Journal of the American Medical Informatics Association, 26(11), 2019, 1247–1254
doi: 10.1093/jamia/ocz149
Advance Access Publication Date: 12 September 2019
Research and Applications

Research and Applications

Evaluating shallow and deep learning strategies for the 2018 n2c2 shared task on clinical text classification

Michel Oleynik ,[1] **Amila Kugic**,[1] **Zdenko Kasáč**,[1] **and Markus Kreuzthaler**[1,2]

[1]Institute for Medical Informatics, Statistics and Documentation, Medical University of Graz, Graz, Austria, and [2]CBmed GmbH - Center for Biomarker Research in Medicine, Graz, Austria

Corresponding Author: Michel Oleynik, Institute for Medical Informatics, Statistics and Documentation, Medical University of Graz, Auenbruggerplatz 2, 8036 Graz, Austria; michel.oleynik@stud.medunigraz.at

Received 16 January 2019; Revised 29 June 2019; Editorial Decision 22 July 2019; Accepted 31 July 2019

ABSTRACT

Objective: Automated clinical phenotyping is challenging because word-based features quickly turn it into a high-dimensional problem, in which the small, privacy-restricted, training datasets might lead to overfitting. Pretrained embeddings might solve this issue by reusing input representation schemes trained on a larger dataset. We sought to evaluate shallow and deep learning text classifiers and the impact of pretrained embeddings in a small clinical dataset.

Materials and Methods: We participated in the 2018 National NLP Clinical Challenges (n2c2) Shared Task on cohort selection and received an annotated dataset with medical narratives of 202 patients for multilabel binary text classification. We set our baseline to a majority classifier, to which we compared a rule-based classifier and orthogonal machine learning strategies: support vector machines, logistic regression, and long short-term memory neural networks. We evaluated logistic regression and long short-term memory using both self-trained and pretrained BioWordVec word embeddings as input representation schemes.

Results: Rule-based classifier showed the highest overall micro F_1 score (0.9100), with which we finished first in the challenge. Shallow machine learning strategies showed lower overall micro F_1 scores, but still higher than deep learning strategies and the baseline. We could not show a difference in classification efficiency between self-trained and pretrained embeddings.

Discussion: Clinical context, negation, and value-based criteria hindered shallow machine learning approaches, while deep learning strategies could not capture the term diversity due to the small training dataset.

Conclusion: Shallow methods for clinical phenotyping can still outperform deep learning methods in small imbalanced data, even when supported by pretrained embeddings.

Key words: natural language processing, data mining, machine learning, deep learning

INTRODUCTION

Background and significance

Clinical narratives stored in electronic health records show considerable variability in format and quality, with their natural language sometimes being described as idiosyncratic.[1] On the one hand, structured data in electronic health records are often created for administrative purposes only and are thus biased toward those diagnoses and procedure codes relevant for billing. On the other hand, semantic tagging of unstructured clinical texts, generally considered the most detailed source of information, is not commonly used and requires prospective planning.[2] Nevertheless, there is an increasing demand to unlock unstructured data to foster primary and secondary uses.[3]

One of such secondary uses would be to support observational research such as cohort, cross-sectional, and case-control studies.[4] A

system that analyzes the content of clinical narratives to recruit patients according to selection criteria could help mitigate sampling bias (eg, by creating a matched control group in a case-control study or by drawing both case and control groups from large cohorts).[5]

To process such data for secondary uses, natural language processing (NLP) techniques must be employed to structure the meaning behind human language into a computer-readable representation. However, researchers cannot easily reuse NLP models from the general domain on clinical text due to significant linguistic differences, especially its tendency toward brevity, often characterized as telegraphic style.

A common challenge in building such systems for the clinical domain is the lack of public corpora of annotated clinical narratives due to privacy concerns. While the availability of huge data silos in the general domain sparked the big data revolution by using complex neural networks to model the diversity of the human language with human-like accuracy, the same has not yet happened in small data scenarios, in which models have commonly been trained from scratch. To address that, there has been a rise of interest in transfer learning methods to reuse models trained on large collections in restricted settings with minimal annotation effort.

Transfer learning

One can transfer knowledge from a larger dataset using various approaches, depending on whether the source and target labels are available and what is reused.[6] One common approach is the so-called feature representation transfer, in which an input representation scheme learned in an unsupervised way in a large corpus is reused in a small annotated dataset.[6] Nonetheless, Goodfellow et al[7] argue that the popularity of this approach has declined, because deep learning achieves human-level performance when large labeled datasets are available and Bayesian methods outperform pretraining on small data.

In the NLP area, Mikolov et al[8] eased feature representation transfer with the release of the word2vec embeddings, trained on around 100 billion words from a Google News corpus. However, clinical text typically shows a low coverage rate in this model due to rare words and misspellings, which has driven the search for alternative input representation schemes.[9]

Bojanowski et al[10] proposed enriching word vectors with subword information to take morphology into account. In parallel, Joulin et al[11] released fastText, an efficient implementation of multinomial logistic regression to allow large-scale linear text classification, in clear contrast to the trend of deep learning approaches. More recently, the National Center for Biotechnology Information used fastText to train word embeddings on around 30 million documents from PubMed and the MIMIC-III (Medical Information Mart for Intensive Care)[12] clinical dataset and released BioWordVec.[13,14] Taken together, these resources may help address clinical idiosyncrasies that hinder transfer learning from big data to the clinical domain.

