UNIVERSIDAD DE CHILE
FACULTAD DE MEDICINA
ESCUELA DE POSTGRADO
PROGRAMA DE GRADOS ACADÉMICOS

THESIS

Master in Medical Informatics Program

"Automatic Referral Classification for the Chilean Waiting List"

Student: Fabián Villena Rodríguez
Thesis Advisor: Jocelyn Dunstan, PhD
Thesis Co-Advisor: Mauricio Cerda, PhD
Thesis Co-Advisor: Matthias Ganzinger, PhD

…………………………………… **Thesis Advisor Signature**

…………………………………… **Thesis Co-Advisor Signature**

……………………………… **Master in Medical Informatics Committee President Signature**

Content

1 Abstract

In Chile, some 80 problems are covered by the Explicit Health Guarantees (which stands for *Garantías explícitas en Salud,* GES in Spanish), which means that there is a guaranteed maximum time between the diagnosis and the treatment of the health problem. Patients have the right to be treated in a prioritized way. Misclassification of patients covered by GES lead to be considered in a non-prioritized waiting list, characterized by prolonged waiting times. Furthermore, hospitals get fines for GES misclassification. Also, urgent cases must be treated as prioritized by the institution even if they are not GES.

We propose to design, construct, train and deploy a web service based automatic system that receives a referral and then classifies it using models for GES and urgency class. This solution is based on machine learning algorithms that train models with human-coded historical data from healthcare services. Those models have the intelligence to classify a referral into GES and urgency.

Natural Language Processing techniques were applied to the free-text of the referrals to code the information into a vectorial representation of words. These dense vectors were inputted as features to train different machine learning models.

The best performing algorithm was Random Forest, reaching an F1-score of 0.91 in GES classification and 0.92 in Urgency classification. The platform has been used for 30 weeks in a hospital and 4472 referrals have been analyzed. Human-machine discrepancies were 129 cases, wherein 87 cases the machine was right.

We were able to deploy a production-ready intelligent system to automatically classify referrals into GES and Urgency categories faster than human classification.

2 Introduction

In Chile, 73% of the population is covered by the government-administered health fund (FONASA) [1,2]. Different from the situation in the private sector, where a patient can go directly to a specialist, patients in public healthcare need to be referred to a specialist from a general practitioner in primary care [3].

As a way to deal with health inequalities that lead to prolonged waiting times and higher risks of health deterioration, the Chilean government implemented in 2006 the Plan for Universal Access to Explicit Health Guarantees (which stands for *Garantías explícitas en Salud,* GES in Spanish), which categorize a series of health problems which must be prioritized over other problems [4]. Hospitals get fines [5] and incentives to deal with referrals covered by this plan, which has led to dramatic differences in waiting times and volume of waiting lists (WL) covered and not covered by GES [6]. One of the explicit guarantees is the opportunity guarantee, which specifies a maximum time to perform some of the processes involved in the health problem (namely diagnostic, treatment, surgery, etc.). Patients with health problems covered by GES must not be present in the national repository of waiting lists and there are two more exclusion criteria for the waiting list: referral classified as an Urgency and some special billing codes. The Chilean Healthcare Superintendence in 2018 sanctioned 83 healthcare institutions for incorrect handling of GES cases, reaching 77% of the entire number of sanctions declared in the same year by the superintendence [5]. An automatic curation process that detects these cases could correct human misclassification.

During 2016, 22,459 patients died while waiting for their first consultation with a specialist, and 2,358 died before the surgery [7]. In a recent study, prolonged waiting time was associated with increased mortality in the WL. These numbers compare to the 993 deaths in the non-WL group during the same year [6].

In October 2018, the Chilean Health Superintendence estimated that around 10% of the cases that belong to the GES group were not receiving the prioritized treatment [8]. In other words, patients that have a diagnosis that classifies as GES were misclassified in the WL group. Due to the dramatic differences in waiting times in both groups, the fate of those misclassified patients is very different.

The reason for referral is in the format of free-text, and every hospital in Chile has a person in charge of uploading the WL to the National Registry, this person (typically a nurse) must review each referral, manually checking if the diagnostic suspicion is classified as GES or not. This project aims to detect misclassification by using a machine learning model that receives the reason for referral in a vector representation to further classify the referral as a WL case or not. Data to train the models come from the referrals done in the Servicio de Salud Metropolitano Sur Oriente (SSMSO) by the healthcare professionals in primary care.

2.1 Background

2.1.1 Healthcare Data

The data generated from healthcare sources can be divided into 2 classes (1) structured data, which is a type of data where we a priori know its syntax and we have a controlled semantic inside each of the data points, for instance, in a laboratory report the syntaxis of the document is conserved across every report and the content of the document is controlled by use of nomenclatures to define each of the findings and the results are explicitly described with a metric unit, and (2) unstructured data that cannot be standardized, the syntax of the document and the content of these data points is noisy and ambiguous, for instance, a free-text note from a healthcare professional [9,10].

2.1.2 The Chilean Waiting List

The Chilean waiting list is a nationwide database containing patient referral information in the form of structured and unstructured data. This repository comprises referrals that were not classified as GES by the sender institution. Every institution must upload their referral to the national repository to consolidate all the cases in one place. We are especially focusing on the surgical waiting list, which contains data from patients who need surgery. The national registry does not have all the referrals generated from the institutions because there are certain cases where the referrals need to be handled inside the institution, for instance, the patients with GES-covered health problems. The GES program explicitly guarantees some pathologies and interventions. One of the guarantees is the opportunity guarantee which specifies maximum waiting time for that disease, so every institution must take into account this time while they are managing their referrals and is for this reason that these cases do not have to be consolidated on the national repository.

In current managing process to send the waiting list cases to the repository a nurse reviews the entire number of internal referrals and decides which case go to the repository, for the GES cases, the nurse analyses the diagnostic suspicion and the age of the patients, and manually compares this information with the list of guaranteed pathologies. For the Urgency cases, the nurse analyzes just the diagnostic suspicion and classifies the referral as an Urgency case based on subjective criteria. To decide if the billing code corresponds to a procedure, the professional must search for the code in a predefined list of procedures codes.

2.1.3 Natural Language Processing

Natural Language Processing (NLP) is a subfield of computer science, information engineering and artificial intelligence concerned with the interactions between computers and human (natural) languages, in particular how to program computers to process and analyze large amounts of natural language data [11].

NLP has been used to systematically extract information from narrative to further structure the information into machine-readable features which can feed ML or rule-based algorithms [12]. From the structured textual features created with NLP methods, several models have been developed: cancer cases identification [13], wound information extraction [14] and normalization of multiple types of clinical data into standardized terminologies [15], to mention some.

In healthcare, there are specific applications of NLP [16], such as:

1. Detection of named entities inside the clinical text: Automatic recognition of entities such as healthcare-associated infections, adverse drug events or cancer symptoms. This task in NLP is known as Named Entity Recognition, conversely in medicine, is called Medical Entity Recognition. [11].
2. Text summarization and translation of patient records: This task has been applied to synthesize large healthcare stays into discharge notes and translating patient records into a language that can be understood by laypeople because medical records are often

difficult to understand due to the broad use of many specialized medical terms and domain-specific language.

3. Automatic coding of diagnostics: For billing and epidemiologic purposes there is a large need for mapping diagnostics written in natural language into standardized nomenclatures. In NLP this task can be divided into 2 subtasks, first, we have to detect in the text where are diagnostics (Named Entity Recognition) and then, this text segment have to be fed into a text classification system which is defined as the assignation of a category to a given natural language text input.

Text Classification

The task of assigning a class to an entire text segment is called Text Classification. One of the most common text classification tasks is the sentiment analysis which extracts the positive or negative orientation of a text segment. This classification scheme when has 2 classes is the simplest one, and is called a binary classifier. The general pipeline that the classification system follows is to first extract important features from the given text segment and then classify the text segment into predefined classes.

In the past, the usage of human-generated rules to classify text segments achieved state-of-the-art performance in this task, but rule-based methods are not robust because unstructured data constantly changes in time and also changes from context to context.

The underlying problem with rule-based methods is that the solution is specifically designed to focus on syntax, but the language itself is the conjunction of pragmatics, semantics, syntax, morphology, phonology and phonetics. A solution capable of working with the entire number of the characteristics of the language is not yet available but there are more methods trying to solve the syntax and semantics of the language, for instance, using machine learning. We are going to use this last approach to solve our tasks along with the usage of neural embedding to extract semantics from words.

Supervised machine learning can be applied to text classification, typically in 2 forms, (1) inputting previously features engineered from the current knowledge of the topic to train models, which is expensive and time-consuming because of the need of experts in the specific fields to come up with the best features from the text and (2) inputting raw text to train models without previous feature engineering, this approach is where the current research is focused on, mainly in the subfield of machine learning named deep learning.

Nowadays these text classification tasks are developed based on supervised machine learning [17], which will be explained in the next sections.

This work is entirely based on a machine learning with neural word embeddings approach to classify text segments into 2 predefined classes.

2.1.4 Word Embeddings

Word embeddings are distributed representations of words that map words from vocabulary to dense vectors of real numbers in a low dimensional embedding space. These approaches represent the meaning of words via geometry such that the relationship between vectors mirrors the linguistic relationship among them [11].

There are several algorithms to obtain word embeddings, some of them use neural networks, such as word2vec [18,19] fastText [20] or co-occurrence matrix decomposition, GloVe [21], and all of them give dense vectors that represent words.

Some of the semantic relationships of words that we can obtain from geometric calculations are the semantic similarity, which is then normalized dot-product between vectors, and the semantic relatedness of words, solved by vector addition and subtraction [22]. For example, the words *disease* and *illness* are similar while *disease* and *hypertension* are related, and these relations can be tested using word embeddings.

To compute the neural word embeddings, as shown in Figure 1, a single layer (hidden layer in the figure) neural network is constructed, and to calculate the weights *h* of the hidden layer, two approaches are used. One approach, called continuous bag of words (CBOW), performs the task of predicting a centre word (output layer) of a sentence given context words (input layer). A second option is called skip-gram (SG), which tries to predict the context words (output layer) given a centre word (input layer) in a sentence. Both approaches calculate the weights of the embedding layer, this embedding layer contains vectors for each word of the vocabulary which retain the semantic relationships described before. In both approaches, the input and output words are one-hot encoded as shown in the input and output layers of the figure.

