Proceedings of the 52nd Hawaii International Conference on System Sciences | 2019

Mining and representing unstructured nicotine use data in a structured format for secondary use

Mandlenkosi Ngwenya
University of South Africa
School of Computing
43615554@mylife.unisa.ac.za

Prof. Felix Bankole
University of South Africa
School of Computing
Bankofo@unisa.ac.za

Abstract

The objective of this study was to use rules, NLP and machine learning for addressing the problem of clinical data interoperability across healthcare providers. Addressing this problem has the potential to make clinical data comparable, retrievable and exchangeable between healthcare providers. Our focus was in giving structure to unstructured patient smoking information. We collected our data from the MIMIC-III database. We wrote rules for annotating the data, then trained a CRF sequence classifier. We obtained an f-measure of 86%, 72%, 69%, 80%, and 12% for substance smoked, frequency, amount, temporal, and duration respectively. Amount smoked yielded a small value due to scarcity of related data. Then for smoking status we obtained an f-measure of 94.8% for non-smoker class, 83.0% for current-smoker, and 65.7% for past-smoker. We created a FHIR profile for mapping the extracted data based on openEHR reference models, however in future we will explore mapping to CIMI models.

1. Introduction

Health care institutions in South Africa still find it difficult to share, compare, reuse and efficiently query patients' health data on their Health Information Systems (HIS). According to [1] HIS are characterized by fragmentation and a lack of coordination, hence these systems are not interoperable. In order to make the interoperability picture clearer, [2] said in the Eastern Cape (South Africa), the South African Society of Cardio-vascular Intervention has observed that different doctors are not able to share their medical notes. As a result, they don't know the history of the patients' treatments and often during consultations, patients would be requested to do lab scans, lab tests, and be prescribed to medicine that another doctor previously prescribed but that did not work. Furthermore, a report by National Department of Health (NDoH) compiled by the Council for Scientific and Industrial Research (CSIR) shows that more than 70% of HIS used in hospitals do not comply with interoperability standards. Some of those that do comply are not able to exchange health records because the target healthcare institution uses a different HIS, and does not comply with the standard from the source healthcare institution [3].

[4] have defined interoperability in health care systems as the ability of information and communication technology (ICT) systems to share and exchange patients' health data. According to [3, 5], there are four different types of interoperability, namely: technical, syntactical, semantical and organizational. In this study the researchers address the issues of organizational, syntactical and semantical interoperability. Syntactical interoperability looks at the exchange of messages from one system to the other, where messages must have a well-defined syntax, vocabulary, and encoding. While Semantic interoperability is meant to get a common understanding between two messages even though they are phrased differently. Lastly, organizational looks at the ability for organisations to effectively communicate and transfer information to other organisations that are not using the same infrastructural architecture. Later in the study we address transportability of messages where we look at Fast Health Interoperability Resources (FHIR) RestFul API.

In health care, standardization concepts have been considered to be the potential solution to the fragmented and *siloed* health systems [6]. Data management standards have enabled seamless exchange of information and have reduced the complexity when sharing data between multiple systems [7–9]. Getting unstructured data to be semantically interoperable could create value in the delivery of healthcare services. It has also been reported that unstructured data constitutes approximately 80% to 85% of business information according to [10–13], and some of this data is dormant in healthcare. In a hospital setting, vital clinical information is recorded in a human-readable language such as English. Recording the information in a human readable language makes it easier and faster for the clinical personnel to record into an EHR (Electronic

URI: https://hdl.handle.net/10125/59811
ISBN: 978-0-9981331-2-6

Health Record) system than to record the data in a structured format [14]. Unstructured data is often easier to read by humans but it is much more difficult to manage via computers [15]. Even in such a case, the volume of this data is overwhelming for clinicians to manage manually, it has no common physical structure, and it is constantly being received and updated. Such data is high in volume, varies a lot in structure, and is high in velocity, and these are the characteristics of big data [13]. For text data, the application of advanced language processing techniques are used to address these kind of data problems. Therefore, the goal of this study is to apply NLP algorithms for extracting patients' smoking data concepts, then apply a coding standard to standardize the extracted information. Achieving this goal makes this study unique in a sense that it extracts smoking concepts via Ruta rules, and applies a sequence classifier which is trained on a sequence-based data like passages and sentences. Then the extracted details are mapped to FHIR profiles and to a health coding standard for secondary use. Doing so creates structure from unstructured data and it ensures that the mapped data is retrievable, comparable, reusable and exchangeable across healthcare systems.

2. Motivation

Table 1. Entities to be extracted from clinical notes

Data Element	Notes from MIMIC-III database	Data value
Quantity	Quit smoking prior to surgery, smoked **one pack** per 3 days for many years prior.	one pack
Type	Social History: Retired machinist. Recently seperated, lives alone. Active smoker - about 3 **cigs**/day. Admits to 45 pack year history of tobacco. Rare ETOH.	cigs
Frequency	SOCIAL HISTORY: She smokes **half a pack a day** for the past 60 years. She is still currently smoking, but denies any alcohol use.	half a pack a day
Event temporal	Social History: quit smoking **30 yrs ago** lives with wife social ETOH investment attorney	30 yrs ago

