Analysis of EHR Free-text Data with Supervised Deep Neural Networks

Duncan Wallace
School of Computer Science
University College Dublin
Dublin, Ireland D04 N2E5
duncan.wallace@insight-centre.org

M-Tahar Kechadi
School of Computer Science
University College Dublin
Dublin, Ireland D04 N2E5
tahar.kechadi@ucd.ie

***Abstract*—In this paper we present an efficient supervised deep neural network architecture to classify patients based solely on free-text notes extracted from their Electronic Health Records (EHRs). In particular, a three-layer Recurrent Neural Network was used in conjunction with the aggregated EHRs of about 127,149 patients from a medical data warehouse. The result forms a key component of an application we name PANNACEA. We evaluated this neural network in the context of competing neural network architectures, comparing the performance of multilayer perceptrons, convolutional neural networks, and recurrent neural networks in relation to the dataset under investigation. We performed evaluation our program to successfully classify the suitability of these patients to the medical service offered based upon a single medical episode.**

***Index Terms*—Bioinformatics, Artificial Neural Networks, Mining text and semi-structured data**

I. Introduction

Data-mining and machine learning in the context of electronic health records (EHRs) promise to advance clinical decision making and the efficacy of treatment offered to patients. However, significant challenges exist in effectively using data stored as free-text notes in patient records for the purposes of predictive modelling. Despite the inherent challenges in using unstructured medical text in an analytical context, the potential value in doing so cannot be underestimated, as these sources are often the most important in effective diagnostics.

Despite the success of feature learning within both textual and image analysis, as well as the rising popularity of deep learning (i.e., learning based on hierarchies of neural networks), these techniques have not been used broadly with EHR data. Recent developments have seen exploration of neural networks in this domain, but as yet no consensus has emerged in relation to the best data format, or network architecture, that should be used in relation to free-text extracted from EHR.

This research will be looking at the data of one of the largest Out-Of-Hours cooperatives (OOH) in Ireland used to facilitate healthcare provided by General Practitioners (GPs). The organisation has provided a corpus of 211000 cases treated in the year 2014.

Our research is fundamentally founded upon providing quantitative models for early identification of patients who are likely to require escalated levels of assistance, in order that these patients be directed towards more appropriate care providers. The next section presents the problems and questions that motivates this research. We discusses state-of-the-art literature in Section III. The proposed solution for addressing the problems is described in Section IV. Electronic health records (EHRs) offer great promise for accelerating clinical research and predictive analysis, and data-mining techniques in order to allow for secondary use of EHRs is a major ongoing area of research.

Our results significantly outperformed those achieved using currently available methodologies available to healthcare providers. These findings indicate that deep learning applied to EHRs can derive patient representations that offer improved clinical predictions, and could provide a machine learning framework for augmenting clinical decision systems.

A. Context

GPs are often the first point of contact for most people when they feel unwell. While the majority of appointments with a GP are made during the working hours, a number of GP Out-of-Hours services also exist for people who require medical treatment when GP practices are closed, such as in the evening, at weekends, and on public holidays. GP OOH services are part of a larger network of unscheduled care providers which also includes emergency ambulances and Accident and Emergency (A&E) Departments.

Out-Of-Hours GP cooperatives deal with a myriad of cases, and in a number of different environments. Despite the non-linear flow of data, all data processed by the OOH concerning a patient ends up in their centralised Electronic Record Patient Database.

However, not all patients who are presented to the OOH are suited to the services that they provide. Non-emergency, chronic conditions are difficult to find adequate redress within an OOH's operative scope [1], [2]. Although such patients may represent only a small minority of the cases presented to co-ops, they can take up a disproportionate amount of an organisation's time. For the year 2014, such cases consumed a total of 6.77 thousand hours of call operators' time in the organisation under investigation.

This project was partly funded by Science Foundation Ireland [SFI/12/RC/2289]

Yet, individual cases may not be recognisable to call handlers as belonging to patients with escalating care requirements. Rather, this may only become obvious after the fact due to excessive interaction between the organisation and such patients on an ongoing basis.

Electronic Health Records used in the operation of OOH coops are composed of numerical data, normalised text, and free-text clinical notes. Free-text medical notes can be any length. Identification of patterns within this data, at either a population or individual level, is currently limited to shallow treatment and narrow parametrisation [3].