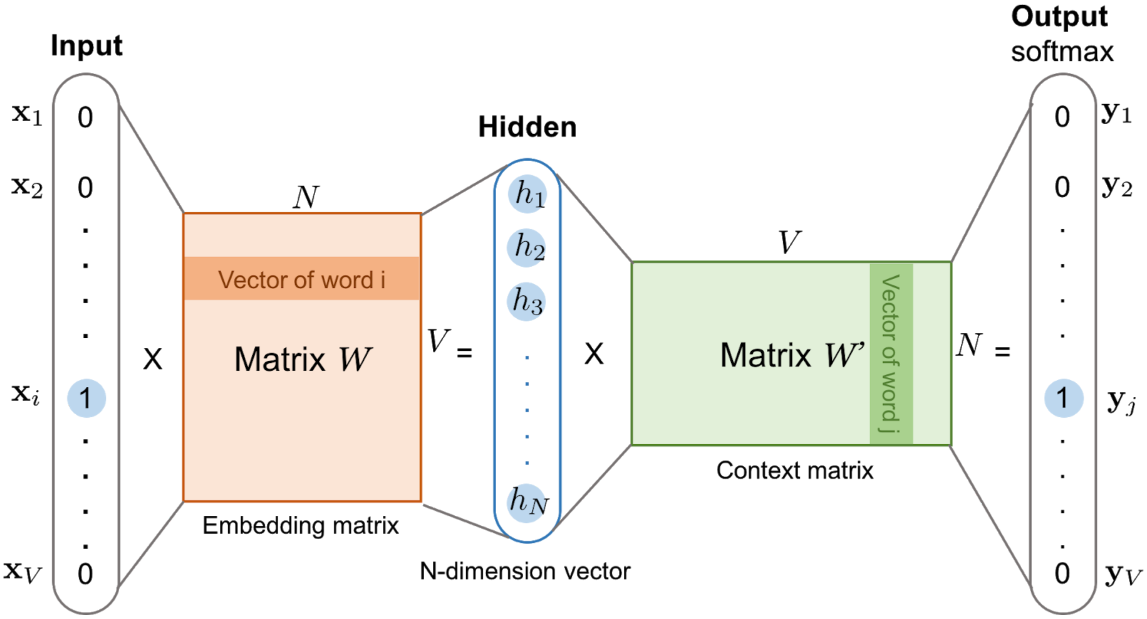


Figure 1: Schematic representation of the neural network architecture to obtain word embeddings. (figure obtained from Lilian Weng Blog: lilianweng.github.io/lil-log)

The reason why the methods for computing word embedding explained previously (CBOW and SG) achieve the goal of extracting the semantics of word is based on the distributional hypothesis of language. The distributional hypothesis states that words that occur in similar contexts tend to have similar meanings, for instance, the words *disease* and *illness* are similar because both of them in the same contexts, for example:

Testicular cancer is a rare disease in men caused…

Testicular cancer is a rare illness in men caused…

Based on this hypothesis, the words disease and illness are going to be very close in the embedding space because the methods to compute word embeddings take into account a centre word and surrounding words [11].

These word representations in a low dimensional embedding space could serve us as inputs to downstream tasks, for example, we can use these vectors to transform natural language text directly into numbers by averaging each of the word vectors in the sentence to retrieve a single vector describing the entire sentence. This sentence vector can be used as features to train supervised machine learning models and perform predictions over natural language text.

Representation Learning

Representation learning is the technique of transforming input data so that the transformed representation can make predictions more easily. For instance, in deep learning, these representations can be derived from the composition of multiple non-linear transformations. The parameters extracted from these non-linear transformations could have basic information about the input data, which can be used to further perform predictions more easily from these transformed data than from the raw data [23]. Data from natural language sources is noisy and ambiguous, therefore using this data for machine learning analysis does not lead us to good results. Transforming raw text sentences using representations, namely word embeddings, lead us to an increased performance of the trained model [24–26] because the usage of high-quality representations to perform predictions is directly related to the final performance of the model.

2.1.5 Supervised Machine Learning

Traditionally, developers have programmed software explicitly to take some data, transform it using a series of rules and then output the result. This approach is time-consuming and poorly scalable when large amounts of data are available.

This rule-based paradigm has been changing with the appearance of machine learning, where we show to a previously-built algorithm the input data, and at the same time we show the output we want, this is called *training* in machine learning, so the program can learn how all the input variables are related to output the desired result. In the end of the training session, we have a trained model which maps the input data to our desired outputs without programming the rules itself for this mapping. This approach of showing the input along with the output data to an algorithm to finally extract a model is called Supervised Machine Learning [27].

Logistic Regression

The logistic regression models the probability that a response belongs to a particular binary class given a set of numerical features. In the logistic model, the probability of response of value 1 is a linear combination of independent variables, namely *features*, which may be binary (coded values as 0 or 1) or continuous (real numbers) [28].

Tree-Based Methods

These methods segment the responses space in regions. The splitting rules to stratify the responses space can be summarized as a tree, so the name of these approaches are known as decision tree methods.

One decision tree is a set of questions (internal node) regarding the value of a certain feature and based on the response, the decision path moves into another internal node. If there are no more questions left, the decision path ends on the node, which has the final decision of the tree (terminal node). The majority class associated with that node is selected as the response class

Random Forests

A single decision tree could not perform well in some settings so another approach emerges. The main idea of Random Forests relies on the construction of a series of decorrelated trees which at the end combines each of the responses to decide the class of the example. To construct decorrelated trees, each one is trained over a subset of features and examples. This idea protects each tree from their individual errors because while some of the trees may be wrong, many of the remaining trees will be right based on the uncorrelation of the entire set of decision trees [29].

Multilayer Perceptron

Using the intuition of the separation of classes using straight lines, the perceptron is a linear classifier which given a set of features, iteratively finds the best set of weights (which multiplies each of the input features) to predict the label of the example. This method only works if the dataset is linearly separable.

The multilayer perceptron utilizes a set of perceptrons to classify non-linearly separable tasks using a layer of perceptrons, each of them receiving the features to further combine each of the outputs of the perceptrons to make a prediction of the label of the example [30].

Support Vector Machines

This method represents each example as points in a space of dimensions equal to the number of predictors to further divide it into two subspaces, one for each class. The basic implementation of the method uses a straight line to divide the space, ensuring that maximization of the distance between each class. This method could perform nonlinear classification distorting the space to ensure a linear separation of the classes before fitting the straight line [28].

2.2 Problem

In Chile, some health problems are covered by Explicit Health Guarantees (GES in Spanish), which means that there is a time limit to diagnose and treat these problems. Moreover, by law, patients have the right to be treated in a prioritized way [4]. Misclassification of patients covered by GES lead to be considered in a non-prioritized waiting list (WL), characterized by prolonged waiting times [6]. Furthermore, hospitals get fines for this GES misclassification and patients die waiting [5]. Also, patients with urgent pathologies must not be in a non-prioritized waiting list [30].

2.3 Clinical Relevance

Managing more efficiently the Chilean WL is an urgent necessity, and the first step to implement an automatized prioritization is finding a representation of the free-text in the reasons for referral. This work creates a representation of diagnostics specifically designed for the Chilean medical context. Implementing machine learning algorithms that operate well with existing tasks such as referral classification inside health institutions, have the potential to make a big difference for the most vulnerable population, empowering justice in the healthcare system at a national level.

2.4 Technical Significance

There is evidence that the quality of word embeddings depends on the specificity of the corpus used to train them, showing that for health applications, a clinical corpus performs better than one trained on the general text [31]. Also, most of the available text for training and published word embeddings are in English. Here, we calculate a word embedding trained over 2 million diagnoses in Spanish using word2vec, a state-of-the-art neural model. The vector representation of reasons for referral is used in a supervised machine learning model. The implementation was performed using an intuitive user interface used by healthcare professionals.

3 Hypothesis

A machine learning model trained over historical human-classified data can separate WL referrals into GES, Urgency and non-prioritized classes with human classification performance, and this classification system is deployable in a real-world hospital context.

4 General Objective

Implement deployable NLP techniques to classify free-text diagnostic suspicions produced in Chile from a patient waiting list to separate referrals into GES, Urgency and non-prioritized classes.

5 Specific Objectives

1. Classify referrals using ML models and compare their performance.
2. Compare the classification proposed by the model with a gold standard that combines the classifications made by healthcare experts.
3. Deploy a pilot classification platform into the workflow of a hospital.

6 Data and Methods

6.1 Data

Labelled data comes from the SSMSO, which attends 1,655,796 patients from 7 communes in Santiago, which corresponds to 8.3% of the Chilean population. The second data source was obtained by Transparency Law (TL) [32], where we requested waiting lists from 29 health services.

The dataset is divided into 3 subsets:

1. Referral subset from the SSMSO containing 2,630,025 data points with previously labelled GES cases.
2. Referral subset from the SSSMO containing 13,249 data points with previously labelled Urgency cases.
3. Waiting list subset obtained through TL, containing data from 23 health services, this database contains 11,826,843 unlabeled referrals.

Labelled data were extracted from the internal repository of referrals of the SSMSO, this repository comprises the entire body of referral documents sent from primary care to secondary care between 2005 and 2018. The healthcare professional who generated the referral must determine if the referral is considered as a GES case or not. The Urgency subset was previously labelled by experts from the health service. The Faculty has a data agreement with the SSMSO, and it is from that agreement that the data for this project is available.

Unlabeled data were extracted from the national repository of waiting lists and this data was requested through TL. The data shared from the SSMSO and TL is administrative data already de-identified and the data agreement and TL specifies the data can be used for research purposes.

6.2 Methods

Word embeddings will be applied to the non-structured diagnostic suspicion. Free text cannot be used directly to perform frequentist analysis or as an input in ML algorithms, so this work will focus on a vector representation approach that systematizes the information contained in the diagnostic suspicion of the waiting list.

We will use word embeddings to define the diagnoses and subsequently use these semantic vectors as input to train ML classification models.

The trained model is going to be deployed as a semi-automatic approach on a web-based platform where humans can use it to review their classifications. When conflicts are found (machine classification differs from human classification) the platform will retrieve the human-corrected class of each referral so the model could be retrained.

The entire pipeline of the data analysis was performed in Python programming language, using the packages Pandas for data structuring and analysis, NLTK for text preprocessing, gensim to model the text representations, scikit-learn for modelling and performance analysis and Flask for the deployment of the web service.

The front-end of the platform which communicates with the classification engine was developed in PHP and JavaScript programming languages using jQuery for communication with the web service and user experience and Bootstrap for the design of the user interface.

6.2.1 Dataset Preprocessing

Raw data from the sources were retrieved as text files in comma-separated values format and were concatenated into a unique dataset. After consolidation, a subset of columns were selected for further usage, the columns selected were:

1. Day of birth
2. Day of the entrance to the waiting list
3. Diagnostic suspicion
4. GES label
5. Urgency label

Columns day of birth and day of the entrance to the waiting lists were parsed into DateTime data formats, taking into account the different date formats used across the dataset.

GES and Urgency labels were normalized into True and False values because of the inconsistency of labelling in the dataset.

Next, rows with undefined values in the columns day of birth, day of the entrance to the waiting list or diagnostic suspicion were dropped.

Three subsets were generated after preprocessing:

1. Corpus subset: This subset is composed only by the free-text data from the diagnostic suspicion column of the consolidated dataset.
2. GES subset: This subset consists of the GES labelled data points of the consolidated subset.
3. Urgency subset: This subset is comprised of the Urgency labelled data points of the consolidated subset.