Clinical text classification

Clinical text classification, also referred to as text-based patient phenotyping,[15–17] aims at automatically assigning a finite set of labels to raw clinical text.[18,19] Historically, several strategies have been employed to address this problem, from rule-based systems, known to provide near-optimal results,[20] to systems based on machine learning (ML), including support vector machines (SVMs),[18] naive Bayes,[21] and decision trees.[21–24] More recently, approaches based on deep learning have been studied, including long short-term memory (LSTM) using hand-engineered features,[25] as well as convolutional neural networks (CNNs) with rule-extracted trigger phrases[26] and word2vec embeddings.[27]

When introducing BioWordVec, Chen et al[14] showed that CNNs trained with fastText embeddings obtained from PubMed and MIMIC-III improved results of a clinical text classification task when compared with models trained with embeddings from each corpus separately or without embeddings at all, even though no comparison was made to embeddings trained on the target dataset. Roberts[28] evaluated the impact of word2vec embeddings trained on multiple corpora (including the target dataset) by applying LSTM and CNN models to 2 downstream tasks: a concept recognition task and a multiclass text classification task, respectively. He showed that models with embeddings trained on several corpora outperformed models with embeddings trained on a single collection but did not consider embeddings with subword information.

2014 i2b2/UTHealth shared task track 2

To promote NLP research in the health domain, the National Center for Biomedical Computing has been organizing since 2006 the Informatics for Integrating Biology and Bedside (i2b2) challenges (https://www.i2b2.org/). The 2014 i2b2/UTHealth shared task track 2 explored the problem of clinical text classification in small data and asked participants to classify patients according to 8 heart disease risk factors (diabetes mellitus, cardiovascular disease [CAD], hypertension, hyperlipidemia, obesity, smoking, family history, and medication).

Participating teams explored several strategies, from rule-based systems to hybrid systems, with different combinations of features and machine learning algorithms.[29] The abundance of hybrid systems led the organizers to conclude there was no consensus about which approaches were better suited for the task. They also found that pseudo-tables encoded in text and CAD indicators were especially hard for most of the teams, resulting in comparatively low F_1 scores.

The best team in 2014 obtained an overall micro F_1 score of 0.9276 by reannotating two-thirds of the training corpus and then training SVM models associated with custom-built lexica to classify triggers for each risk factor.[30] Documents were preprocessed to identify section headers, negation markers, modality words, and other output from the ConText tool,[31] but did not use other syntactic and semantic cues. They also showed that such fine-grained annotations could have helped other automated systems.

Kotfila and Uzuner[32] performed a systematic comparison of feature spaces, weighting schemes, kernels, and training data sizes regarding the efficiency of SVM classifiers trained on the same data. They reported that minimal feature spaces (only lowercased alphabetic tokens) performed as well as combinations with lexically normalized tokens and semantic concepts extracted via MetaMap;[33] *tf-idf* was not a significant factor to determine efficiency (compared with a count-based weighting scheme); and linear kernels were not statistically significantly worse than radial kernels. Finally, they concluded that larger corpora might not be necessary to achieve high efficiency with SVM models.

In 2018, the Department of Medical Informatics of the Harvard Medical School assumed the organization of what is now called National NLP Clinical Challenges (n2c2) (https://n2c2.dbmi.hms.harvard.edu).

The first track focused on the problem of cohort building for clinical trials and framed it as a multilabel binary text classification task.

Research problem

It is challenging to automatically classify clinical text because word-based features quickly turn this task into a high-dimensional problem. On top of that, large corpora are seldom shared due to ethical concerns, while training complex models on small datasets may lead to overfitting. Pretrained embeddings might solve this issue by reusing unsupervised input representation schemes trained on a larger dataset and fine-tuning them using a small annotated dataset.

Therefore, considering (1) previous satisfactory results with shallow methods[11,32] that challenge approaches based on deep learning and (2) the recent availability of pretrained embeddings with subword information in the clinical domain[14] that supports training deeper models in small data scenarios, we decided to participate in the 2018 n2c2 shared task track 1 and use its data to contribute results that may help elucidating these topics. To the best of our knowledge, this is the first study to assess the BioWordVec pretrained embeddings in a text classification task apart from the original work.

Objective

We sought to evaluate shallow and deep learning text classifiers and the impact of pretrained embeddings with subword information in a small clinical dataset.

Hypothesis

We hypothesize that shallow strategies for text classification outperform deep learning strategies in small clinical datasets and that pretrained embeddings increase classification efficiency in the clinical domain.

MATERIALS AND METHODS

Data

We participated in the 2018 n2c2 shared task track 1 and received a small training dataset (70%) with 202 annotated files (patients) containing 887 medical narratives written in English. Two months afterward, we received an extra test set (30%) with 86 new patients (377 narratives) so that we could run our systems and send our results.