The researchers have derived the smoking behaviour challenge from a study by [16] whose findings provided the guidance of how unstructured nicotine use information could be represented as structured data elements. [16] have concluded that there is a need for the use of NLP methods for extracting clinical values from clinical notes and standardizing them. Later on, [17] were able to use a rule-based NLP methods for extracting the use of substances such as drugs, nicotine and alcohol from clinical notes. Their study also mapped the use of each substance to data elements, for example Table 1 shows a real-life example of extracted details from clinical notes. In this study we look at methods for smoking details standardization, and how to exchange the data between organizations. Clinical Element Model (CEM) is an Intermountain Healthcare's initiative that is used for defining the conformant structures and the semantics of clinical data through detailed clinical models. These models are used to normalize clinical text documents including not only Health Level 7 (HL7) messages and HL7 Consolidated Clinical Document Architecture (C-CDA) documents. CEMs enable secondary use of health data and also makes data comparable between multiple health care systems [18]. On the Strategic Health IT Advanced Research Projects (Sharpn), [19] have derived a set of generic CEMs for capturing clinical-based information such as Medications, Signs/symptom, Disease/disorder, Procedures, Labs, and Diagnoses into granular and computable models. Although CEMs were successfully used on the Sharpn project, the developers of CEM have reported that CEM was a short-term project, and they intend to replace it with Clinical Information Modelling Initiative (CIMI) [20]. However, CIMI are still under development, therefore in this study we have opted for the use of OpenEHR standard. OpenEHR is an open standard that can be used for storing, querying and partly exchanging clinical data. It uses a reference model for defining the semantics, data structures, identifiers, data types and more of an EHR system [21]. Therefore, the researchers have used openEHR guidelines for defining smoking details model. Furthermore, openEHR is not primarily concerned with data exchange, therefore we propose FHIR resources and profiles to be used for exchanging data. FHIR represents health information in a form of resources which are used for categorizing medical concepts, for instance the observation resource is used for managing and capturing demographic information, monitoring progress, and for supporting diagnostics. As for FHIR profiles they are used for defining data elements, constraint and relationships between heath data elements that become a building block of the FHIR resource.

3. Materials and methodology

In this section the researchers define the processes and steps that were followed in order to conduct this study. Firstly, we collected data and pre-processed it. Then we wrote rules in a Ruta language in order to annotate clinical notes. Then we trained a classifier, thereafter we tested the model that was produced on unannotated clinical data, and the details are covered in this section.

3.1 Data collection

We used Medical Information Mart for Intensive Care Version 3 (MIMIC-III) database for our experiments. The database is not open-source, however it is accessible to researchers under a data usage agreement [22] and is accessible on the http://mimic.physionet.org website. This database contains patients' demographic information, laboratory tests, medications, ICD9 diagnoses, admitting notes, discharge summaries and pharmacotheraphy, demographics, and a medical history dictionary. It should also be noted that this database complies with the HIPPA (Health Insurance Portability and Accountability Act) regulations, and patient-identifying information was removed. For the purpose of this study, discharge summary data from the noteevents table was used.

The researchers sampled 288 unique records based on the subject_id. These records were filtered by the "*discharge summary*" category and by whether they contained patient's smoking information. Additional filters were applied to exclude: deceased patients, patients younger than 18 years of age, and to exclude records with a true flag for the iserror attribute. The researchers ensured that the retrieved results for all the queries are unique based on every sample that was selected, the uniqueness of a record was based on the subject_id which is unique per patient on the MIMIC-III database.

3.2. Data Preprocessing and components

Unstructured data is said to be difficult to search, classify, and to use. The researchers have proposed the use of NLP tools such as CLAMP or cTakes in order to analyse clinical text data. CLAMP is also known as Clinical Language Annotation, Modelling and Processing, while cTakes is clinical Text Analysis and Knowledge Extraction System and is open-source. Both systems use UIMA (Unstructured Information Management Architecture) as the underlying framework. UIMA is an open-source framework that was originally developed by IBM for processing text, sound and video. Although cTakes offers similar functionality to CLAMP, CLAMP has been reported to provide modern and advanced NLP components and a user-friendly graphical user interface for analysing clinical text [23]. Hence in this study an academic version of the tool was used. Below is a list of NLP components that we have used:

- **Sentence boundary detector**. This component was used for detecting the beginning and the end of a sentence through punctuation marks such as a full-stop or a question mark.
- **Tokenizer**. This component has two sub-functions, firstly it breaks the sentences into tokens that can be analysed further. Then it merges the tokens in order to create date, fraction, measurement, person title, range, roman numerals, and time-based tokens. We used the OpenNLP chunker which is inherent on the CLAMP toolkit.
- **Normalizer**. It is used to produce tokens based on punctuation, spelling variants, stop words, and symbols just to mention a few. Part of speech (POS) functionality detects the type of grammar used on the text data, it assigns tags of tokens such as *patient* to a noun tag.
- **Chunker**. It is used for tagging noun phrases, verb phrases and more.
- **Named Entity Recognition (NER)**. NER is used to extract entities from the given text through rule-based and machine learning approaches. This is one of the most important functions because it is a building block for understanding the semantics of a language [24]. This component allowed the researchers to add a dictionary that helps to map acronyms, abbreviations, and synonyms to common words that will be used throughout the experiments. For instance, words such as "former, past, h/o, hx, quit" were mapped to the "*history*" tag. Furthermore, words were stemmed so that "*smoked*" and "*smoking*" represent a common word which is "*smoke*".
- **Assertion identifier**. This checks if there is a negation associated with a clinical concept. It checks for the absence or opposite of a positive observation, e.g. "*Patient's father has history of alcohol abuse, but patient does not drink alcohol"*. In this case the second part of the sentence regarding the patient is negated, while the first passage about patient's father is not. Therefore, similarly to the NER component, we added a dictionary of words for identifying negated words and phrases.
- **UMLS encoder**. The encoder is used to match the clinical concept terms into UMLS Concept Unique Identifier (CUI) code. Once a term has been mapped to a CUI code, it is then easier to map that term to LOINC or SNOMED or to any coding standard. For instance, nicotine is mapped to the CUI code of C0028040 which has a LOINC code of 3854-7 for the presence of nicotine in urine.

