from clinical notes: a perspective on the tension between structure and flexible documentation. *J Am Med Inform Assoc*, 18(2):181–186.

Guergana K Savova, James J Masanz, Philip V Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C Kipper-Schuler, and Christopher G Chute. 2010. Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17(5):507–513.

Donia Scott, Rossano Barone, and Rob Koeling. 2012. Corpus annotation as a scientific task. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uur Doan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may. European Language Resources Association (ELRA).

Lisa Seyfried, David A Hanauer, Donald Nease, Rashad Albeiruti, Janet Kavanagh, and Helen C Kales. 2009. Enhanced identification of eligibility for depression research using an electronic medical record search engine. *International journal of medical informatics*, 78(12):e13–e18.

WHO. 1994. Schedules for clinical assessment in neuropsychiatry: version 2.

Meliha Yetisgen-Yildiz, Cosmin Adrian Bejan, and Mark M Wurfel. 2013. Identification of patients with acute lung injury from free-text chest x-ray reports. *ACL 2013*, page 10.

Qing T Zeng, Sergey Goryachev, Scott Weiss, Margarita Sordo, Shawn N Murphy, and Ross Lazarus. 2006. Extracting principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system. *BMC medical informatics and decision making*, 6(1):1.