Each file had a sequence of 2-5 narratives and was annotated at the patient level with "met" or "not met" for 13 criteria (see Table 1). Of the 13 criteria, 6 were highly imbalanced in the dataset (1 class with <10% of the 202 training samples), 1 semibalanced (with "met" only in approximately 20% of the samples), and the remaining 6 balanced (minority class with at least one-third of the examples). Moreover, 2 criteria were value-dependent: *Hba1c* and *Creatinine*.

Table 1. Overview of the target classification criteria in the 2018 n2c2 shared task track 1

Criterion	Balance	Description
Abdominal	Balanced	History of intra-abdominal surgery.
Advanced-cad	Balanced	Presence of advanced cardiovascular disease.
Alcohol-abuse	Imbalanced	Current weekly alcohol use over recommended limits.
Asp-for-mi	Semibalanced	Use of aspirin to prevent myocardial infarction.
Creatinine	Balanced	Serum creatinine above the normal limit.
Dietsupp-2mos	Balanced	Use of dietary supplements in the last two months.
Drug-abuse	Imbalanced	Drug abuse.
English	Imbalanced	The patient can speak English.
Hba1c	Balanced	Glycated hemoglobin levels between 6.5% and 9.5%.
Keto-1yr	Imbalanced	Ketoacidosis in the last year.
Major-diabetes	Balanced	Major complication due to diabetes.
Makes-decisions	Imbalanced	The patient can make decisions by himself.
Mi-6mos	Imbalanced	Myocardial infarction in the last six months.

Balanced criteria had the minority class with at least one-third of samples; the semibalanced criterion *Asp-for-mi* had "met" in around 20% of samples and imbalanced criteria had 1 class with <10% of the training samples.

Evaluation metrics

Participating teams were evaluated using precision, recall, and F_1 score across the thirteen criteria and the 2 possible classification outputs: "met" and "not met." Overall F_1 was the simple mean of the individual F_1 scores for the classes "met" and "not met." The final metric used for ranking was overall micro F_1 score in the test set. We additionally considered "met" and "not met" as positive and negative outcomes, respectively, and thus also report overall accuracy per criterion.

Preprocessing

We preprocessed input text as follows: (1) removal of spurious whitespaces (as defined by the Java programming language), (2) sentence detection using a customized rule-based algorithm that deals with common abbreviations and artificial new lines, (3) tokenization by the Unicode Text Segmentation algorithm (http://unicode.org/reports/tr29/) as implemented by Lucene 7.5.0 (https://lucene.apache.org/), (4) lowercasing, (5) stop words removal using the SMART[34] system's list of 524 common words, and (6) punctuation removal.

Word embeddings

We used the BioWordVec (https://github.com/ncbi-nlp/BioSentVec) embeddings with subword information pretrained on PubMed and MIMIC-III as available online. To evaluate its impact, we also trained word embeddings from scratch in the target dataset (n2c2) using fastText with the same hyperparameters: window size of 20, learning rate of 0.05, negative sample size of 10, and maximum length of word n-grams set to 6.

Methods

Table 2 shows an overview of the assessed strategies. We evaluated only orthogonal (nonhybrid) strategies to ease comparison among methods. Our baseline was a majority classifier, which always assigns the dominant class seen in training data. We also built a rule-based classifier (RBC) to better understand the data and the noise present therein. We then trained ML-based classifiers using

Table 2. Overview of the evaluated methods and their characteristics

Acronym	Classification method	Word embeddings
Baseline	Majority	N/A
RBC	Rule-based classifier	N/A
SVM	Support vector machine	N/A
SELF-LR	Logistic regression	Self-trained
PRE-LR	Logistic regression	Pretrained
SELF-LSTM	Long short-term memory	Self-trained
PRE-LSTM	Long short-term memory	Pretrained

Pretrained word embeddings were obtained from BioWordVec.
N/A: not applicable.

SVMs, logistic regression (LR), and long short-term memory (LSTM) recurrent neural networks. Additionally, we explored both self-trained (SELF) and pretrained embeddings (PRE) with subword information for approaches using word embeddings as the input representation scheme (LR and LSTM). We contributed the results of RBC, SVMs, and a variant of LSTM (described in Supplementary Appendix A) as official runs for participation at the n2c2 shared task.

Our extensible Java framework is available on GitHub at https://github.com/bst-mug/n2c2 under the open-source Apache License version 2.

Rule-based classifier

We developed a rule-based approach using both regular expressions and textual markers, extended in 4 criteria (*Advanced-cad*, *Asp-for-mi*, *Major-diabetes*, and *Mi-6mos*) with negation and context detection. For the value-dependent criteria *Creatinine* and *Hba1c*, we extracted the corresponding value using a regular expression and compared it to manually defined thresholds, namely $1.4 < \text{creatinine} < 10$ and $6.5 \leq hba1c \leq 9.5$. For the remaining criteria, we manually identified typical text snippets from the training set (such as "elevated creatinine") that, when found, would classify a patient for a given criterion. We enriched these text markers with negative "lookaround" regular expressions to invalidate the marker when it referred to (1) a negated context (eg, "denies ischemia"), (2) drug allergies ("allergy to aspirin"), (3) distant history ("STEMI in 2008"); or (4) family history ("FH with NSTEMI").