- **UIMA Ruta Rule**. It is a rule engine that is used for identifying, creating and modifying annotations. The rules are used to speed up the process of corpus annotation and they help with feature extraction instead of manually extracting features from corpora. Encompassed in Ruta rule engine is a scripting language that allows for execution of conditional statements, control structures declaration of variables and more [25].
- **Temporal recognizer and relation**. The temporal is able to extract time-specific information such "*last month*, *3rd of August*, *2011-01-02"* and more. There is also a temporal relation which is used for creating relations between the event and the time, e.g. for the passage: "*smoked for five months"*, *smoked* is the event, while *five months* is the temporal recognized.

3.3 UIMA Ruta rule engine

```
TYPESYSTEM ClampTypeSystem;

// 1. rules to parse past smokers;
BLOCK(ForEach) Sentence{} {
// pattern: history of smoking;
ClampNameEntityUIMA{ FEATURE( "semanticTag",
"History") }
    ClampNameEntityUIMA{ FEATURE( "semanticTag",
"Smoker") -> SETFEATURE( "semanticTag",
"PastSmoker" ) };

// 2. rules to parse non-smokers;
BLOCK(ForEach) Sentence{} {
    ClampNameEntityUIMA{ FEATURE( "semanticTag",
"Smoker"), FEATURE( "assertion", "absent" )
        -> SETFEATURE( "semanticTag", "Non-smoker"
) };
}

// 3. rules to parse current smokers;
BLOCK(ForEach) Sentence{} {
// currently smokers
ClampNameEntityUIMA{ FEATURE( "semanticTag",
"TimeModifier") }
    ClampNameEntityUIMA{ FEATURE( "semanticTag",
"Smoker") -> SETFEATURE( "semanticTag",
"CurrentSmoker" ) };
}
```

Figure 1. Rules written in Ruta scripting language for classifying notes to a past smoker, non-smoker and the current smoker class

In this study we use the words corpora and corpus to represent the sampled datasets, corpora is the datasets of text data, and the annotated set of this data is referred to as annotated corpus [26]. The corpora will be annotated using Ruta rules, see Figure 1 for sample rules written via Ruta rule scripting language. The researchers have identified two main tasks which includes: smoking information extraction and smoking status classification. These tasks only covered cigarette smoking behaviour from the clinical notes. Smoking therapy details such nicotine patch or gum are outside the scope of this study and therefore were not explored.

Smoking information extraction task

Subtask 1. Extract nicotine

The researchers created a dictionary of all substances that could be smoked by patients, e.g. cigarette, cigar, pipe and more. The dictionary included abbreviations and acronyms of smoked substances. All cigarette substances were tagged as "*nicotine*".

Subtask 2. Extract quantity and range

Regular expressions were used to extract the amount or range of cigarettes that the patient smokes at a given time. These were tagged using the "*qty*" tag. In a large portion of the clinical notes the quantity is measured in "*packs"* as shown in the example below. However, there are also notes that explicitly state the number of cigarettes smoked without mentioning the pack. Therefore, some of the regular expressions that were used had to check for the occurrence of a numeric value that was succeeded by any of the possible expressions "*pack*|*pck*|pk|cigs|pack of", or one that was succeeded by "*nicotine*" tag. The quantity of smoke often goes together with range, for example, "*patient smokes 3 to 4 packs*". Therefore, range was also extracted through a regular expression such as "-|*to*|*and*", which was preceded and succeeded by a numeric value (also in text). In other instances the quantity was preceded by a symbol, for example "*< 1 pack*". In such cases the researchers used a "-" to tag as less than value and a "+" for opposite case.

Subtask 3. Extract temporal based information

There were three types of time-based values that we wanted to extract based on smoking event. That is "*date: when an event happened*", "*frequency: how often it happened*" and "*duration: how long did it happened*". Below are part of the rules we wrote for extracting these time-based values:

- **Date/Time:** This is a point in time value which represents both relative and absolute time. An example of relative time is "*last year*", while absolute could be "*2010-01-01*". We tagged these values as "*date*". In addition, we have also tagged a range where there is a start and an end date of date as "*interval*" tag.
- **Frequency**: We used frequency as a determiner for the number of cigarettes that the patient has smoked in a day or week or month or in a year. According to [27] this is known as "*pack-year*" whereby each pack contains 20 cigarettes. The pack information is often recorded with acronym "*ppd*", *"ppy"*, "*pyh*" which respectively stands for pack per day, pack per

year, and pack year history. As thus, frequency was extracted with rules that identified the amount tag that was followed by a subordinating conjunction e.g. "*3 packs per day*". Some of the rules tagged "*frequency*" whenever a "*qty*" tag appeared next to temporal value. In some cases the frequency is recorded in the following format, "*2 packs 3 x a week*". Therefore, we added these frequency representations from "*1x*" to "*10x*" in NER dictionary and we mapped them to the "*frequency*" tag.