Text Preprocessing

For the correct analysis of the diagnostic text corpus we have to preprocess the text, we translate all the corpus letters to lowercase, delete all the special characters through a regular expression, and then normalize special letters by their base form (every accented word was translated to its non-accented version and *ñ* was translated to *n*), then we tokenize sentences by splitting the diagnostic suspicions into a list of words.

Training and Testing Subsets Construction

GES and Urgency datasets were divided into training and testing subsets. The training dataset was balanced, where the majority class was subsampled to have the same proportion of True and False labels for the GES subset and upsampled on the Urgency dataset. The testing subset was not balanced.

The ratio of the division of training and testing subsets was 7:3.

6.2.2 Descriptive analysis

For data exploration, the subsets were described using different metrics according to the data content. GES and Urgency subsets were described based on the number of data points, class balance and the date span contained in the dataset. Corpus subset was described by the number of diagnostic suspicions, the mean number of words per diagnostic suspicion, number of tokens in the corpus and vocabulary size.

6.2.3 Features

Based on the current human-based pipeline for referral classification, to categorize the referral according to their GES class, the features used are diagnostic suspicion and patient's age, the age was calculated based on the day of birth and day of entrance to the waiting list, this age was normalized using a min-max scaler, where the highest age value was mapped to 1 and the lowest value to 0. For the referral Urgency class, only the diagnostic suspicion is going to be used and for procedure classification, the billing code is going to be used to filter by administrative rules.

6.2.4 Word and Sentence Embeddings Computation

Word embeddings representations were computed with the Word2Vec architecture using the skip-gram method and the hyperparameters of the computing phase were set to the defaults proposed by the Word2Vec authors.

This part of the workflow allowed us to construct a mapping dictionary which translates each word into a dense vector representation.

For the computation of word embeddings, the corpus subset was used.

To represent each diagnostic (a construction of words, namely a sentence) the method used was a weighted average of word embeddings, where each word vector of the sentence was multiplied by their corresponding Inverse Document Frequency (IDF) and then these vectors were summed and finally divided by the norm of the vector.

The IDF metric is computed as follows,

$$IDF_w = log\left(\frac{n_d}{FD_{d,w}}\right) + 1 ,$$

where n_d is the number of diagnostic suspicions in the corpus and $FD_{d,w}$ is the raw frequency of diagnostic suspicions d which contains the word w. The sentence embedding is calculated as

$$SE_d = \frac{\sum_{i=1}^{n} WE_{w_i} * IDF_{w_i}}{\left| \sum_{i=1}^{n} WE_{w_i} * IDF_{w_i} \right|},$$

where SE_d is the sentence embedding of the diagnostic suspicion d, n is the index of the word in the diagnostic suspicion, WE_{W_i} is the word embedding of the word w in the index i.

Each of the diagnostic suspicions of both the Urgency and GES subsets was transformed into their corresponding vector representations using the method explained above. We used sentence embeddings as the text feature for model training.

6.2.5 Modelling

A variety of machine learning algorithms along with multiple hyperparameters were used to select the best configuration to train two final models, one for GES and other for urgency task. The algorithms selected to train the models were Support Vector Classifier, Random Forest Classifier, Logistic Regression, and Multi-Layer Perceptron. Random number generators were locked to a predefined seed across all the experiments to perform fair comparisons between models.

Initial Grid-Search

Different hyperparameter values were selected to train models with each of the algorithms mentioned. Each combination of the grid search was trained and tested using a 3-fold cross-validation method, AUCROC metric was used to compare between models and selecting the best set of hyperparameters. Each of the hyperparameters used to train the algorithms is described next.

- Support vector classifier:

  - C: 1, 10, 100, 1000
  - Gamma: 1, 0.1 ,0.001 ,0.0001
  - Kernel: linear, RBF
- Random forest classifier
  - Bootstrap: True, False
  - Max depth: 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, None
  - Max features: auto, sqrt
  - Min samples leaf: 1, 2, 4
  - Min samples split: 2, 5, 10
  - N estimators: 200, 400, 600, 800, 1000, 1200, 1400, 1600, 1800, 2000
- Logistic regression:
  - C: 1.e-05, 1.e-04, 1.e-03, 1.e-02, 1.e-01, 1.e+00, 1.e+01, 1.e+02, 1.e+03, 1.e+04, 1.e+05
  - Penalty: l1, l2
- Multi-layer perceptron classifier:
  - Hidden layer sizes: (50, 50, 50), (50, 100, 50), (100, 1)
  - Activation: tanh, relu
  - Solver: sgd, adam
  - Alpha: 0.0001, 0.05
  - Learning rate: constant, adaptive

Best Model Selection

For the purpose of selecting the best algorithm to train a final model, each algorithm was trained using the best set of hyperparameters extracted from the last step, AUCROC metric was used in order to select the best trained model. The method used for model comparison was 10-fold cross-validation.

Statistical significance assessment

With the intention to establish statistical differences across the trained models for the selection of the best performing model, different statistical tests were applied to the results. Shapiro Wilk test was used to determine if the results are normally distributed, paired student's t-tests (or Wilcoxon signed-rank test if there is no normal distribution) were performed across all the combinations of model results to confirm if there are statistical differences in the mean AUCROC across models, finally, retrieved p values were corrected using Bonferroni's multiple tests correction method. An alpha value of 0.05 was selected for null hypothesis rejection.

Best model training

The best algorithm selected using the method described before was used to train a model using the entire training datasets. The trained model was tested over the testing subset and different metrics were reported, this performance assessment is described in the next sections.

6.2.6 Gold Standard Construction and Human Performance

For the creation of a gold standard testing dataset, a random subset of diagnostic suspicions were extracted. The mentioned subset was labelled by three different experts. Where discrepancies were found, the author decided the label based on the information contained in the official documents of the Healthcare Superintendence. The mentioned subset is named the gold standard.

This subset is going to be used to perform classifications by healthcare professionals and by the best model to assess the performance of human and machine classification.

For each of the human-labelled subsets, human performance metrics were calculated using the constructed gold standard to further compare them with the machine metrics.

6.2.7 Performance Assessment

Performance of each trained model is going to be compared using the following metrics [33]:

Precision: The ratio between the number of relevant retrieved data points and the total number of retrieved data points so,

$$Precision = \frac{|(relevant\ data\ points) \cap (retrieved\ data\ points)|}{|(retrieved\ data\ points)|} .$$

This metric can be interpreted as the probability that a retrieved data point is relevant.

Recall: Recall is the ratio between the number of relevant retrieved data points and the total number of relevant data points so,

$$Recall = \frac{|(relevant\ data\ points) \cap (retrieved\ data\ points)|}{|(relevant\ data\ points)|} .$$

This metric can be interpreted as the probability that a relevant data point is retrieved.

F-Score: Metric which returns a unique value that weighs the precision and recall metrics as

$$F_{\beta} = (1 + \beta^{2}) * \frac{Precision * Recall}{(\beta^{2} * Precision) + Recall} ,$$

and the most used F-Score is the F_1 score which is defined as a harmonic mean between precision and recall, in which case beta is equal to 1.

Area Under the Receiver Operating Characteristics Curve (AUCROC): Graphical representation of the recall of the positive class (sensitivity) and the recall of the negative class (specificity) for a binary classifier in which we modulate the threshold where we classify a data point as positive. The area under this curve is used as a measure of the performance of the binary classifier.

6.2.8 Deployment

This method is going to be deployed into production through a Python web service that receives features as a JavaScript Object Notation (JSON) encoded message, processes the features, makes the classification and responds with another message containing the predicted referral classes. This web service is going to be integrated with a web-based graphical user interface that interacts with the healthcare professional and the backend will be integrated with the classification web service.

7 Results

Methodologies described before were applied to the raw dataset and the results are described for both Urgency and GES subsets.

7.1 Dataset preprocessing

Data flow for preprocessing and general data structure is described in Figure 2.

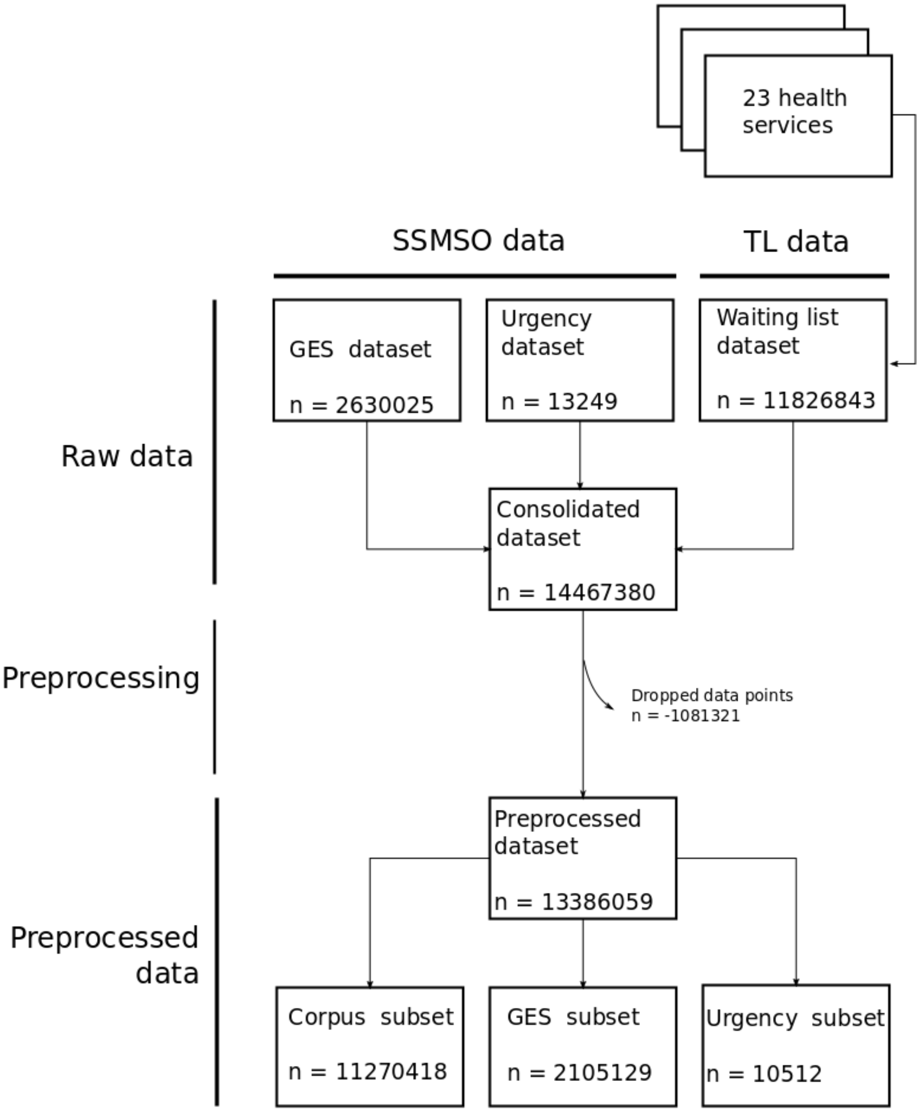

23 health services
SSMSO data
TL data
Raw data
GES dataset n = 2630025
Urgency dataset n = 13249
Waiting list dataset n = 11826843
Consolidated dataset n = 14467380
Preprocessing
Dropped data points n = -1081321
Preprocessed data
Preprocessed dataset n = 13386059
Corpus subset n = 11270418
GES subset n = 2105129
Urgency subset n = 10512

Figure 2: Data flow and description of sources and target preprocessed subsets.