Extracting features from non-annotated, unstructured text creates distinct challenges. Clinical text is a patois designed to be human interpretable by persons with health care backgrounds who can infer meaning from context. The manner in which information is recorded is not only inconsistent from person to person, but moreover, is dirty: typically exhibiting spelling mistakes, shorthand, colloquial phrases, and truncated grammar. Although best practise guides exist which describe ways in which medical terms should be recorded by healthcare professionals, e.g. [4], there is no standard for the industry as a whole and adoption by individuals may be discretionary. Typically, as is the case in the dataset under investigation, medications, signs, and diseases can be written using a multitude of variants. The terse writing style and abbreviation of many words consequently restricts the number of Natural Language Processing (NLP) approaches which would be suitable to use upon the corpus [5]. Entirely ignoring syntactic context in favour of lexeme specific approaches (e.g. Bag-of-Word variants) [6] would not necessarily be appropriate. Not only is context important in developing robust means of cleaning the data, but the environment within which lexemes appear may have a substantial bearing on their importance (for instance, contextual negation) [7].

Traditional healthcare models, whereby medical treatment was limited to a single option provided somewhat procrustean solutions to patient needs, are slowly giving way to concepts concerning both personalised healthcare and community based intervention. Nowadays, the development of decentralised regional programmes is increasing the avenues of treatment available for patients. The thesis of this research is that through the mining of Electronic Medical Records containing mixed types of data, and extracting patterns from the processed data, patients can be successfully categorised through means of supervised machine learning early in their engagement with healthcare providers. This categorisation has quite narrow parameters: the aim of which is to filter patients that are less suitable to the healthcare provider being examined in the course of this research; specifically, an Out-Of-Hours Coop.

The patient records treated in this research may be presented in media res, having had prior interaction with healthcare that is not recorded, and possible further interaction with health infrastructure, that is likewise unavailable within the scope of this project. This is particularly the case if the patient is referred to hospital, for while the data that is recorded by the co-op on the patient may be forwarded to the hospital, the reverse is not true. As such, if diagnoses are made in relation to these patients in subsequent analysis, these diagnoses are unlikely to feature within the data possessed by the co-op, unless specifically mentioned by the patient in potential subsequent interaction between the co-op and the patient.

While this factor relating to the data in question is to some degree a technical deficiency owing to the manner in which telemedicine works in this particular environment, even were an exhaustive medical history of the patient available to the call-handler, its practical value would be diminished by the necessity for call-handlers to promptly handle and conclude calls. Manually parsing large volumes of medical information relating to the patient in question would hamper the performance of a triage. From the point of view of any algorithm that is developed within the scope of our research, the potential features available are inherently limited by the incomplete medical histories that are recorded about patients.

II. Problem definition

The aim of this paper is to provide a key component of a decision support, namely the means to predict patients that are likely to require elevated levels of care. For training purposes, this is defined as any patient who called or was referred to the OOH in question more than 40 times in a calendar year.

Decision support systems based upon predicative analysis have had limited application in real world environments. One of the reasons for this is that EHR data is challenging to represent and model due to its high dimensionality, noise, heterogeneity, sparseness, incompleteness, random errors, and systematic biases [8].

A. Feature Selection

The success of predictive analytics is, strictly speaking, largely dependent upon the features that are used. An approach which has typically been employed in the area of EHR analysis is for domain experts to specify clinical variables in an ad-hoc manner. Consequently such approaches have tended to focus on normalised data fields, such as the age, sex, or weight of patients. Due to the difficulty of obtaining reliable results, because of diverse nomenclature and inconsistent typography, less emphasis has conventionally been placed on free-text to this end. Moreover, even in cases where a domain expert has adequately allowed for heterogeneity in the feature search space of the particular non-normalised data he is using, it is nonetheless the case that manual definition scales poorly, does not generalise well, and misses opportunities to discover novel patterns and features.

B. Artificial Neural Networks

Machine learning has widely known industrial applications. Increased adoption of data mining and machine learning in competitive industry largely stems from the increased efficiency offered by knowledge extraction. Whilst much work has been done on disease diagnoses based upon symptoms described, less work has been published on the categorisation of patients.