- **Duration**: Duration in our case is the length of time since the patient started smoking. [27] have emphasized the complexity of extracting duration from the corpora. They said that duration could be represented in multiple ways, which often leads to an overwhelming task when designing patterns to extract the data. However recent studies [28–32] have shown progress in duration extraction. They have also addressed common mistakes to be watchful of during the annotation process. In [31], they have shown how "*duration*" could be confused with "*frequency*". For example the phrase "*every three days*", should be tagged as "*frequency*" however, the presence of "*three days*" alone indicates "*duration*". Therefore, the determiner "*every*" is a distinction between the two phrases. Another example where frequency and duration is used: "*smokes one pack of cigarettes per day x over 50 years*", "*one pack of cigarettes per day*" should be tagged as frequency, whereas "*over 50 years*" is the duration.

Smoking status extraction task

Subtask 4. Extract smoking status

This task is about annotating the given corpus into one of the three classes, namely: current smoker, non-smoker and past smoker. We applied rules at a document and sentence-level and used extraction methods as shown in [33–35]. Shown in Figure 1 is an excerpt of the rules where the first rule states that if the "*history"* tag is followed by the "*smoker*" tag then a new tag past smoker was created as a feature. The second rule states that if there is a "*smoker*" tag followed by a negated tag "*absent*", then tag it as non-smoker. The "*absent*" keyword indicates that the tag is negated, for instance, when the clinical note states that "*the patient denies tobacco use*". The third rule extracts information about the current smokers. A current smoker was tagged for every corpus that stated that the patient has been smoking in the past year. The rules tagged corpus with "*month < 12*" or "*year < 1*" as current smokers and otherwise as past smokers. Part of the rules were constructed by first identifying a temporal value which in this case are time-based adverbs such as currently, momentarily, presently and more.

4. Modelling

We used the components covered from the previous section to annotate clinical notes so that they can be used to train a machine learning classifier. Annotated text provides more information about the text, hence it makes it the metadata of the text. As thus a gold standard was defined in order to annotate the clinical notes.

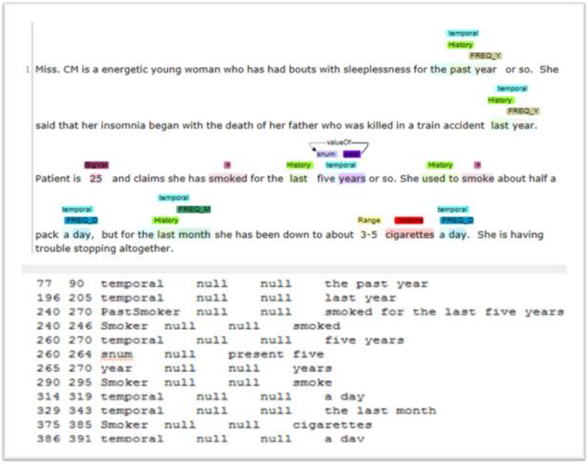
4.1. Gold standard

The annotation process is driven by the expert's advice, for instance, extracting smoking-related information from clinical notes is done by a health informatics annotator. However, [36] have discovered that non-expert annotators can achieve the same performance on a larger training sets as experts do when done on limited set. Experts perform annotation on a limited set because the task is time-consuming, and expensive. Therefore, due to difficulty in finding an already annotated smoking data corpus, and health experts for the annotation task. We have resorted into following the guidelines provided by [16, 17, 33] for creating annotations for patient smoking details, and smoking status and for creating the gold standard. Furthermore, we followed the annotation development cycle as defined by [26]. The gold standard in an annotation development lifecycle is defined as the benchmark and the final version of the annotated corpus which is then used to train the machine learning classifier [26]. Prior to training the classifier, we created rules using UIMA RUTA engine. We executed the rules for the purpose of annotating the top 50 clinical notes that covered task 1 through to task 4. Then we manually observed if the rules captured the annotations as prescribed from the smoking details and smoking status guidelines. We revised the rules until we were satisfied with the outcomes, and usually the correctness of the annotations is calculated from the Inter-Annotator Agreement (IAA) scores. However, in our case rules were used in place of human annotators.

4.2. Selecting features

We extracted features through the following word representation (WR) features were used: (1) clustering-based feature; (2) distributional feature; (3) and word embeddings features. The list of word representation features is as shown below:

- **Brown clustering.** It is a clustering-based word representation algorithm that groups related words into clusters based on the context that these words are in. Then the algorithm partitions the words and outputs the partitions into clusters of words. Lastly it generates an agglomerative hierarchical cluster which is a cluster that implements a bottom up approach [37].
- **Word embedding.** Word embedding feature has the capability to represent words as vectors. Words that are contextually related to one another are represented closer while nonrelated words appear far apart from each other. For instance, tobacco, alcohol and smoking are paired closer to one another on a vector space. We have used a pre-trained word vector "*Wikipedia 2014+Gigaword*" dataset from the https://nlp.stanford.org/projects/glove website. The dataset was trained through the unsupervised GloVe word embedding model. This model has been reported to outperform other models for word analogy, word similarity and NER tasks [38].
- **Random indexing.** This is a form of a distributional word representation technique that has been reported to have human cognitive features such as the ability to make judgements about the quality of an essay or any text-based material that one wants to analyse. [39] have used it for assessing the coherence of words used in a student's essay.