Support vector machines

We explored SVMs trained on a bag-of-words representation of the input documents using *tf-idf* (term frequency – inverse document frequency). As text is typically linearly separable,[18] we then applied SVM with a linear kernel. We used the Weka[35] 3.8.2 framework with a LibSVM[36] wrapper to train the SVM classifier. We could not significantly improve the overall micro F_1 score in the training set by using any of the following: (1) cost hyperparameter optimization[37] (default: 1), (2) a lower number of features to avoid overfitting (default: 1000), or (3) L2 normalization on the SVM objective function. Thus, we kept the default values wherever possible.

Logistic regression

LR is a linear machine learning method that is equivalent to a single-layer perceptron (a single-layer feedforward neural network) with a logistic function as output instead of a step function. We trained LR using fastText with 100 epochs, learning rate of 0.50, window size of 5, cross entropy loss function, and a single thread to make results reproducible. We represented the input text as the average of either self-trained word embeddings (SELF-LR) or pretrained BioWordVec embeddings (PRE-LR).

Long short-term memory

We explored a deep learning approach based on a recurrent neural network composed of LSTM cells,[38] a type of architecture that is generally used to model time series events,[25] but can also be used to model natural language for domain-specific tasks.[39,40] Our network had a single LSTM layer with 64 cells and a RNN output layer with sigmoid activation and cross entropy loss function. We employed the Deeplearning4j (https://deeplearning4j.org) 0.9.1 framework to model the LSTM neural network. We used Adam[41] for gradient-based optimization with a learning rate of 0.02 and trained the network for 25 epochs with a dropout rate of 50% to prevent overfitting. Similar to LR, we represented the input text as a sequence of either self-trained word embeddings (SELF-LSTM) or pretrained BioWordVec embeddings (PRE-LSTM).

RESULTS

Tables 3 and 4 depict overall F_1 score and accuracy per criterion, on the test set of the proposed methods when compared with the baseline, a majority classifier. We present detailed results by target class in Supplementary Appendix B.

Baseline

As expected, individual F_1 scores for the majority classifier did not exceed 0.5000, but due to an imbalance among target classes (the most extreme example being *Keto-1yr*), accuracy and overall micro F_1 scores reached high values, thereby setting the baseline at $F_1 = 0.7608$ and $A = 0.7648$.

Rule-based classifier

With respect to individual F_1 scores and accuracies, the RBC showed better efficiency than the baseline for every criterion except *Alcohol-abuse*, on which the RBC had a single false positive due to a missed negation. Rules improved *Dietsupp-2mos* the most, with absolute accuracy (F_1) increase of 0.4070 (0.5800). Imbalanced criteria such as *Alcohol-abuse* and *Makes-decisions* presented the worst F_1 scores; however, owing to micro-averaging, such criteria did not significantly affect the overall efficiency. Conversely, we observed the lowest accuracies in the criteria *Advanced-cad*, *Creatinine*, and *Major-diabetes*.

Support vector machines

Considering individual F_1 scores, SVMs performed equal or better than the baseline on each criterion; however, considering individual accuracies, SVMs performed worse than the baseline for *Asp-for-mi* due to an increase in false negatives (shown as a decrease in recall for "met" and precision for "not met" in Supplementary Table B3). Imbalanced criteria such as *English*, *Alcohol-abuse*, and *Makes-decisions* presented the lowest F_1 scores. Considering accuracy, *Dietsupp-2mos* had the lowest results, followed by *Abdominal* and *Hba1c*.

Table 3. Overall F_1 score per criterion on the test set of the evaluated strategies when compared with the baseline, a majority classifier

Criterion	Baseline	RBC	SVM	SELF-LR	PRE-LR	SELF-LSTM	PRE-LSTM
Abdominal	0.3944	0.8720	0.6028	0.5681	0.5959	0.4930	0.5146
Advanced-cad	0.3435	0.7902	0.7281	0.7109	0.6838	0.5865	0.4788
Alcohol-abuse	0.4911	0.4881	0.4911	0.4911	0.4911	0.4911	0.4881
Asp-for-mi	0.4416	0.7095	0.6063	0.5962	0.6060	0.4948	0.4416
Creatinine	0.4189	0.8071	0.6532	0.7180	0.7399	0.4788	0.5322
Dietsupp-2mos	0.3385	0.9185	0.5814	0.6150	0.6261	0.5903	0.4640
Drug-abuse	0.4911	0.6910	0.4911	0.4911	0.4881	0.4850	0.4881
English	0.4591	0.8644	0.4591	0.4591	0.4591	0.5253	0.5176
Hba1c	0.3723	0.9382	0.6267	0.5393	0.5770	0.4682	0.5137
Keto-1yr	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000
Major-diabetes	0.3333	0.8369	0.7555	0.7518	0.7420	0.4883	0.5435
Makes-decisions	0.4911	0.4911	0.4911	0.4911	0.4911	0.4911	0.4881
Mi-6mos	0.4756	0.8752	0.6815	0.4756	0.4756	0.4658	0.4691
Overall (macro)	0.4270	0.7525	0.5899	0.5698	0.5751	0.5045	0.4953
Overall (micro)	0.7608	0.9100	0.8035	0.8017	0.8063	0.7362	0.7377

Overall F_1 score is the simple mean of the F_1 scores for the classes "met" and "not met."