7.2 Descriptive analysis

A general view of the dataset is going to be described in the next lines. Specific descriptions per subsets are going to be reported.

7.2.1 Corpus Subset

The corpus that was used to train word embeddings consisted in 11,270,418 diagnostic suspicions, the mean number of words per diagnostic suspicion is 4.98 (5.12 SD). The corpus is composed of 56,079,828 word-tokens where the vocabulary length is 252,513 different words.

7.2.2 GES Subset

GES subset is comprised of 2,105,129 referrals spanning from 2005 to 2018, the mean age in the subset is 45.2 (26.0 SD) years. The ratio between GES and no-GES referrals is 1:4.6. Age and date of entry to the waiting list is not normally distributed and their distributions are reported in Figures 3 and 4.

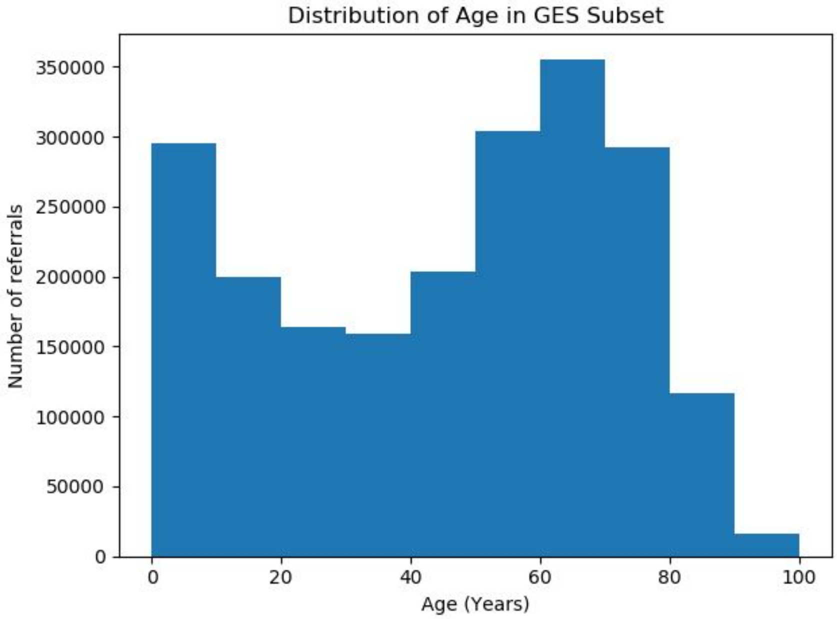


Figure 3: Histogram of the distribution of age of the referred patient in the GES subset.

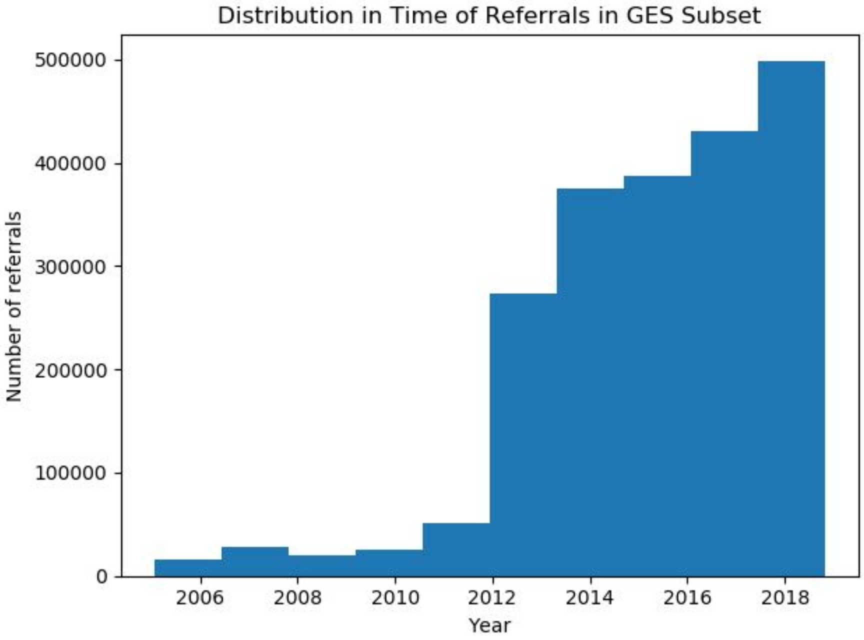


Figure 4: Histogram of distribution of the number of referrals as a function of the date where the referral was generated in the GES subset.

7.2.3 Urgency Subset

Urgency subset is comprised of 10,512 referrals spanning from 2011 to 2018. The ratio between Urgency and no-Urgency referrals is 1:7.1. The date of entry distribution is reported in Figure 5.

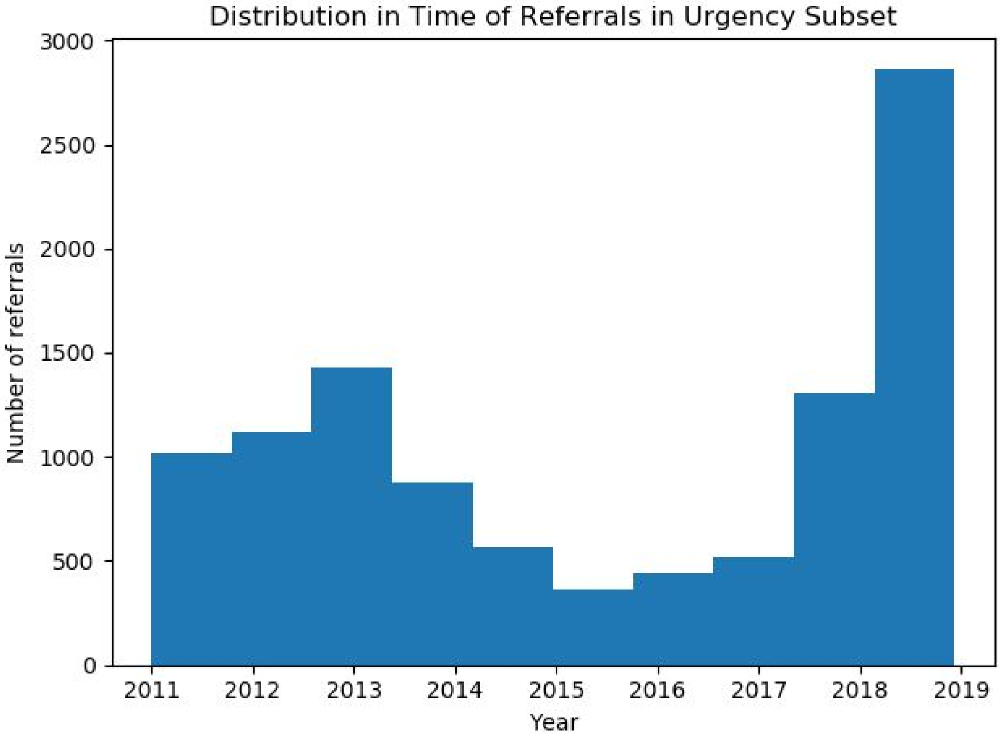


Figure 5: Histogram of distribution of the number of referrals as a function of the date where the referral was generated in urgency subset.

7.3 Word embeddings

Word embeddings were computed using corpus subset, the computed word embedding consists of 57,112 dense vectors of 300 dimensions. Qualitative results are reported in Table 1, where we show the most similar words to certain medical words to verify the embedding is returning semantically close words.

Word	**Most similar words**
diente	dte pieza pza
faringitis	rinofaringitis faringoamigdalitis adenoiditis
paracetamol	tramadol naproxeno ketoprofeno

Table 1: Most similar words for given words corresponding to the groups body part, disease, and medication.

A two-dimensional t-stochastic neighbour embedding (t-SNE) was projected from the word embedding to visualize relations between words (Figure 5.a), words surrounding the word *diente* are displayed in Figure 5.b.

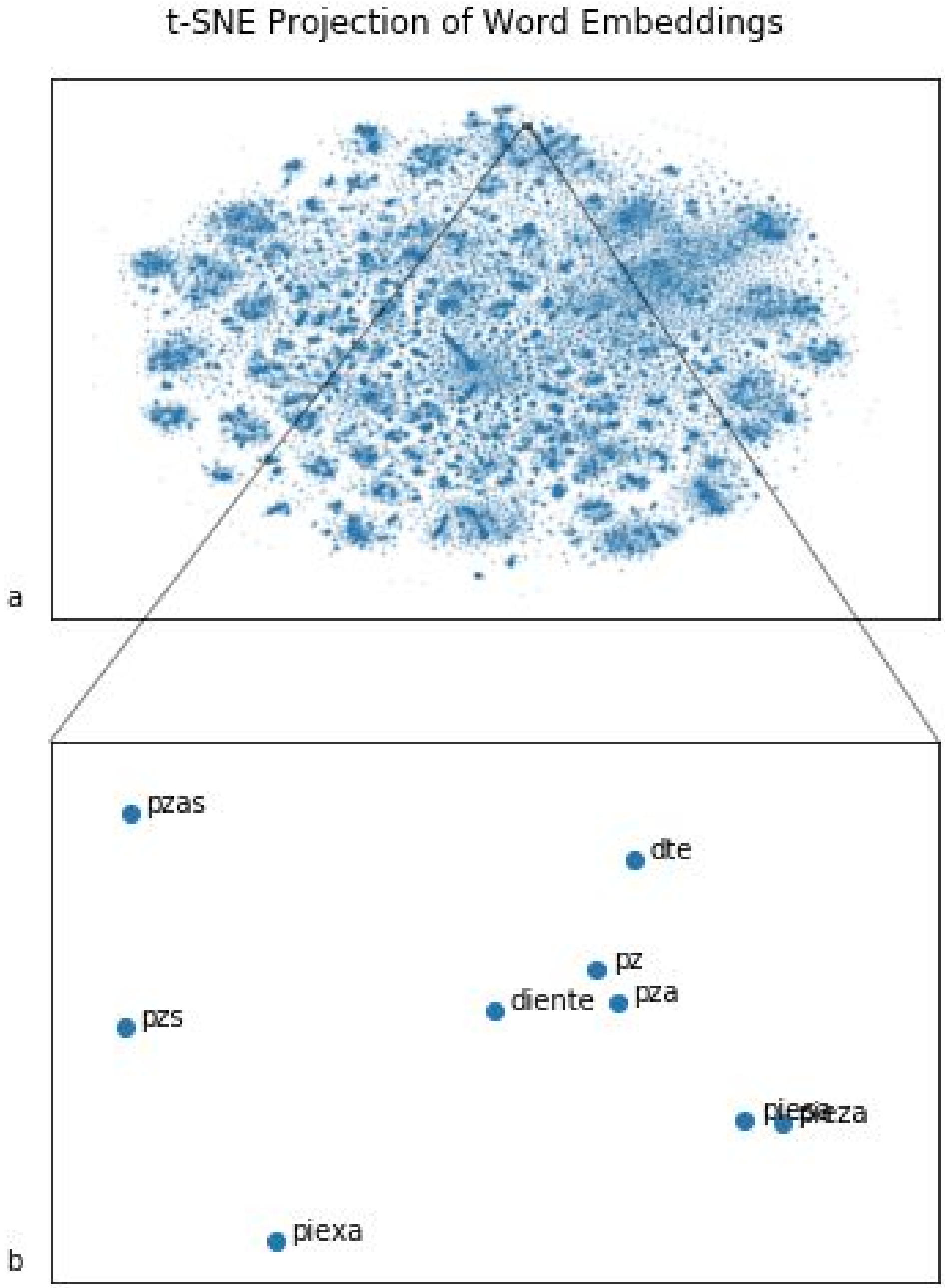


Figure 5: t-SNE projection of word embeddings (a) showing a region where similar words to *diente* are located (b)

7.4 Initial grid-search

A grid search was performed using different hyperparameters for each of the algorithms, the results for the GES and Urgency task are displayed on the next sections.