Miss. CM is a energetic young woman who has had bouts with sleeplessness for the past year or so. She said that her insomnia began with the death of her father who was killed in a train accident last year. Patient is 25 and claims she has smoked for the last five years or so. She used to smoke about half a pack a day, but for the last month she has been down to about 3-5 cigarettes a day. She is having trouble stopping altogether.

```
77  90  temporal    null    null    the past year
196 205 temporal    null    null    last year
240 270 PastSmoker  null    null    smoked for the last five years
240 246 Smoker  null    null    smoked
260 270 temporal    null    null    five years
260 264 snum    null    present five
265 270 year    null    null    years
290 295 Smoker  null    null    smoke
314 319 temporal    null    null    a day
329 343 temporal    null    null    the last month
375 385 Smoker  null    null    cigarettes
386 391 temporal    null    null    a dav
```

Figure 2. Extracted and the representation of featured from clinical text

4.3. Conditional Random Fields

Conditional Random Fields (CRFs) is task-specific type of a probabilistic graphical modelling framework. It is used for classifying sequential data through segmentation and annotation. CRFs train a model discriminatively, meaning it learns how to make a conditional prediction of a class (or hidden state) from the given features (or observable states). The framework employs the "*BIO notation*" whereby "*B*" indicates the beginning of the named-entity phrase, "*I*" indicates the inside or the end of the named-entity phrase and "*O*" is other, which indicates that the word is not part of the named-entities [40]. We have also used the above mentioned word representation features. In addition, we also used lexical features where words are represented by their lemmas, part-of-speech, chunking, tokens and the presence of a negation tag for training the CRF classifier.

4.4. Setup and Evaluations

The experiments were executed on Windows 10 Lenovo machine, with the following specifications: Intel (R) Core (TM) i7 7500U CPU, at 2.90GHz; 8GB RAM; 64-bit Operating System. We used the training and test data that was annotated according to the defined gold standard. Then we measured the performance of the CRF classifier using recall, precision, f-measure, macro and micro-average. Recall is the proportion of true positives against the proportion of the sum of true positives and false negatives. True positives and true negatives show agreement between the classifier's predictions and the gold standard, whereas false positive and false negative is an indication of a disagreement. Then precision is the proportion of true positives against the proportion of the sum of true positives and false positives. However, there is a trade-off between recall and precision. An algorithm that achieves a very high precision has low recall and vice versa [41]. Therefore f-measure is used to combine the measures of both precision and recall and calculates the harmonic means of each. Now since task 4 is multi classification problem, therefore we used macro and micro-averaging. Macro-averaging is used for calculating the average precision and recall for all the classes (current-smoker, non-smoker and past smoker). Whereas micro-averaging is used for summing up all the true positives, false positives, and false negative for each class, and this sum is further computed for effectiveness on large classes on the test data [42]. From the annotated corpus, we partitioned the training data into 65%, and the test data to 35% in order to train the CRF classifier.

5. OpenEHR data models to FHIR profiles

We propose the use of openEHR model as a guideline for creating the smoking details FHIR profile. For instance, the extracted details would be represented as shown in Table 2. Table 2 shows data type, value, coding standard code, and the source coding standard.

Temporal information could be parsed via SUTime and stored on the database.

Table 2. FHIR profile data elements

FHIR profile				
Element Name	**Type**	**Value**	**Code**	**Standard**
Amount	int	1	C1265611	SNOMED-CT
Frequency	Date	day	P1D	ISO 8601
Temporal	Date	20 years	P20Y	ISO 8601
Substance	Value Set	Tobacco	C0040329	SNOMED-CT
Smoking status	Codeable	Smoker	72166-2	LOINC

6. Results

Table 3. Smoking status results from CRF classifier

Output from customized rules						
	P	**R**	**F1**	**TP**	**Prd**	**G**
Past Smoker	0.72	0.75	**0.73**	84	116	**112**
Current Smoker	0.57	0.68	0.62	39	68	**57**
Non-Smoker	0.71	0.65	0.68	**60**	**84**	**92**
Macro Avg.	0.67	0.69	0.68			
Micro Avg.	0.68	0.70	0.69			

In this section of the study, the researchers aim to show the results obtained from the application of classification rules and the machine learning sequence classifier. The tasks involved extracting smoking details from clinical notes and classifying each note to one of the three smoking statuses. Therefore the classifier is evaluated on its ability to correctly assign an appropriate class on the correct document based on the established gold standard. If the gold standard matches with the predictions made by the classifier, then that is regarded as a correct prediction. The CLAMP software uses the CRFSuite library to train the CRF classifier, it outputs precision, recall and F1-measure score [43]. We ran five folds of cross-validation for selecting the best model and for optimizing parameters. In each fold, Limited-memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) algorithm was used for estimating the CRF parameters, and the default settings were used for the CRF parameters. Table 3 represents the summarized results which executed for a minimum of three hours for each model. Since our predictions focused on multiple classes. Therefore micro and macro-averaging for the precision, recall and f-measure scores were used instead of a confusion matrix. The results show precision (P), recall (R), f-measure (F1), true positives (TP), predicted (Prd) and the gold standard (G). There were 112 past smoker annotations that met the gold standard. It can be observed on Table 3 that the past smoker had the highest f-measure of 0.73 as compared to the other classes. While the lowest was the current smoker class. The precision and recall results give more information about the class distribution and the correctness of the methods used for identifying correct classes. The class with both the highest precision and recall was a sign that the rules we used were able to detect the smoking statuses in the given corpus. In addition, the test data had enough tests for the calculation of predictions for the same class, meaning there was a good class coverage.