PRE-LR: pretrained logistic regression; PRE-LSTM: pretrained long short-term memory; RBC: rule-based classifier; SELF-LR: self-trained logistic regression; SELF-LSTM: self-trained long short-term memory; SVM: support vector machine.

Table 4. Overall accuracy per criterion on the test set of the evaluated strategies when compared with the baseline, a majority classifier

Criterion	Baseline	RBC	SVM	SELF-LR	PRE-LR	SELF-LSTM	PRE-LSTM
Abdominal	0.6512	0.8837	0.6512	0.6279	0.6628	0.5233	0.6047
Advanced-cad	0.5233	0.7907	0.7326	0.7209	0.6977	0.5465	0.5465
Alcohol-abuse	0.9651	0.9535	0.9651	0.9651	0.9651	0.9535	0.9651
Asp-for-mi	0.7907	0.8605	0.7558	0.7674	0.7791	0.7442	0.7791
Creatinine	0.7209	0.8372	0.7209	0.7674	0.7907	0.5698	0.6395
Dietsupp-2mos	0.5116	0.9186	0.5814	0.6163	0.6279	0.6047	0.4651
Drug-abuse	0.9651	0.9651	0.9651	0.9651	0.9535	0.9651	0.9651
English	0.8488	0.9419	0.8488	0.8488	0.8488	0.8372	0.8488
Hba1c	0.5930	0.9419	0.6512	0.5814	0.6047	0.6047	0.5465
Keto-1yr	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
Major-diabetes	0.5000	0.8372	0.7558	0.7558	0.7442	0.5349	0.5465
Makes-decisions	0.9651	0.9651	0.9651	0.9651	0.9651	0.9651	0.9651
Mi-6mos	0.9070	0.9651	0.9302	0.9070	0.9070	0.9070	0.9070
Overall	0.7648	0.9123	0.8095	0.8068	0.8113	0.7504	0.7522

Overall accuracy is calculated with "met" and "not met" being considered as positive and negative outcomes, respectively.

PRE-LR: pretrained logistic regression; PRE-LSTM: pretrained long short-term memory; RBC: rule-based classifier; SELF-LR: self-trained logistic regression; SELF-LSTM: self-trained long short-term memory; SVM: support vector machine.

Logistic regression

Both SELF-LR and PRE-LR showed equal or better F_1 scores than the baseline for all criteria (except *Drug-Abuse* for PRE-LR); considering accuracy, SELF-LR (PRE-LR) had better results on 4 (6) criteria: *Advanced-cad*, *Creatinine*, *Dietsupp-2mos*, and *Major-diabetes* (*Abdominal*, *Advanced-cad*, *Creatinine*, *Dietsupp-2mos*, *Hba1c*, and *Major-diabetes*); slightly worse results on 3 (2) criteria: *Abdominal*, *Asp-for-mi*, and *Hba1c* (*Asp-for-mi* and *Drug-abuse*); and was in a tie in the remaining 6 (5) criteria. Similar to SVM, imbalanced criteria had the lowest F_1 scores and *Hba1c* showed the lowest accuracy results for both SELF-LR and PRE-LR.

Long short-term memory

With respect to individual F_1 scores, SELF-LSTM (PRE-LSTM) showed worse results than the baseline in the criteria *Drug-Abuse* and *Mi-6mos* (*Alcohol-Abuse*, *Drug-Abuse*, *Makes-decisions*, and *Mi-6mos*). Conversely, with respect to the accuracy, SELF-LSTM (PRE-LSTM) showed better accuracy results than the baseline in the criteria *Advanced-cad*, *Dietsupp-2mos*, *Hba1c*, and *Major-diabetes* (*Advanced-cad* and *Major-diabetes*). As expected, SELF-LSTM (PRE-LSTM) mostly improved F_1 score for balanced criteria, with *Dietsupp-2mos* (*Major-diabetes)* showing the most substantial absolute increase in F_1 score: 0.2518 (0.2102).

DISCUSSION

The results in the previous section showed that RBC had the highest classification efficiency, followed by shallow methods (SVM and LR), which had similar scores among themselves and above the baseline. Apart from overall macro F_1 score, the deep learning method (LSTM) showed results worse than the baseline. We could not show however a significant difference in classification efficiency of using pretrained word embeddings when compared with embeddings trained on the n2c2 dataset.

We further analyzed false positives and false negatives to obtain deeper insights about the data, discuss our limitations, and propose future work.

False positives and negatives analysis

A high textual diversity compared with the amount of available balanced training data contributed to setting the limits of our approaches. For example, although in our context "chest pain" could be considered a synonym for angina (and thus a predictor for *Advanced-cad*), it was commonly used in negated sentences such as "denies [...] chest pain," "negative for [...] chest pain," and "chest pain free." To keep our methods orthogonal, we did not explore the impact of rule-based or automated negation detection.