7.4.1 GES Subset

The best hyperparameters for the GES task for each algorithm are summarized in Tables 2 - 5.

Logistic Regression	
Hyperparameter	**Best Value**
Penalty	l2
C	10000

Table: 2 Best hyperparameters for Logistic Regression in the GES task.

Multilayer Perceptron	
Hyperparameter	**Best Value**
Solver	sgd
Learning rate	adaptive
Hidden layer sizes	(50, 100, 50)
Alpha	0.0001
Activation	relu

Table: 3 Best hyperparameters for Multilayer Perceptron in the GES task.

Random Forest	
Hyperparameter	**Best Value**
Number of estimators	1600
Minimum samples split	5
Max features	auto

Max depth	100
Bootstrap	true

Table: 4 Best hyperparameters for Random Forest in the GES task.

Support Vector Machine	
Hyperparameter	**Best Value**
Kernel	RBF
Gamma	1
C	10

Table: 5 Best hyperparameters for Support Vector Machine in the GES task.

7.4.2 Urgency Subset

The best hyperparameters for the Urgency task for each algorithm are summarized in Tables 6 - 9.

Logistic Regression	
Hyperparameter	**Best Value**
Penalty	l2
C	100

Table: 6 Best hyperparameters for Logistic Regression in Urgency task.

Multilayer Perceptron	
Hyperparameter	**Best Value**
Solver	adam
Learning rate	constant
Hidden layer sizes	(50, 100, 50)
Alpha	0.05

Activation	relu

Table: 7 Best hyperparameters for Multilayer Perceptron in Urgency task.

Random Forest	
Hyperparameter	**Best Value**
Number of estimators	1000
Minimum samples split	10
Max features	auto
Max depth	70
Bootstrap	true

Table: 8 Best hyperparameters for Random Forest in Urgency task.

Support Vector Machine	
Hyperparameter	**Best Value**
Kernel	RBF
Gamma	1
C	10

Table: 9 Best hyperparameters for Support Vector Machine in Urgency task.

7.5 Best model selection

Using the hyperparameters found in the last section, the best model is going to be selected based on the best AUCROC.

7.5.1 GES Subset

The best performing algorithm was Random Forest, achieving a normally distributed mean AUCROC of 0.961 [0.960, 0.962 CI95%], the results of the remaining algorithms are summarized in Figure 6.

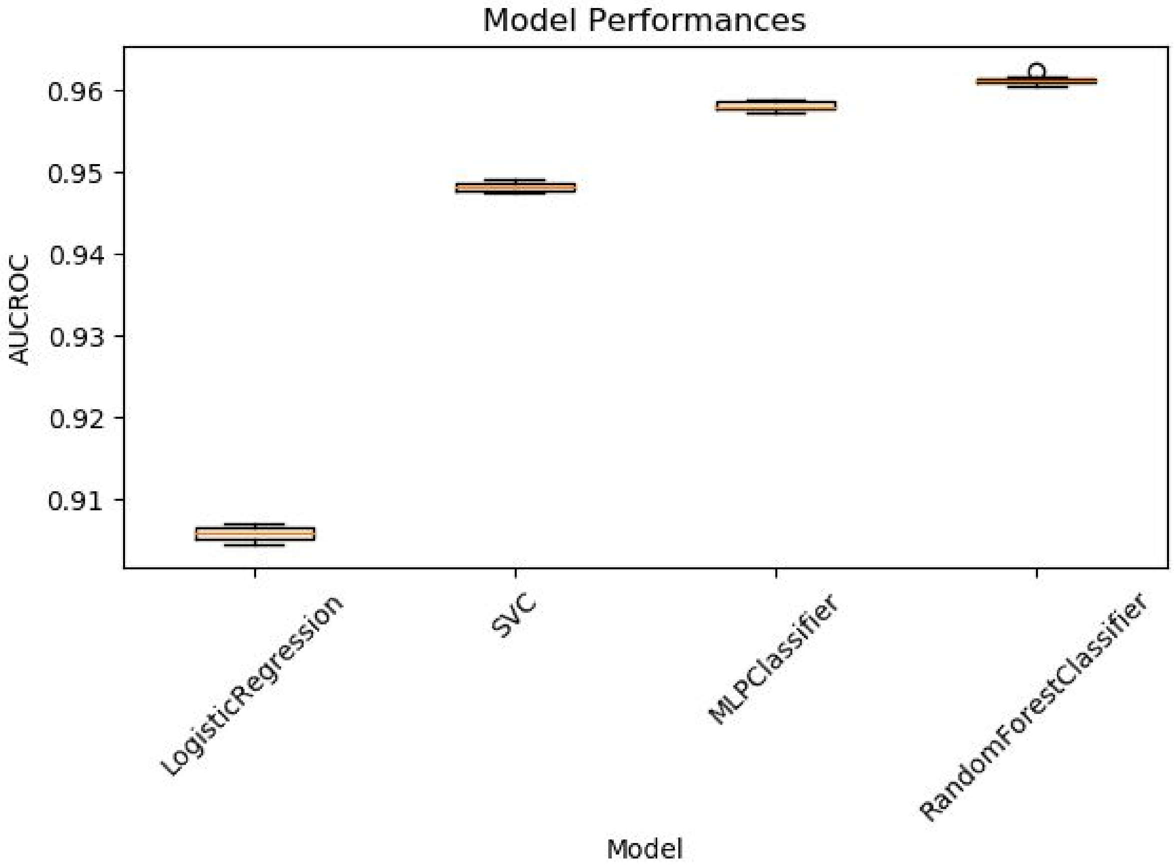


Figure 6: Boxplot of model performances after 10-fold cross-validation in the GES task. Statistical differences in the mean AUCROC were found between all model combination pairs.

7.5.2 Urgency Subset

The best performing algorithm was Random Forest, achieving a not normally distributed mean AUCROC of 0.999 (SD 0.0008), the results of the remaining algorithms are summarized in Figure 7.

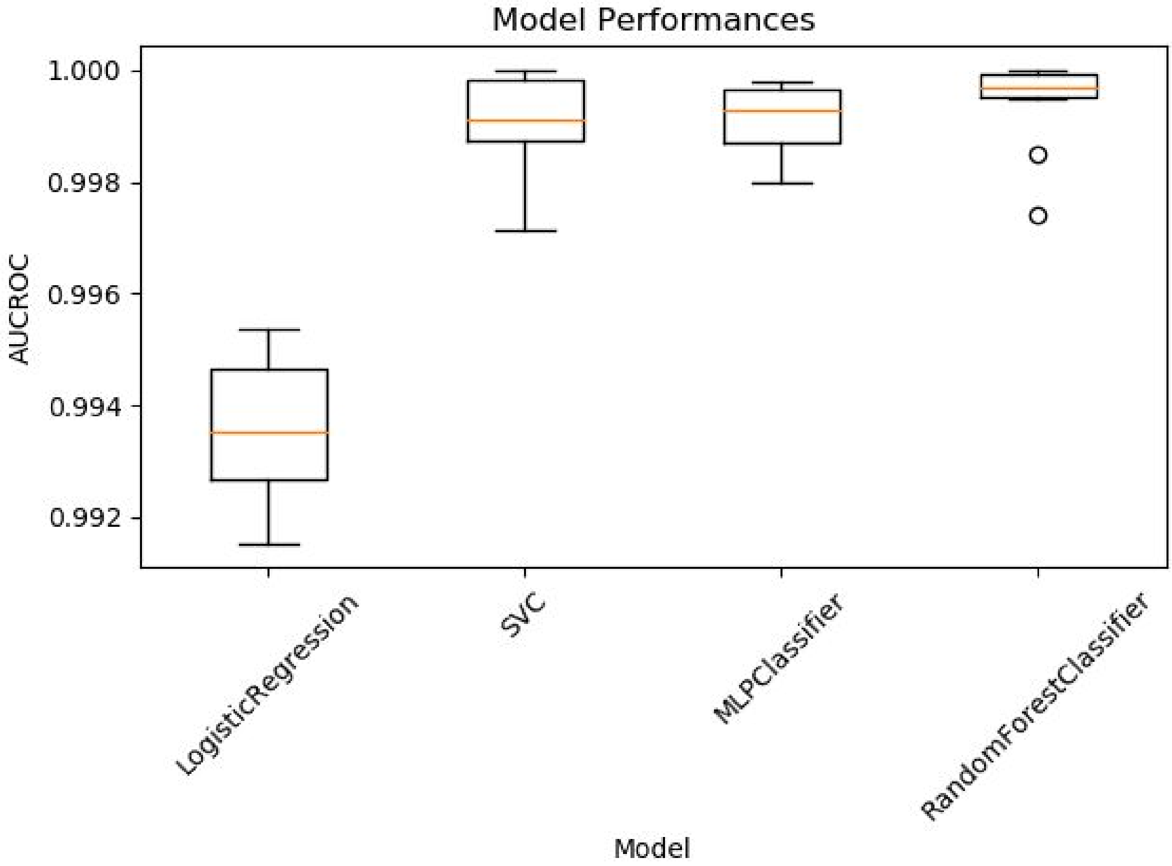


Figure 7: Boxplot of model performances after 10-fold cross-validation in Urgency task. Statistical differences on the mean AUCROC were found only between Random Forest - Logistic Regression, Logistic Regression - Multilayer Perceptron, Logistic Regression - Support Vector Machine model combinations.

7.6 Best model performance

The performance of Random Forest was tested over the testing subset of each of the tasks, the results are displayed in the next section.

7.6.1 GES Subset

Metrics for each class are summarized in Table 10 and the ROC curve is displayed in Figure 8.