Artificial Neural Networks, due to their capacity to model any function, have had proven applicability in relation to data that has traditionally proven problematic in probability analytics and machine learning. In particular, ANNs have shown promising results in the context of natural language processing. Notwithstanding the quality of EHR text, and domain specificity of the language used, ANNs have demonstrated their suitability in processing this data. We used the popular open-source library Tensorflow in the development of the ANN architectures described below.

A key objective of this paper is to derive the optimal neural network architecture for the purposes of medical diagnostics within our planned PANNACEA platform. To this end we conducted a series of comparisons between three different types of neural network over a range of parameters and hyperparameters, in relation to the classification of textual notes, and measured the optimal performance of each. The textual notes themselves were presented to each of these classifiers in different data forms in order to ascertain the ideal format for their processing, and whether a marked difference between the different classifiers emerges based upon the format of the data provided to them.

III. Related Research

The use of machine learning in relation to medical data is by no means new. However, free-text has proven a problematic source of features for classic machine learning techniques. Even when advanced feature extraction methodologies are employed, the inherent noise and heterogeneity of medical natural language remains an issue. Current research within medical informatics have begun to broach ANNs for use with data from Electronic Health Records in order to overcome these challenges.

However, the actual use of neural networks in conjunction with Electronic Health Records (EHR) has displayed wide variance in contemporary research. For instance, some researchers maintain feature extraction techniques, and use neural networks for classification purposes using these extracted features, while others use neural networks for the purposes of feature transformation, and have other types of machine learning algorithms use these features. The architectures and designs of the ANNs similarly are multivariate, with no single structure proving the de facto standard across extant research.

Geraci et al. for instance use multi-layer perceptrons as part of H2O.ai's Deep Learning platform [9]. Similar to our objectives, Geraci et al. seek to produce a binary classification of patients, in their case, as either suitable or unsuitable to be participants for a study on youth depression. The data Geraci et al. were considering was unstructured medical textual notes with limited application of diagnosis codes.

Geraci et al. accepted that searching using a fixed list of terms could potentially achieve a reasonably high accuracy viz-a-viz the correct identification (or exclusion) of potential candidates. Contextual negation was also taken into account for list based searches. On very select training sets this "brute-force" methodology could obtain 80% sensitivity and 88% specificity. However, as expected, this approach did not generalise, and in some cases during cross validation it achieved no better than random accuracy.

Geraci et al. were faced with a relatively small dataset, as classification had to be performed by domain experts, leaving a total of 861 patient documents with which to work. The actual features of the documents were refined into a document term matrix, with the value for each term recorded by its tf-idf ratio (term frequency - inverse document frequency). Somewhat unusually, Geraci et al. produced two separate neural network models, one which had high specificity and poor sensitivity, and one which happened to have high sensitivity but poor specificity. By combining the results of these two neural networks Geraci et al's model achieved a sensitivity of 75% and specificity of 87%. The actual structure of the neural networks was that of a three layer network, with relu activation function on hidden layers. The significant difference between the high sensitivity and specification networks was the number of nodes, with the high specificity network having 758 input nodes, with the high sensitivity network having 102 input nodes. Other aspects relating to the architecture of either network aren't looked at at length.

The use of neural networks using unsupervised learning for the purposes of preprosessing of features is fairly typical in NLP applications. This is increasly common for natural language in medical domains, as can be seen, for instance in Miotto et al.'s DeepPatient classifier project which uses a unsupervised multilayer perceptron (denoising autoencoder) [10]. Although DeepPatient ultimately uses a Random Forest for the purposes of classifying patients, the use of autoencoders clearly improved the quality of the medical textual features presented. DeepPatient's preprocessing saw dramatic improvements in detection of certain disorders, over that achieved using raw text, but this was not shared uniformly. Similarly the use of unsupervised ANNs to produce word embeddings, typically using Google's word2vec, has been employed to improve results of systems such as the named-entity recognition described by Habibi et al. [11]

Rasmya et al. use the Reverse Time AttentIoN (RETAIN) RNN model for the prediction of heart disease based on the extracted information from Electronic Health Records [12]. While this objective is similar to our planned capacity to detect frequent-flier patients, Rasmya et al.'s data does not seem to include natural language. Importantly, Rasmya et al. demonstrated the generalizability of RNNs, given a corpori from different hospitals, and the superior performance of the RNN used in such classification comapred to a more traditional model such as logistic regression.