Similarly, context also played a role. For instance, "renal transplant" was a common indicator for an intra-abdominal surgery (and thus a predictor for *Abdominal*); we found, however, that its meaning was sometimes changed by nearby words, as in the excerpt "postponing her renal transplant." Even though a bag-of-bigrams approach might have helped, training such a model would have needed more data due to the larger dimensionality.

A specific case of context is family history. It showed more prominently in the selection criteria *Makes-decisions* (in which "dementia" would have been an important feature if it had not been used in sentences such as "Father had dementia") and *Mi-6mos* (in which "MI" was an indicator for myocardial infarction not only for the patient but also for relatives, such as in the sentence "father died of MI").

Another common source of errors for our strategies were value-based criteria such as *Hba1c* and *Creatinine*. To handle numbers in ML strategies without rule-based normalization, we would have needed (1) a large amount of data to capture all values and (2) a nonlinear approach to model both a low and a high threshold. Conversely, we also needed to define a clear classification threshold for rule-based approaches, for which we found inconsistent examples in the training set (eg, creatinine in serum for values close to 1.4 mg/dL).

Limitations

It is known that rule-based approaches do not generalize well and present a maintenance burden. To avoid that, we kept our markers to the bare minimum and used regular expressions only when needed. Together with automated negation and context detection, we believe our method could be reused in other English-speaking institutions with minimal effort. Meanwhile, our shallow approaches (SVM and LR) are language and domain independent; therefore, their reuse in other institutions could be even simpler.

Likewise, we selected features for the SVM classifier based on word frequency only. A better approach would have been to employ output-based or statistical-based methods such as chi-square and information gain. The lack of feature selection may have been the reason the criterion *Dietsupp-2mos* had the lowest SVM results: there is a vast quantity of dietary supplements, often unique in the collection. Nevertheless, our LR approach used a 200-dimensional word vector as input representation, a vector space which should keep semantically-close words nearby, and similarly showed low results for the mentioned criteria.

Furthermore, we did not completely investigate the impact of text normalization. Even though our ML-based methods employed basic tokenization, lowercasing, and stemming, we did not resolve short forms nor misspellings. Our manual data inspection and the high RBC results showed that these were not crucial issues in this dataset, but other domains and languages might require further preprocessing.

Finally, some documents had pseudo-tables for laboratory data, from which we could not extract the proper pieces of information. Our manual analysis of training data showed, however, that physicians would usually emphasize abnormal laboratory results in the text and thus we could capture it using straightforward strategies. In a real clinical setting, structured data may be directly accessible from laboratory systems and thus constitute a better exploratory avenue.

Future work

Future work might explore the dependence between criteria (eg, an episode of myocardial infarction in the last 6 months would trigger not only *Mi-6mos*, but potentially also *Advanced-cad* and *Asp-for-mi*). Experienced NLP researchers might also experiment with sentence parsing to unlock the meaning behind sentences such as "BUN and creatinine were 32 and 1.2."

We opted for independent strategies to ease method comparison and promote interpretability in the shared-task scenario. A real system could benefit from a hybrid approach, using (1) an ensemble of methods (eg, weighted linear combination of signals provided by each approach), (2) stacking strategies (eg, SVM trained on top of count features extracted by the rule-based approach), or (3) a mixed approach (eg, RBCs for imbalanced and value-based criteria and ML-based classifiers for more balanced and complex criteria).

Finally, recent developments in transfer learning that allow reuse of full NLP models trained on large data may help the clinical domain more than input pretraining alone. Special attention should be devoted to Google's BERT[42] and Universal Sentence Encoder,[43] built on top of the ULMFiT[44] and the ELMo[45] models, with pretrained models released not only for the general domain but quite recently also for the clinical domain, the so-called ClinicalBERT.[46]

CONCLUSION

We participated in the 2018 n2c2 shared task track 1 and used its dataset to evaluate shallow and deep learning strategies and the impact of recently released pretrained embeddings for multilabel text classification in small clinical data. We also built a rule-based classifier to provide us with a deeper understanding of the underlying data. We submitted to the shared task the results of 3 orthogonal strategies (RBC, SVM, and a variant of LSTM) to support method comparison. We also contributed our extensible Java framework to the community under an open-source license.

Our rule-based classifier showed the highest overall micro F_1 score of 0.9100, with which we finished first in the shared task. Shallow strategies showed lower overall micro F_1 scores (SVM: 0.8035, SELF-LR: 0.8017, PRE-LR: 0.8063), but still higher than the deep learning strategy (SELF-LSTM: 0.7362, PRE-LSTM: 0.7377) and the baseline (0.7608) set to a majority classifier. We could not show however a significant difference in classification efficiency of using pretrained word embeddings when compared with embeddings trained on the n2c2 target dataset.

Together with the inter-rater agreement scores released by the task organizers, our top-ranking RBC contributed to practical upper bounds for each selection criteria, which might guide other researchers with directions for further improvement. It also provides the community with a reliable method for clinical phenotyping, which can also be reused in a fine-grained way and thus allow further experiments with deep learning approaches.