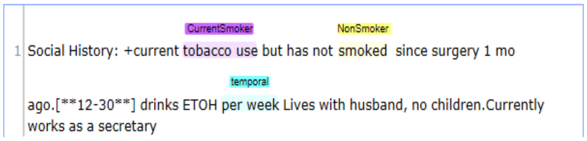
Table 4. Smoking details results from CRF classifier

Output from customized rules						
	P	**R**	**F1**	**TP**	**Prd**	**G**
Type	0.83	0.91	0.87	291	350	319
Amount	0.78	0.38	0.52	109	139	280
Frequency	0.90	0.62	0.73	480	536	773
Temporal	0.80	0.71	0.75	922	1155	1303

Table 4 represents results that were obtained when smoking details were extracted. The results of the second task yielded the highest measure for the type of substance smoked. While the duration measure remained low, however the number of temporal values that met the gold standard was high. One can also observe from these results that micro and macro-averaging was not used since task two is based on a binary classification problem.

7. Discussions

In this study, we had two tasks whereby we needed to classify patient’s smoking status from clinical text, and thereafter extract smoking details where applicable. For both tasks we used a CRF sequence classifier to train the annotated corpus. Few of the things we discovered for task one was that there were sometimes two classes representing the same corpus as shown Figure 4. This reduced the accuracy of the produced model because the number of gold standard records would increase. In return this increased the number of false negatives and ultimately making the recall value lower than what it is supposed to be. Therefore, the researchers had to revise the rules, and also improve the training time so that it becomes easier and efficient to train the model. For the double class annotation problem, we wrote rules for identifying only two classes at a time.



Social History: +current tobacco use but has not smoked since surgery 1 mo ago.[**12-30**] drinks ETOH per week Lives with husband, no children.Currently works as a secretary

Figure 4. Double class annotation where a single document is represented by two classes

Thus, the first set of rules were between non-smoker class and current smoker. Then non-smoker and past smoker class, and lastly between current smoker and past smoker. Another problem was the length of time it took to train the model. This was due to long documents with an average of 3500 words when we were only interested in two or less paragraphs. Therefore, we wrote a small program for extracting smoking information from the clinical notes. Extracting relevant information did not only help with the inefficient processes, however it also gave us the opportunity to train with more and relevant data, since it is known from a classic study by [44] that the classifier's performance improves as more relevant data is added.

The training time was shortened to less than 2 minutes for more than 300 records. We noticed that when more relevant data was added the accuracy of the results improved, the f-measure score for the non-smoker increased to 0.95, while current smoker increased to 0.82 and the only decrease was from past smoker class with 0.54. The performance of our algorithm for the non-smoker class has surpassed that of [35] for document-level classification by a percent. [35] did a similar study where they focused on transferability of the smoking status detection module at different institutions, however they only covered smoking status without focusing on smoking details. On the other side, an earlier study by [34] had obtained a much higher f-measure of 97% for the non-smoker detection class at a document-level. As we were adding more training data, the f-measure of the past smoker increased from 0.54 to 0.66, we also noticed that the rules were robust because we were able to reuse the same rules for an unknown set of test data. This implied that they could be implemented for extracting smoking status from other health institutions. However, it is worth mentioning that these results were as good as the data that was used, in this case the MIMIC-III data. Therefore, the results might be influenced or biased by the manner in which the health clinician captured the data. As for the smoking details extraction task, we observed that the classes were properly balanced for the type of substance smoked, and frequency had a high precision because it is usually represented as "*ppd*" which made it simple to extract. However, it was more challenging to identify the amount because it is often concatenated to the frequency value, e.g. "*1ppd*", hence the low recall. Low recall means that there were not enough training examples for the amount tag. Our result come short when comparing with those obtained by [17] for the smoking details except the temporal which in our case was 80% while they obtained 78.4%. More time of this study was spent in writing the rules in UIMA Ruta, in addition the learning curve for this engine was steep, and it contributed to the results obtained in this study.

8. Conclusion

This study was aimed at extracting smoking details and classifying clinical notes to non-smoker, current smoker and past smoker classes, then ultimately standardize the data. The researchers have used NLP methods to create value from data that is difficult to use for secondary purposes. We have explored various data modelling standards and we ended up using openEHR models because of their accessibility, interoperability, and openness. However, we used the openEHR as a guideline for creating a FHIR profile. In Future we would like to cover abstinence goals in substance abuse which includes alcohol, smoking, and drug, and then map the extracted information to CIMI model and to a FHIR profile. Furthermore, since the grammar used to represent smoking information is sometimes incorrect and this study did not cover grammar issues in clinical notes. However, this could be a future study where one uses sentence-level grammatical error identification concepts.