In counterpoint, C. Yao et al. use a Convolutional Neural Network (CNN) in order to predict correct answers to users' questions in an online Q&A system [13]. The specific structure used in this CNN is inherited from the work by Yoon Kim [14]. A large amount of preprocessing of textual data is performed on the user submitted data, with an emphasis on extracting symptoms, though these features are also vectorized as word embeddings. The accuracy of the model that C. Yoa et al.

create is 71%. It is shown in this research that the quality of the predictions in this case is largely dependent upon the number of symptoms that are present in user data.

Gehrmann et al. also use Kim's CNN architecture to attempt to identify particular patient phenotypes from patterns discovered in their medical notes [15]. In particular, as is the case in our research, specific emphasis was placed on the identification of "frequent flyer" patients; that is to say, patients that have an unusually high amount of contact with the healthcare facility in question. Due to the necessity of manual annotation, the dataset in Gehrmann et al.'s research was small, with some 1610 cases available for both training and testing purposes.

Gehrmann et al. automatically extracted clinical features using cTAKES [15], which were then used as the features in the CNN. Like many of the other papers presented here, these features were provided to the classifier as a bag-of-words. However, prior to their treatment by the CNN, the words were transformed into word embeddings using Word2Vec. The CNN achieved a high F-score across all tests, though results differed significantly, depending on the specific phenotype being tested.

All of the literature discussed above shows the potential strength of artificial neural networks in connection with EHR. However, while many of these papers compare artificial neural networks with other machine learning methodologies, less focus is placed on the specific type, parameters, or hyperparameters of the artificial neural network used within the scope of their analyses.

IV. Methodology

A. Data Processing

In the scope of the investigation conducted by this paper we excluded all normalised data, such as information relating to the age, or sex of the patient, or information concerning the urgency of or amount of time taken in handling patients' calls. This paper exclusively focuses on free-text data, which, notwithstanding its huge potential in medical analytics, has traditionally been the most difficult source of medical data for use in the context of machine learning. The exclusive aim of this paper is to determine the optimal configuration of data and neural network hierarchy for the classification of free-text derived from EHR systems.

In processing the data we cleaned and tokenised the already anonymized free-text notes. Information relating to call frequency of individual patients was derived during extraction, and retained for testing purposes. Cleaning of free-text included the removal of duplicated information, as call handlers may copy-paste case records in order to facilitate access to patient histories. Cleaning also featured the removal of noise (such as the use of symbols to emphasise certain words within the text). We developed our own tokenization program as our data proved too dissimilar from the data that the majority of off-the-shelf tokenenisers were designed for (e.g. typical Twitter text). Measures that were developed included the disambiguation of words, numbers, and symbols. Where appropriate, symbols and letters which represented words were substituted by their real-world equivalents. As such, a section of text that reads "temp+++! - 1/2diazepam, tdsx5" would be changed to " temp ++ 1/2 diazepam , tds * 5". These cleaning and tokenization processes were designed to be conservative, and were by no means exhaustive in their approach. For instance, within the scope of this paper's investigation no attempt was made to alter the free-text in order to correct spelling errors, disambiguate drug names, or unroll acronyms. Each of the data forms that the free text was converted into (such as vectors or list of characters) used this cleaned and tokenised data as its base.