Class	Precision	Recall	F1-score	Support
no-GES	0.98	0.91	0.94	173011
GES	0.67	0.90	0.77	37502
Weighted Average	0.92	0.90	0.91	210513

Table 10: Best model performance metrics for the GES task.

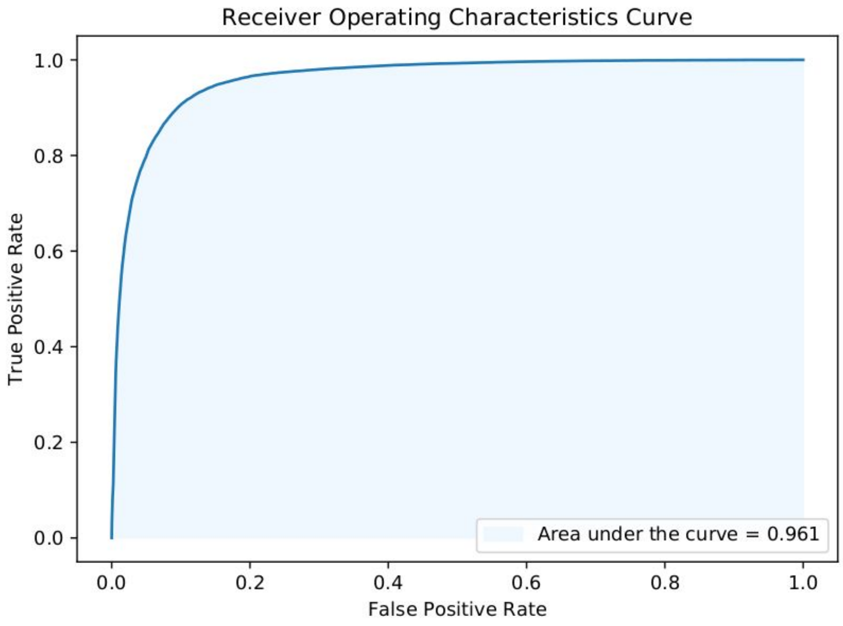


Figure 8: ROC curve for the GES task.

7.6.2 Urgency Subset

Metrics for each class are summarized in Table 11 and the ROC curve is displayed in Figure 9.

Class	Precision	Recall	F1-score	Support
no-Urgency	0.97	0.99	0.98	3679
Urgency	0.96	0.76	0.85	526
Weighted Average	0.97	0.88	0.92	4205

Table 11: Best model performance metrics for Urgency task.

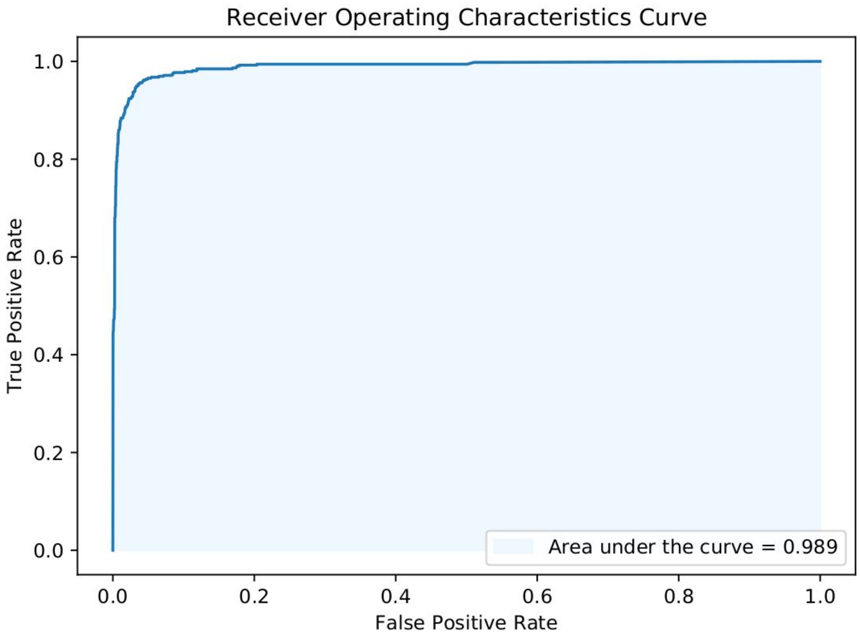


Figure 9: ROC curve for Urgency task.

7.7 Human - Machine Comparison and Performance over Gold Standard

The subset used for human-classification contained 941 diagnostic suspicions and this subset was labelled by 3 experts. In 829 diagnoses there were no discrepancies between the experts in the GES task and 798 in the Urgency task, these results are summarized in the Venn diagrams shown on Figures 10 and 11 for the GES and Urgency tasks, respectively.

Venn diagrams are figures that show all possible logical relations between sets. For example, in Figure 10, we have 3 sets, one for each expert and in each logical intersection, we show the number of agreements that the experts had: Between Human 1 and Human 2 were 33 agreements, Between Human 2 and Human 3 were 34 agreements, between Human 3 and Human 1 were 45 agreements and so on.

The experts achieved a 0.80 Fleiss-Kappa inter-expert agreement performance for the GES task and 0.64 for Urgency task, which both are a substantial agreement.

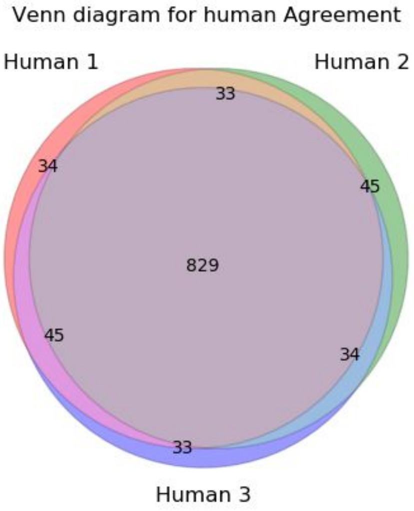


Figure 10: Venn diagram for GES human classification agreement.

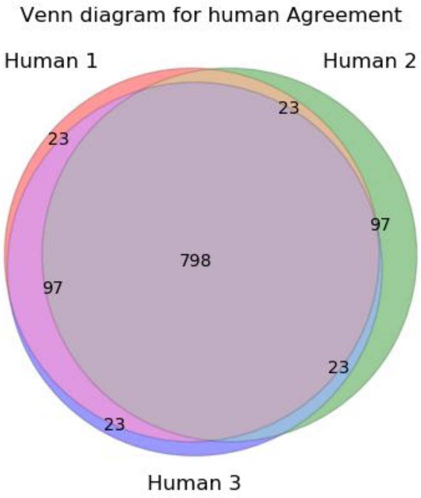


Figure 11: Venn diagram for Urgency human classification agreement.

7.7.1 GES Task

Performance metrics for each expert after gold standard construction are described in Table 12.

Expert	Class	Weighted Average			Support
		Precision	Recall	F1-Score	
1	no-GES	0.95	0.99	0.97	681
	GES	0.98	0.88	0.93	260
	Weighted	0.96	0.96	0.96	941
2	no-GES	0.99	0.93	0.96	681
	GES	0.85	0.97	0.91	260
	Weighted	0.95	0.94	0.94	941
3	no-GES	0.97	0.96	0.97	681
	GES	0.90	0.93	0.92	260
	Weighted	0.95	0.95	0.95	941
Average		0.95	0.95	0.95	2823

Table 12: Expert performance over ground truth

Machine performance over the gold standard subset is described in Table 13 and the ROC curve is displayed in Figure 12.

Class	Precision	Recall	F1-score	Support
no-GES	0.85	0.98	0.91	681
GES	0.92	0.55	0.69	260
Weighted Average	0.87	0.86	0.85	941

Table 13: Machine performance over ground truth

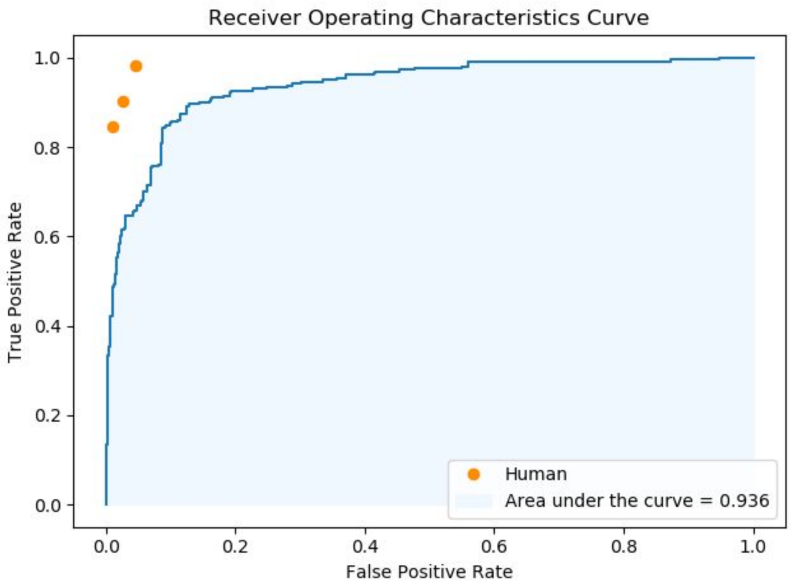


Figure 12: Machine ROC curve over ground truth and human performance

7.7.2 Urgency Task

Performance metrics for each expert after gold standard construction are described in Table 14.

Expert	Class	Weighted Average			Support
		Precision	Recall	F1-Score	
1	no-Urgency	0.99	0.96	0.97	822
	Urgency	0.78	0.90	0.84	119
	Weighted	0.96	0.96	0.96	941
2	no-Urgency	0.98	0.90	0.94	822
	Urgency	0.56	0.87	0.68	119
	Weighted	0.93	0.90	0.91	941
3	no-Urgency	0.99	0.96	0.98	822
	Urgency	0.77	0.95	0.85	119
	Weighted	0.96	0.96	0.96	941
Average		0.95	0.94	0.94	2823

Table 14: Expert performance over ground truth

Machine performance over the gold standard subset is described in Table 15 and the ROC curve is displayed in Figure 13.