We also discussed that clinical context, negation, and value-based criteria hindered shallow machine learning approaches. Even though pretrained word embeddings could not alleviate these issues,

we see potential in novel transfer learning techniques that allow reuse not only of feature representation schemes but also full classification models, thus bridging the gap between big and small data.

Taken together, our study suggests that rule-based and shallow methods for clinical phenotyping can still outperform deep learning methods in small imbalanced data, even when augmented with pretrained embeddings with subword information.

FUNDING

MO is funded by the Brazilian National Research Council - CNPq (project number 206892/2014-4).

AUTHOR CONTRIBUTIONS

MO analyzed false positives/negatives and implemented the code framework and the rule-based, support vector machine, and logistic regression approaches. AK analyzed the data, implemented zoning, and presented the work at the conference. MK implemented the long short-term memory and the bidirectional long short-term memory approach. ZK analyzed false positives/negatives, revised the manuscript, and gave overall clinical feedback.

SUPPLEMENTARY MATERIAL

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

ACKNOWLEDGMENTS

We thank our group leader, Stefan Schulz, and Alexandra Pomares for the article revision. This work is part of the IICCAB (Innovative Use of Information for Clinical Care and Biomarker Research) project within the K1 COMET Competence Center CBmed, funded by the Austrian Federal Ministry of Transport, Innovation and Technology; the Austrian Federal Ministry of Science, Research and Economy; the Austrian state of Styria (Department 12, Business and Innovation); the Styrian Business Promotion Agency; and the Vienna Business Agency. The COMET program is executed by the Austrian Research Promotion Agency.

CONFLICT OF INTEREST STATEMENT

None declared.

REFERENCES

1. Meystre SM, Savova GK, Kipper-Schuler KC, *et al*. Extracting information from textual documents in the electronic health record: a review of recent research. *Yearb Med Inform* 2008; 17: 128–44.
2. Hebal F, Nanney E, Stake C, *et al*. Automated data extraction: merging clinical care with real-time cohort-specific research and quality improvement data. *J Pediatr Surg* 2017; 52 (1): 149–52.
3. Safran C, Bloomrosen M, Hammond WE, *et al*. Toward a national framework for the secondary use of health data: an American medical informatics association white paper. *J Am Med Inform Assoc* 2007; 14 (1): 1–9.
4. Mann CJ. Observational research methods. Research design II: cohort, cross sectional, and case-control studies. *Emerg Med J* 2003; 20 (1): 54–60.
5. Geneletti S, Richardson S, Best N. Adjusting for selection bias in retrospective, case–control studies. *Biostatistics* 2008; 10 (1): 17–31.
6. Pan SJ, Yang Q. A survey on transfer learning. *IEEE Trans Knowl Data Eng* 2010; 22 (10): 1345–59.
7. Goodfellow I, Bengio Y, Courville A. *Deep Learning*. Cambridge, MA: MIT Press; 2016.
8. Mikolov T, Chen K, Corrado G, *et al*. Efficient estimation of word representations in vector space. *arXiv* 2013 Sep 7 [E-pub ahead of print].
9. Arnold S, Gers FA, Kilias T, *et al*. Robust named entity recognition in idiosyncratic domains. *arXiv* 2016 Aug 24 [E-pub ahead of print].
10. Bojanowski P, Grave E, Joulin A, *et al*. Enriching word vectors with subword information. *Trans Assoc Comput Linguist* 2017; 5: 135–46.
11. Joulin A, Grave E, Bojanowski P, *et al*. Bag of tricks for efficient text classification. In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*; 2017: 427–31. https://www.aclweb.org/anthology/papers/E/E17/E17-2068/ Accessed May 3, 2019.
12. Johnson AEW, Pollard TJ, Shen L, *et al*. MIMIC-III, a freely accessible critical care database. *Sci Data* 2016; 3 (1): 160035.
13. Zhang Y, Chen Q, Yang Z, *et al*. BioWordVec, improving biomedical word embeddings with subword information and MeSH. *Sci Data* 2019; 6 (1): 52.
14. Chen Q, Peng Y, Lu Z. BioSentVec: creating sentence embeddings for biomedical texts. *arXiv* 2019 Jun 19 [E-pub ahead of print].
15. Shivade C, Raghavan P, Fosler-Lussier E, *et al*. A review of approaches to identifying patient phenotype cohorts using electronic health records. *J Am Med Inform Assoc* 2014; 21 (2): 221–30.
16. Hripcsak G, Albers DJ. Next-generation phenotyping of electronic health records. *J Am Med Inform Assoc* 2013; 20 (1): 117–21.
17. Pathak J, Kho AN, Denny JC. Electronic health records-driven phenotyping: challenges, recent advances, and perspectives. *J Am Med Inform Assoc* 2013; 20: e206–11.
18. Joachims T. Text categorization with support vector machines: learning with many relevant features. In: Nédellec C, Rouveirol C, eds. *Machine Learning: ECML-98*. Berlin: Springer; 1998: 137–42.
19. Lewis DD, Schapire RE, Callan JP, *et al*. Training algorithms for linear text classifiers. In: *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information retrieval-SIGIR '96*. Zurich: ACM Press; 1996: 298–306.
20. Chiticariu L, Li Y, Reiss FR. Rule-based information extraction is dead! Long live rule-based information extraction systems! In: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Seattle: Association for Computational Linguistics; 2013: 827–32. http://www.aclweb.org/anthology/D13-1079 Accessed January 11, 2019.
21. Wilcox A, Hripcsak G. Classification algorithms applied to narrative reports. *Proc AMIA Symp* 1999; 1999: 455–9.
22. Khan A, Baharudin B, Lee LH, *et al*. A review of machine learning algorithms for text-documents classification. *J Adv Inf Technol* 2010; 1: 4–20.
23. Yang Y. An evaluation of statistical approaches to text categorization. *Inf Retr* 1999; 1 (1/2): 69–90.
24. Schütze H, Hull DA, Pedersen JO. A comparison of classifiers and document representations for the routing problem. *In: Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York: ACM; 1995: 229–37.
25. Lipton ZC, Kale DC, Elkan C, *et al*. Learning to diagnose with LSTM recurrent neural networks. *arXiv* 2017 Mar 21 [E-pub ahead of print].
26. Yao L, Mao C, Luo Y. Clinical text classification with rule-based features and knowledge-guided convolutional neural networks. *BMC Med Inform Decis Mak* 2019; 19 (S3): 71.
27. Karimi S, Dai X, Hassanzadeh H, *et al*. Automatic diagnosis coding of radiology reports: a comparison of deep learning and conventional classification methods. In: *BioNLP*. 2017: 328–32.
28. Roberts K. Assessing the corpus size vs. similarity trade-off for word embeddings in clinical NLP. In: *Proceedings of the Clinical Natural Language Processing Workshop (ClinicalNLP)*. 2016: 54–63. https://aclweb.org/anthology/papers/W/W16/W16-4208/ Accessed June, 25 2019.
29. Stubbs A, Kotfila C, Xu H, *et al*. Identifying risk factors for heart disease over time: overview of 2014 i2b2/UTHealth shared task Track 2. *J Biomed Inform* 2015; 58: S67–77.