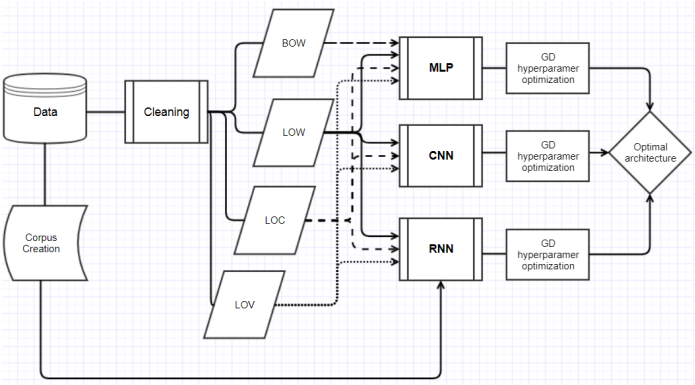

Fig. 1. Model for ANN testing

Data was converted into four mutually exclusive structures: namely bag-of-words, list of characters, list of words and list of vectors. Each of the data formats has potential merits and demerits in terms of their potential capacity to provide useful representations of patient cases for the purposes of classification. Most of the formats include either implicit or explicit contextual information, with bag-of-words alone being formed of an entirely unstructured representation of patient cases. As mentioned in the related research, bag-of-words is a very popular model for use with this type of task. However, due to its unstructured nature, bag-of-words was only suitable for use in conjunction with the multilayer perceptron ANNs.

List-of-words and list-of-characters are representations of the data at either a lexeme or character level. Unlike bag-of-words, the structure inherited from the patient cases is maintained.

B. Word Embeddings

Word Embeddings are low dimensional distributed representations of words in a real-valued vector space. These are richer data than available in atomic representations of lexemes (i.e one-hot vectors). Due to their capacity to exhibit semantic similarity, the use of word embeddings as a form of representation have become increasingly popular in the area of NLP [16].

We created word embeddings using Word2Vec, using the Skip-Gram model [17], with a preprocessed version of the corpus itself being used as the training data. This preprocessing converted all words to lowercase and removed all non alphabetic characters, after tokenization. Sentence disambiguation was also employed on this dataset. Although this

methodology is non-typical in the training of the Word2Vec unsupervised neural network, with most researchers using a pretrained version, the rationale for using the corpus for the purpose of was due to the high frequency of domain specific terms, and local jargon, that would appear infrequently in corpora populated from sources such as news outlets. In particular, contractions and acronyms may exactly resemble common words habitually present in most corpora, but have distinct meaning within the context of the medical institution in question (e.g. "sob" a word which typically refers to the process of crying, specifically means "shortness of breath" in the dataset in question). Our data of word embeddings also maintained the inherent structure present in patient medical notes, as this would likely prove beneficial for neural network architectures which would have the capacity to recognise such structural patterns.

C. Hyperparameter optimization

Choice of hyperparameters can be crucial to the effectiveness of a particular ANN architecture. This becomes a limiting factor due to the fact that suitable hyperparameters may not be known *a priori*. Unfortunately, choice of hyperparameters is subject to high dimensionality. Furthermore, tuning of a single hyperparameter may affect the performance of another hyperparameter; therefore tuning in isolation may not be effective.

The most simple approach to finding the correct combination of hyperparameters is to perform an exhaustive search across all combinations of hyperparameter. This methodology, known as Grid Search, is only suitable within a narrow subset of hyperparamers. Instead we performed Bayesian optimization for hidden layers in relation to the number of layers, number of hidden units, learning rate, activation function, and optimizaion function used within each of the three types of ANN, with batch size alone being subject to manual tweaking. The search space for both Recurrent Neural Network (RNN) and CNN featured dropout, while CNN optimization also included filter, kernel and pool sizes.

D. Long Short-Term Memory

Long Short-Term Memory (LSTM) [18], originally developed to help solve the vanishing gradient problem common in simple Recurrent Neural Networks, is a block that has the capacity to store representations of recent input events. This has proven application in NLP problems due to capacity of the ANN to remember term dependencies, and could be useful in the interpretation of contextual information in patient cases. We implemented a standard LSTM unit, featuring an input gate, an output gate and a forget gate, but the number of LSTM units was variable, with the depth to be determined during hyperparameter optimization.

V. Results

The outputs of the classifiers were a binary truth value based on the determined likelihood that the patient case being tested belonged to someone who either currently, or in the near future, would have escalating levels of required care. A natural issue to arise when classifying rare occurrences is for a machine learning algorithm to fail to learn representations of this occurrence, as assuming an absence of the occurrence will be correct the vast majority of the time. Therefore we randomly split the number of training cases into a even ratio of high-demand and non high-demand patients.

Training ran up to a thousand epochs for each ANN configuration. Testing used a quarter of the corpus, using the best configuration, as derived during hyperparameter optimization.