Class	Precision	Recall	F1-score	Support
no-Urgency	0.92	0.91	0.94	822
Urgency	0.55	0.74	0.63	119
Average	0.90	0.89	0.90	941

Table 15: Machine performance over ground truth

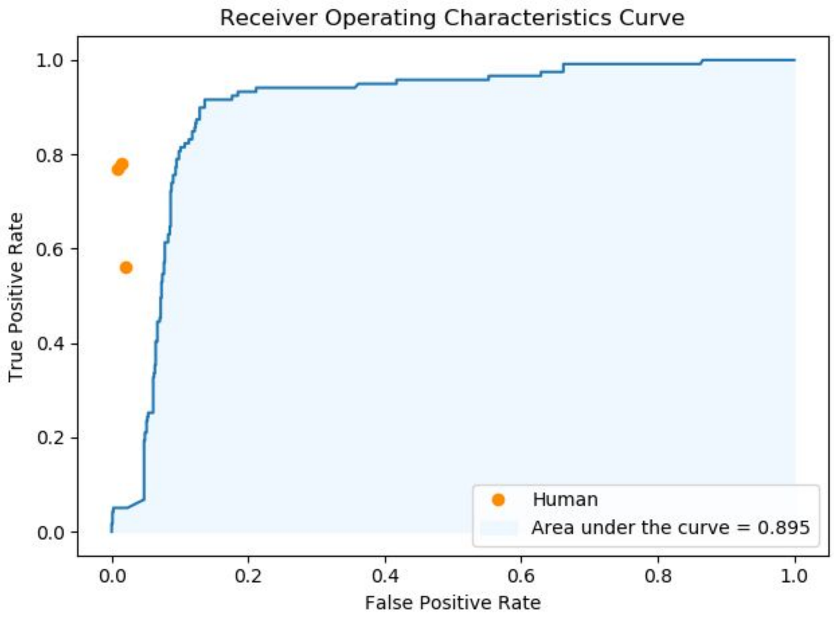


Figure 13: Machine ROC curve over ground truth and human performance

7.8 Deployment

To describe the deployment of the system in the hospital we divided it into (1) a backend classification engine and (2) a frontend classification interface. An overview of the semi-automatic pipeline of referral classification is summarized in Figure 14 and the results of the usage are described in the next section.

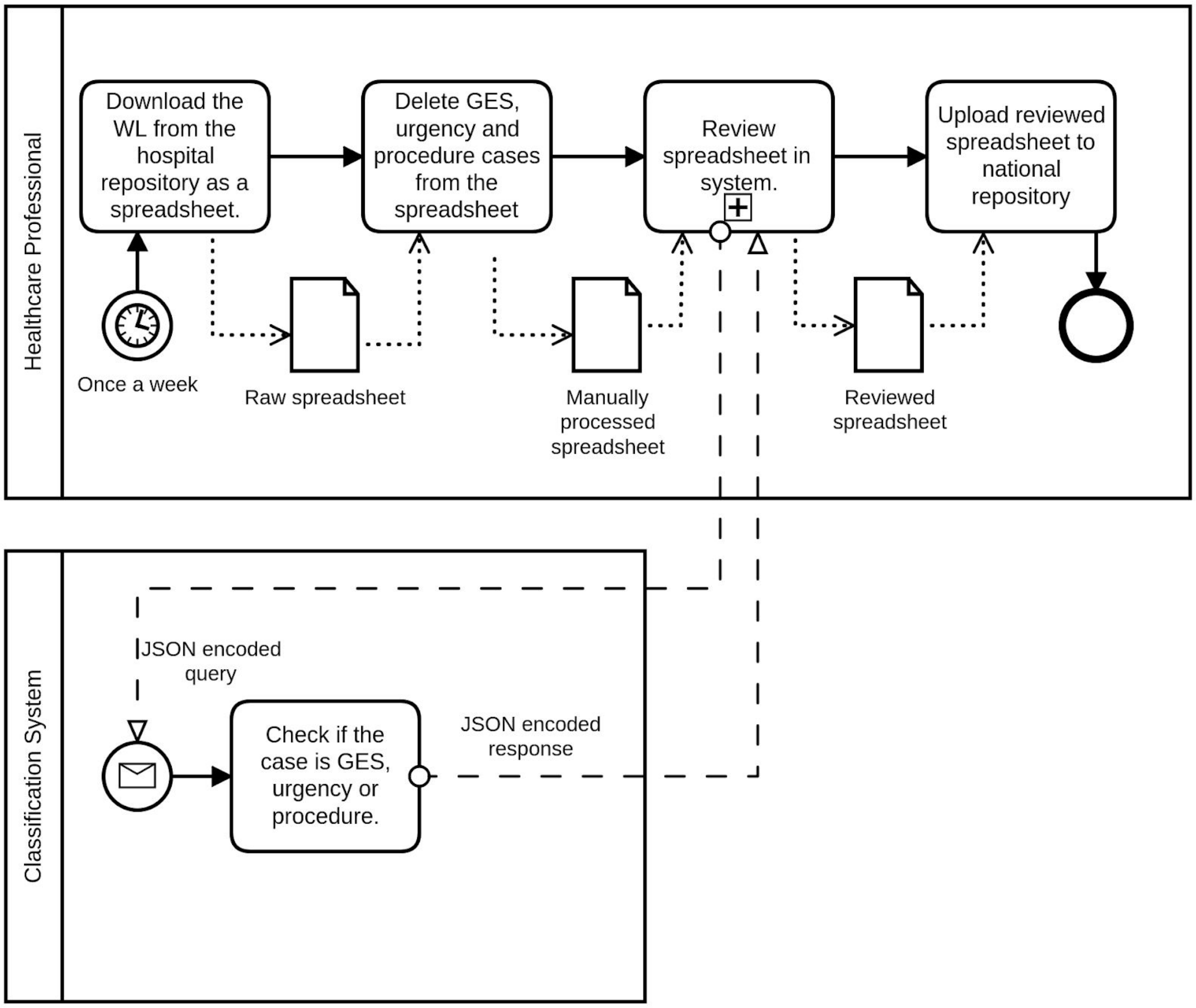

Healthcare Professional
Download the WL from the hospital repository as a spreadsheet.
Delete GES, urgency and procedure cases from the spreadsheet
Review spreadsheet in system.
Upload reviewed spreadsheet to national repository
Once a week
Raw spreadsheet
Manually processed spreadsheet
Reviewed spreadsheet
Classification System
JSON encoded query
Check if the case is GES, urgency or procedure.
JSON encoded response

Figure 14: Overview of the classification platform.

7.8.1 Classification engine

The best classification models were deployed in a server as a web service using Python as backend, this web service receives queries in JavaScript Object Notation (JSON) format containing the diagnostic suspicion, age of the patient and the billing code associated with the referral, the query is then parsed and passed through the model to retrieve the classes associated with the referral, the response is compiled as a JSON message and sent back to the sender. An example of the query and response messages are displayed in Figure 15.

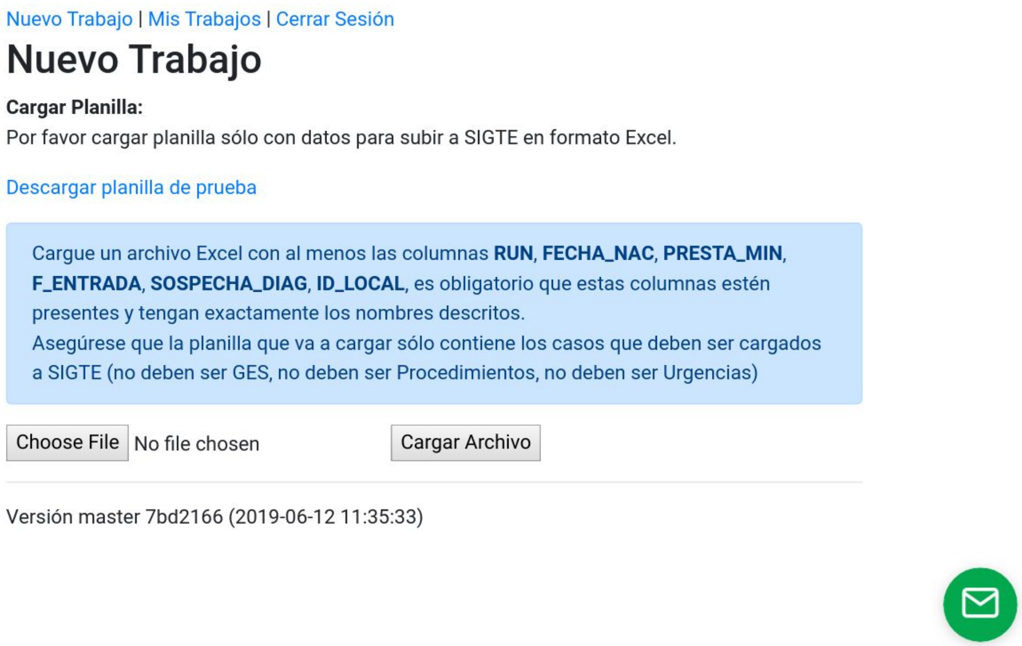
Query message	**Response message**

{
“diagnostic”: "cáncer cervocouterino",
“age”: 58,
“presta_min”: “12-05-26”
}
	{
	"ges":true,
	"proc":false,
	"urg":false
	}

Figure 15: Description of a query message regarding a referral in which the diagnostic suspicion is *cáncer cervocouterino* (intentional misspelling), the patient is 58 years old and the billing code is 12-05-26. The response message describes that the referral is classified only as a GES case.

7.8.2 Classification interface

A web-based classification interface was developed in PHP, JavaScript, CSS and HTML to receive the spreadsheets in Microsoft Excel format, parse the spreadsheet and classify each of the referrals using the classification engine. Next, the proposed classifications are displayed to the healthcare professional to verify the discrepancies. After the discrepancies are solved, a corrected spreadsheet is available to download. A graphical description of the procedure is shown in Figures 16 - 19.



Nuevo Trabajo | Mis Trabajos | Cerrar Sesión

Nuevo Trabajo

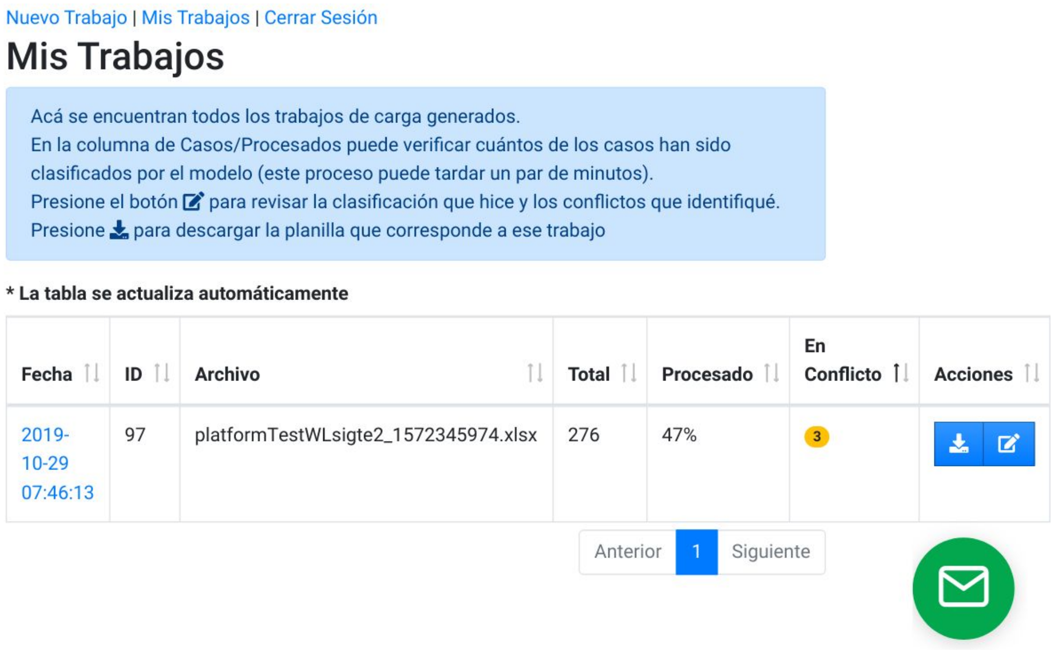
Cargar Planilla:
Por favor cargar planilla sólo con datos para subir a SIGTE en formato Excel.