References

1. Masilela, T. C., Foster, R., & Chetty, M. (2013). *The eHealth Strategy for South Africa 2012-2016: how far are we? South African Health Review Review*.
2. The Competition Commission South Africa. (2016). Private Healthcare Inquiry Public Hearings Day 3, 18 February Live at CSIR #HMI. Retrieved April 17, 2018, from https://youtu.be/-cn_elW48X0

3. CSIR, & NDoH. (2014). National Health Normative Standards Framework for Interoperability in eHealth in South Africa: Version 2.0, (March). Retrieved from http://hufee.meraka.org.za/Hufeesite/staff/the-hufee-group/paula-kotze-1/hnsf-complete-version
4. Mxoli, A., Mostert-Phipps, N., & Gerber, M. (2014). Personal Health Records: Design considerations for the South African context. In *Design, Development and Research* (Vol. 16, pp. 124–245). doi:http://researchspace.csir.co.za/dspace/handle/10204/7712
5. Lopez, D., & Blobel, B. (2009). A development framework for semantically interoperable health information systems. *International Journal of Medical Informatics*, *78*(2), 83–103. doi:10.1016/j.ijmedinf.2008.05.009
6. Smith, J., Fridsma, D., & Johns, M. (2014). Igniting an Interoperable Healthcare System.
7. Adebesin, F., Kotzé, P., Greunen, D. van, & Foster, R. (2013). Barriers & challenges to the adoption of E-Health standards in Africa. *Proceedings of Health Informatics South Africa 2013 (HISA 2013) Conference*. Retrieved from http://researchspace.csir.co.za/dspace/handle/10204/6910
8. Nagy, M., Preckova, P., Seidl, L., & Zvarova, J. (2010). Challenges of interoperability using HL7 v3 in Czech healthcare. *Studies in health technology and informatics*, *155*, 122–128. doi:10.3233/978-1-60750-563-1-122
9. Gruenheid, A., Dong, X. L., & Srivastava, D. (2014). Incremental record linkage. *Proceedings of the VLDB Endowment*, *7*(9), 697–708. doi:10.14778/2732939.2732943
10. Gharehchopogh, F. S., & Khalifelu, Z. a. (2011). Analysis and evaluation of unstructured data: text mining versus natural language processing. In *2011 5th International Conference on Application of Information and Communication Technologies (AICT)* (pp. 1–4). IEEE. doi:10.1109/ICAICT.2011.6111017
11. Abdullah, M. F., & Ahmad, K. (2013). The mapping process of unstructured data to structured data. In *International Conference on Research and Innovation in Information Systems, ICRIIS* (Vol. 2013, pp. 151–155). doi:10.1109/ICRIIS.2013.6716700
12. Doan, A., Naughton, J. F., Ramakrishnan, R., Baid, A., Chai, X., Chen, F., … Vuong, B. (2008). Information Extraction Challenges in Managing Unstructured Data, *37*(4), 14–20.
13. Raghupathi, W., & Raghupathi, V. (2014). Big data analytics in healthcare: promise and potential. *Health Information Science and Systems*, *2*(1), 3. doi:10.1186/2047-2501-2-3
14. Rosenbloom, S. T., Stead, W. W., Denny, J. C., Giuse, D., Lorenzi, N. M., Brown, S. H., & Johnson, K. B. (2010). Generating Clinical Notes for Electronic Health Record Systems. *Applied clinical informatics*, *1*(3), 232–243. doi:10.4338/ACI-2010-03-RA-0019
15. Barbulescu, M., Grigoriu, R., Halcu, I., Neculoiu, G., Sandulescu, V. C., Marinescu, M., & Marinescu, V. (2013). Integrating of structured, semi-structured and unstructured data in natural and build environmental engineering. In *2013 11th RoEduNet International Conference* (pp. 1–4). IEEE. doi:10.1109/RoEduNet.2013.6511738
16. Chen, E. S., Manaktala, S., Sarkar, I. N., & Melton, G. B. (2011). A multi-site content analysis of social history information in clinical notes. *AMIA ... Annual Symposium proceedings. AMIA Symposium*, *2011*, 227–36. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/22195074
17. Wang, Y., Chen, E. S., Pakhomov, S., Arsoniadis, E., Carter, E. W., Lindemann, E., … Melton, G. B. (2015). Automated Extraction of Substance Use Information from Clinical Texts. *AMIA ... Annual Symposium proceedings. AMIA Symposium*, *2015*, 2121–30. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/26958312
18. Rea, S., Pathak, J., Savova, G., Oniki, T. A., Westberg, L., Beebe, C. E., … Chute, C. G. (2012). Building a robust, scalable and standards-driven infrastructure for secondary use of EHR data: The SHARPn project. *Journal of Biomedical Informatics*, *45*(4), 763–771. doi:10.1016/j.jbi.2012.01.009
19. Pathak, J., Bailey, K. R., Beebe, C. E., Bethard, S., Carrell, D. S., Chen, P. J., … Chute, C. G. (2013). Normalization and standardization of electronic health records for high-throughput phenotyping: the SHARPn consortium. *Journal of the American Medical Informatics Association*, *20*(e2), e341–e348. doi:10.1136/amiajnl-2013-001939
20. Oniki, T. A., Zhuo, N., Beebe, C. E., Liu, H., Coyle, J. F., Parker, C. G., … Huff, S. M. (2016). Clinical element models in the SHARPn consortium. *Journal of the American Medical Informatics Association*, *23*(2), 248–256. doi:10.1093/jamia/ocv134
21. Beale, E. T., Heard, S., Kalra, D., & Lloyd, D. (2008). EHR Information Model.
22. Johnson, A. E. W., Pollard, T. J., Shen, L., Lehman, L. H., Feng, M., Ghassemi, M., … Mark, R. G. (2016). MIMIC-III, a freely accessible critical care database. *Scientific data*, *3*, 160035. doi:10.1038/sdata.2016.35
23. Soysal, E., Wang, J., Jiang, M., Wu, Y., Pakhomov, S., Liu, H., & Xu, H. (2017). CLAMP – a toolkit for efficiently building customized clinical natural language processing pipelines. *Journal of the American Medical Informatics Association*. doi:10.1093/jamia/ocx132
24. Savova, G. K., Masanz, J. J., Ogren, P. V, Zheng, J., Sohn, S., Kipper-Schuler, K. C., & Chute, C. G. (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. doi:10.1136/jamia.2009.001560
25. Kluegl, P., Toepfer, M., Beck, P. D., Fette, G., & Puppe, F. (2016). UIMA Ruta: Rapid development

of rule-based information extraction applications. *Natural Language Engineering*, *22*(1), 1–40. doi:10.1017/S1351324914000114