TABLE I
Testing Results

	Artificial Neural Network		
Data	***MLP***	***CNN***	***LSTM***
BOW	0.8 [a]		
LOC	0.64	0.72	0.73
LOW	0.71	0.75	0.76
LOV	0.73	0.73	**0.82**

[a]Bag of words was only used with MLP

As shown in Table I above, Multilayer Perceptrons were inferior to both CNN and RNN in relation to contextual information. However, MLP performed well using Bag-of-words, underpinning the validity of the frequent usage of this combination, as discussed in Section III.

Character level encoding proved an inferior representation regardless of ANN type. The high dimensionality of feature space was not offset by quality of the data presented to the classifiers, thus generating both long training times and inferior accuracy.

One-hot encoding of lexemes with contextual conservation generated adequate results, and in testing proved the most suitable representation for Convolutional Neural Networks. Although the CNN model achieved good results in all categories, it was nonetheless consistently outperformed by Recurrent Neural Networks using LSTM.

TABLE II
Optimal configuration

Model	μ	units	layers	φ	opt
MLP	0.73639	48	2	leaky relu	GradientDescent
CNN	0.002257	76	3	relu	Adam
LSTM	0.003186	100	3	tanh	GradientDescent

Recurrent Neural Networks using LSTM performed well using all data representations. However, the use of word embeddings saw a significant increase in the potential output of this classifier. While significantly more expensive to train than an MLP network, this combination was the most successful in classifying patients.

The optimal hyperparameters, listed above in Table II relate to the Bag-of-Words model for MLP, the List-of-Words model for CNN, and list of vectors for LSTM. The optimal hyperparameters for CNN included a pool size of 3 and 18 dense units. As the Recurrent Neural Network, with vectors, performed best, its results in testing are shown above in the

TABLE III
TESTING RESULTS

		Prediction outcome: p	Prediction outcome: n	total 13287
actual value	p′	443	12844	P′
	n′	108	60189	N′
total 60297		P	N	

confusion matrix in Table III, while the training cost using this configuration is shown in below in fig 2. A dropout rate of 0.136 was also applied to this network in order to reduce the occurrence of overfitting.

Our deep LSTM network achieved precision of 0.82, recall of 0.8, and an overall F1-score of 0.81.

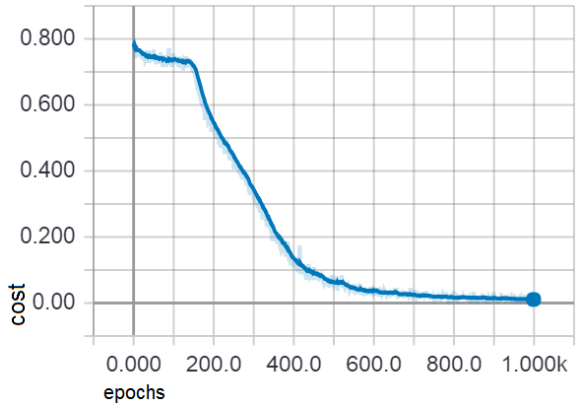

Fig. 2. Training cost for RNN

VI. Conclusion

This paper has not only shown how ANNs can achieve high quality results in classifying patients from medical free-text, but that not all data representations performed as well as one another. A single data representation was likely to perform differently, depending on the type of ANN it was provided to. Moreover, the performance of any given ANN was largely tied to the specific hyperparameters in use. By using Bayesian optimization we were able to discover excellent hyperparameters for use with each of the ANNs, and as such determine the most suitable ANN for use with our data. Notwithstanding the mixed quality of data, and even in the absence of individual patient histories ANNs performed well. This is particularly relevant in our research where the context concerns the classification of a cohort for whom there is not a well defined set of symptoms. As such we did not attempt to extract potentially useful features, but rather allow the neural networks to decide which features were deemed most representative.

Whilst the classifier performed well in classification, the issue remains that the black box model creates challenges in interpretability. The capacity of a classifier to not merely provide the service of identifying members of a cohort, but to inform medical professions the phenotypes likely to represent certain patients is of specific importance within the domain of medical treatment. While providing this capability is outside the scope of this paper, it must be noted that using a more complicated ANN architectures, and more detailed data representations, such as word embeddings, is likely to add to this challenge.