30. Roberts K, Shooshan SE, Rodriguez L, *et al*. The role of fine-grained annotations in supervised recognition of risk factors for heart disease from EHRs. *J Biomed Inform* 2015; 58: S111–9.
31. Harkema H, Dowling JN, Thornblade T, *et al*. Context: an algorithm for determining negation, experiencer, and temporal status from clinical reports. *J Biomed Inform* 2009; 42 (5): 839–51.
32. Kotfila C, Uzuner Ö. A systematic comparison of feature space effects on disease classifier performance for phenotype identification of five diseases. *J Biomed Inform* 2015; 58: S92–102.
33. Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proc AMIA Symp* 2001; 2001: 17–21.
34. Salton G. *The SMART Retrieval System—Experiments in Automatic Document Processing*. Upper Saddle River, NJ: Prentice-Hall; 1971.
35. Hall M, Frank E, Holmes G, *et al*. The WEKA data mining software: an update. *SIGKDD Explor Newsl* 2009; 11 (1): 10–8.
36. Chang C-C, Lin C-J. LIBSVM. *ACM Trans Intell Syst Technol* 2011; 2 (27): 1–27.
37. Hsu C-W, Chang C-C, Lin C-J. A practical guide to support vector classification. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.224.4115&rep=rep1&type=pdf Accessed November 28, 2018.
38. Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Comput* 1997; 9 (8): 1735–80.
39. Gao S, Young MT, Qiu JX, *et al*. Hierarchical attention networks for information extraction from cancer pathology reports. *J Am Med Inform Assoc* 2017; 25: 321–30.
40. Jagannatha AN, Yu H. Bidirectional RNN for medical event detection in electronic health records. *Proc Conf* 2016;2016:473–82. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5119627/ Accessed May 10, 2017.
41. Kingma DP, Ba J. Adam: a method for stochastic optimization. *arXiv* 2017 Jan 30 [E-pub ahead of print].
42. Devlin J, Chang M-W, Lee K, *et al*. 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, Volume 1 (Long and Short Papers)*. 2019: 4171–86. https://aclweb.org/anthology/papers/N/N19/N19-1423/ Accessed June 29, 2019.
43. Cer D, Yang Y, Kong S, *et al*. Universal Sentence Encoder. *arXiv* 2018 Apr 12 [E-pub ahead of print].
44. Howard J, Ruder S. Universal language model fine-tuning for text classification. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2018: 328–39. https://www.aclweb.org/anthology/papers/P/P18/P18-1031/ Accessed May 3, 2019.
45. Peters M, Neumann M, Iyyer M, *et al*. Deep contextualized word representations. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics, Human Language Technologies, Volume 1 (Long Papers)*. 2018: 2227–37.
46. Alsentzer E, Murphy J, Boag W, *et al*. Publicly available clinical BERT embeddings. In: *Proceedings of the 2nd Clinical Natural Language Processing Workshop*. 2019: 72–8. https://aclweb.org/anthology/papers/W/W19/W19-1909/ Accessed June 29, 2019.