26. Pustejovsky, J., & Stubbs, A. (2013). *Natural language annotation for machine learning: A guide to corpus-building for applications*. O'Reilly Media. Retrieved from http://www.amazon.com/Natural-Language-Annotation-Machine-Learning/dp/1449306667

27. De Silva, L., Ginter, T., Forbush, T., Nokes, N., Fay, B., Mikuls, T., … Health, V. A. (2011). Extraction and Quantification of Pack-years and Classification of Smoker Information in Semi-structured Medical Records. In *Proceedings of the 28 th International Conference on Machine Learning*. Retrieved from http://www.cs.utah.edu/~alnds/papers/smokers_2011.pdf

28. Lee, H.-J., Xu, H., Wang, J., Zhang, Y., Moon, S., Xu, J., & Wu, Y. (2016). UTHealth at SemEval-2016 Task 12: an End-to-End System for Temporal Information Extraction from Clinical Notes. *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, 1292–1297. Retrieved from http://www.aclweb.org/anthology/S16-1201

29. Chang, A. X., & Manning, C. D. (n.d.). SUTIME: A Library for Recognizing and Normalizing Time Expressions. Retrieved from https://nlp.stanford.edu/pubs/lrec2012-sutime.pdf

30. Madkour, M., Benhaddou, D., & Tao, C. (2016). Temporal data representation, normalization, extraction, and reasoning: A review from clinical domain HHS Public Access. *Comput Methods Programs Biomed*, *128*, 52–68. doi:10.1016/j.cmpb.2016.02.007

31. Kovacevic, A., Dehghan, A., Filannino, M., Keane, J. A., & Nenadic, G. (2013). Combining rules and machine learning for extraction of temporal expressions and events from clinical narratives. *Journal of the American Medical Informatics Association : JAMIA*, *20*(5), 859–66. doi:10.1136/amiajnl-2013-001625

32. Lin, Y.-K., Chen, H., & Brown, R. A. (2013). MedTime: A temporal information extraction system for clinical narratives. *Journal of Biomedical Informatics*, *46*, S20–S28. doi:10.1016/j.jbi.2013.07.012

33. Uzuner, O., Goldstein, I., Luo, Y., & Kohane, I. (2008). Identifying patient smoking status from medical discharge records. *Journal of the American Medical Informatics Association : JAMIA*, *15*(1), 14–24. doi:10.1197/jamia.M2408

34. Sohn, S., & Savova, G. K. (2009). Mayo clinic smoking status classification system: extensions and improvements. *AMIA ... Annual Symposium proceedings. AMIA Symposium*, *2009*, 619–23. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/20351929%5Cnhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2815365

35. Liu, M., Shah, A., Jiang, M., Peterson, N. B., Dai, Q., Aldrich, M. C., … Xu, H. (2012). A study of transportability of an existing smoking status detection module across institutions. *AMIA ... Annual Symposium proceedings. AMIA Symposium*, *2012*, 577–86. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/23304330

36. Kwitt, R., Hegenbart, S., Rasiwasia, N., Vécsei, A., & Uhl, A. (2014). *Do We Need Annotation Experts? A Case Study in Celiac Disease Classification*. Retrieved from https://s.yimg.com/ge/labs/v2/uploads/86740455-2.pdf

37. Collins, M. (2011). Lecture 10 : Discriminative Training for MT / the Brown et al . Word Clustering Algorithm Discriminative Training for MT. *Word Journal Of The International Linguistic Association*. Retrieved from http://www.cs.columbia.edu/~mcollins/courses/6998-2011/lectures/lec11.pdf

38. Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1532–1543). doi:10.3115/v1/D14-1162

39. Higgins, D., & Burstein, J. (2007). Sentence similarity measures for essay coherence. *Proceedings of the 7th International Workshop on Computational Semantics IWCS*, (January), 1–12. Retrieved from http://scholar.google.com/scholar?hl=en&q=Sentence+similarity+measure+for+essay+coherence&btnG=Search&as_sdt=2000&as_ylo=&as_vis=0#2

40. Koller, D., & Friedman, N. (2009). *Probabilistic Graphical Models- Principles and Techniques*. *Journal of Chemical Information and Modeling* (Vol. 53). doi:10.1017/CBO9781107415324.004

41. Gorunescu, F. (2011). Data Mining: Concepts, Models and Techniques. *Data mining - Concepts, Models and Technique*, 1–357. doi:10.1007/978-3-642-19721-5

42. Manning, C. D., Raghavan, P., & Schütze, H. (2009). *An Introduction to Information Retrieval*. Retrieved from https://nlp.stanford.edu/IR-book/pdf/irbookonlinereading.pdf

43. Okazaki, N. (2007). CRFsuite: a fast implementation of Conditional Random Fields (CRFs). Retrieved from http://www.chokkan.org/software/crfsuite/

44. Banko, M., & Brill, E. (2001). Scaling to very very large corpora for natural language disambiguation. *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics - ACL '01*, 26–33. doi:10.3115/1073012.1073017