Changes in relation to weight initialization (e.g. He-et-al. initialization) could reduce the initial high learning cost depicted in Fig 2.). This paper leaves significant scope for preprocessing of the data, which can improve the ultimate accuracy of the ANN used. However, the fact that only general preprocessing methods were used with our corpus helps underline the generalizability of the approach taken within the scope of this paper. A potential additional feature than could be developed on foot of this work is to highlight borderline cases which narrowly fall into positive predictions. This could help eliminate false positives, and improve clinician efficiency when interpreting the outputs of the model.

References

[1] Lone Flarup, Grete Moth, Morten Bondo Christensen, Mogens Vestergaard, Frede Olesen, and Peter Vedsted. Chronic-disease patients and their use of out-of-hours primary health care: a cross-sectional study. *BMC family practice*, 15(1):1, 2014.

[2] Dianne den Boer-Wolters, Mirjam J Knol, Kien Smulders, and Niek J de Wit. Frequent attendance of primary care out-of-hours services in the netherlands: characteristics of patients and presented morbidity. *Family practice*, page cmp103, 2009.

[3] Advanced Health and Care Ltd. Helpsheet ce-cen. 2014.

[4] HSE. Health service executive code of practice for healthcare records management. 2010.

[5] Junichi Tsujii. Computational linguistics and natural language processing. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 52–67. Springer, 2011.

[6] Yin Zhang, Rong Jin, and Zhi-Hua Zhou. Understanding bag-of-words model: a statistical framework. *International Journal of Machine Learning and Cybernetics*, 1(1-4):43–52, 2010.

[7] Peter L Elkin, Steven H Brown, Brent A Bauer, Casey S Husser, William Carruth, Larry R Bergstrom, and Dietlind L Wahner-Roedler. A controlled trial of automated classification of negation from clinical notes. *BMC medical informatics and decision making*, 5(1):13, 2005.

[8] Taxiarchis Botsis, Gunnar Hartvigsen, Fei Chen, and Chunhua Weng. Secondary use of ehr: data quality issues and informatics opportunities. *Summit on Translational Bioinformatics*, 2010:1, 2010.

[9] Joseph Geraci, Pamela Wilansky, Vincenzo de Luca, Anvesh Roy, James L Kennedy, and John Strauss. Applying deep neural networks to unstructured text notes in electronic medical records for phenotyping youth depression. *Evidence-based mental health*, 20(3):83–87, 2017.

[10] Riccardo Miotto, Li Li, Brian A Kidd, and Joel T Dudley. Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. *Scientific reports*, 6:26094, 2016.

[11] Maryam Habibi, Leon Weber, Mariana Neves, David Luis Wiegandt, and Ulf Leser. Deep learning with word embeddings improves biomedical named entity recognition. *Bioinformatics*, 33(14):i37–i48, 2017.

[12] Laila R Bekhet, Yonghui Wu, Ningtao Wang, Xin Geng, Wenjin Jim Zheng, Fei Wang, Hulin Wu, Hua Xu, and Degui Zhi. A study of generalizability of recurrent neural network-based predictive models for heart failure onset risk using a large and heterogeneous ehr data set. *Journal of biomedical informatics*, 2018.

[13] Cuili Yao, Yue Qu, Bo Jin, Li Guo, Chao Li, Wenjuan Cui, and Lin Feng. A convolutional neural network model for online medical guidance. *IEEE Access*, 4:4094–4103, 2016.

[14] Yoon Kim. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*, 2014.

[15] Sebastian Gehrmann, Franck Dernoncourt, Yeran Li, Eric T Carlson, Joy T Wu, Jonathan Welt, John Foote Jr, Edward T Moseley, David W Grant, Patrick D Tyler, et al. Comparing deep learning and concept extraction based methods for patient phenotyping from clinical narratives. *PloS one*, 13(2):e0192360, 2018.

[16] Yoav Goldberg. Neural network methods for natural language processing. *Synthesis Lectures on Human Language Technologies*, 10(1):1–309, 2017.

[17] David Guthrie, Ben Allison, Wei Liu, Louise Guthrie, and Yorick Wilks. A closer look at skip-gram modelling. In *Proceedings of the 5th international Conference on Language Resources and Evaluation (LREC-2006)*, pages 1–4. sn, 2006.

[18] Martin Sundermeyer, Ralf Schlüter, and Hermann Ney. Lstm neural networks for language modeling. In *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.