Descargar planilla de prueba

Cargue un archivo Excel con al menos las columnas **RUN**, **FECHA_NAC**, **PRESTA_MIN**, **F_ENTRADA**, **SOSPECHA_DIAG**, **ID_LOCAL**, es obligatorio que estas columnas estén presentes y tengan exactamente los nombres descritos.
Asegúrese que la planilla que va a cargar sólo contiene los casos que deben ser cargados a SIGTE (no deben ser GES, no deben ser Procedimientos, no deben ser Urgencias)

Versión master 7bd2166 (2019-06-12 11:35:33)

Figure 16: Webpage to upload the manually processed spreadsheet in Microsoft Excel Format.



Nuevo Trabajo | Mis Trabajos | Cerrar Sesión

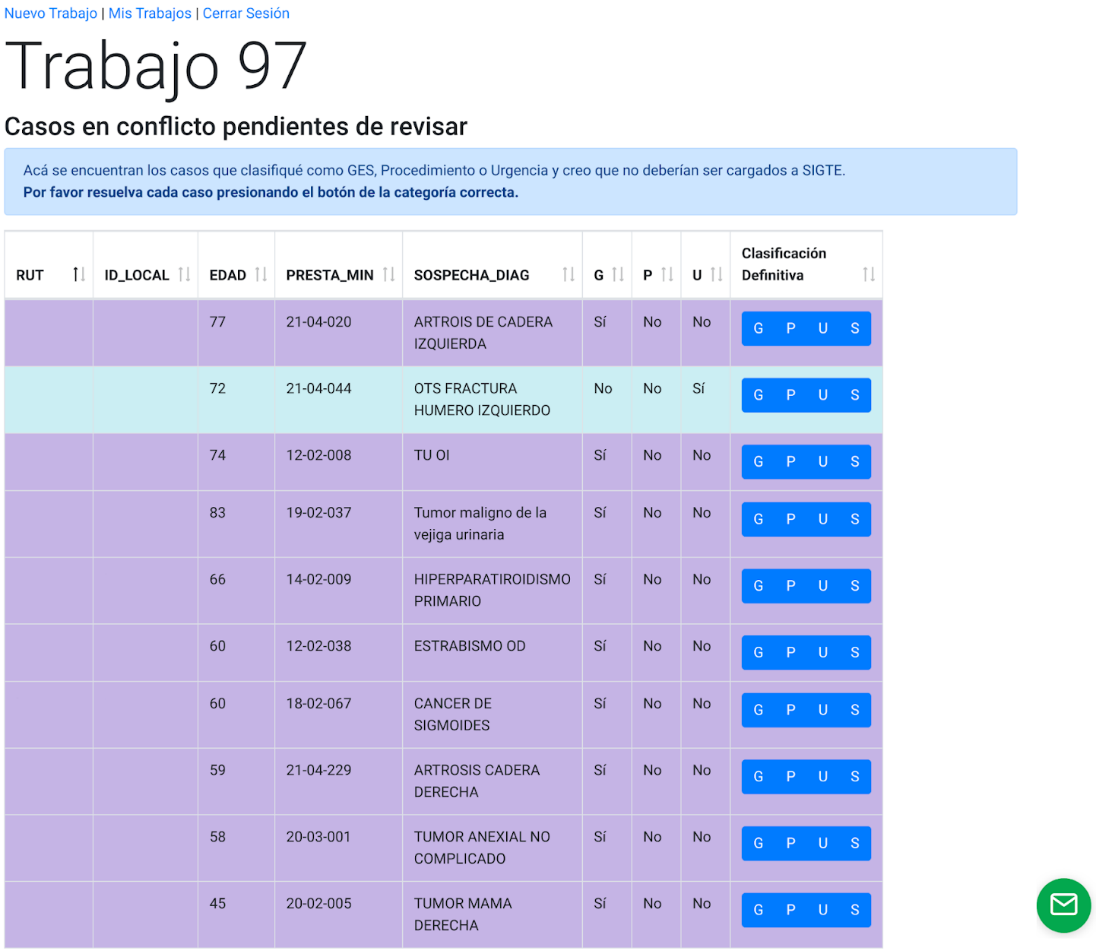
Mis Trabajos

Acá se encuentran todos los trabajos de carga generados.
En la columna de Casos/Procesados puede verificar cuántos de los casos han sido clasificados por el modelo (este proceso puede tardar un par de minutos).
Presione el botón ✎ para revisar la clasificación que hice y los conflictos que identifiqué.
Presione ⬇ para descargar la planilla que corresponde a ese trabajo

*** La tabla se actualiza automáticamente**

Fecha	ID	Archivo	Total	Procesado	En Conflicto	Acciones
2019-10-29 07:46:13	97	platformTestWLsigte2_1572345974.xlsx	276	47%	3	⬇ ✎

Figure 17: Webpage showing the current spreadsheet being processed by the backend.



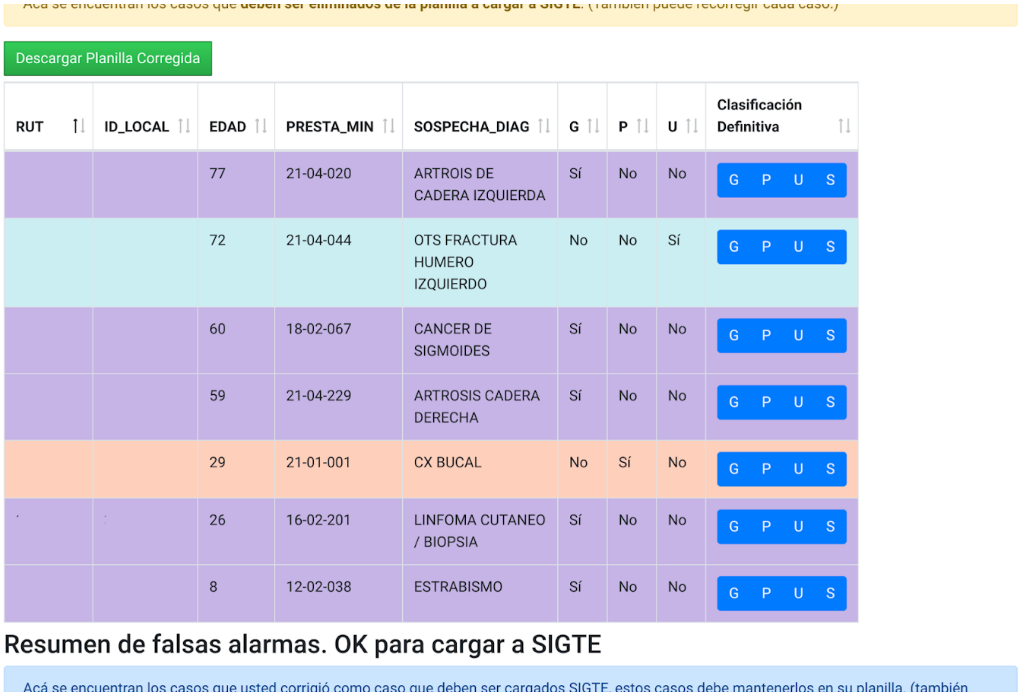
Trabajo 97

Casos en conflicto pendientes de revisar

Acá se encuentran los casos que clasifiqué como GES, Procedimiento o Urgencia y creo que no deberían ser cargados a SIGTE.
Por favor resuelva cada caso presionando el botón de la categoría correcta.

RUT	ID_LOCAL	EDAD	PRESTA_MIN	SOSPECHA_DIAG	G	P	U	Clasificación Definitiva
		77	21-04-020	ARTROIS DE CADERA IZQUIERDA	Sí	No	No	G P U S
		72	21-04-044	OTS FRACTURA HUMERO IZQUIERDO	No	No	Sí	G P U S
		74	12-02-008	TU OI	Sí	No	No	G P U S
		83	19-02-037	Tumor maligno de la vejiga urinaria	Sí	No	No	G P U S
		66	14-02-009	HIPERPARATIROIDISMO PRIMARIO	Sí	No	No	G P U S
		60	12-02-038	ESTRABISMO OD	Sí	No	No	G P U S
		60	18-02-067	CANCER DE SIGMOIDES	Sí	No	No	G P U S
		59	21-04-229	ARTROSIS CADERA DERECHA	Sí	No	No	G P U S
		58	20-03-001	TUMOR ANEXIAL NO COMPLICADO	Sí	No	No	G P U S
		45	20-02-005	TUMOR MAMA DERECHA	Sí	No	No	G P U S

Figure 18: Conflicts are shown to the user to manually solve each one pressing the desired correct referral class.


Acá se encuentran los casos que **deben ser eliminados de la planilla a cargar a SIGTE**. (También puede recorregir cada caso.)

RUT	ID_LOCAL	EDAD	PRESTA_MIN	SOSPECHA_DIAG	G	P	U	Clasificación Definitiva
		77	21-04-020	ARTROIS DE CADERA IZQUIERDA	Sí	No	No	G P U S
		72	21-04-044	OTS FRACTURA HUMERO IZQUIERDO	No	No	Sí	G P U S
		60	18-02-067	CANCER DE SIGMOIDES	Sí	No	No	G P U S
		59	21-04-229	ARTROSIS CADERA DERECHA	Sí	No	No	G P U S
		29	21-01-001	CX BUCAL	No	Sí	No	G P U S
		26	16-02-201	LINFOMA CUTANEO / BIOPSIA	Sí	No	No	G P U S
		8	12-02-038	ESTRABISMO	Sí	No	No	G P U S

Resumen de falsas alarmas. OK para cargar a SIGTE

Acá se encuentran los casos que usted corrigió como caso que deben ser cargados SIGTE, estos casos debe mantenerlos en su planilla. (también

Figure 19: Classifications reviewed by the use where the machine was right are displayed on a different table. A button to download the corrected spreadsheet is available.

7.8.3 Implementation results

Until September 2019, the platform has been used for 30 weeks and 4,472 referrals have been analyzed. Human-machine discrepancies were found in 129 cases, wherein 87 cases the machine was right.

8 Discussion

Currently, available classifiers rely only on nomenclature codifications made by general practitioners, but there is a high risk of bias on relying on this data because the general practitioners are not expert coders, so there is a high proportion of miscoded diagnostics [34]. Coded diagnostics are only a minimum part of the whole volume of the referrals, a low number of healthcare institutions in the healthcare service are using ontologies to code their diagnoses, therefore, the data retrieved from these institutions are in the form of free-text. Taking into account the argument previously stated, approaches that use nomenclatures are not completely suitable for this purpose, on the other hand, these rule-based or dictionary-based approaches are highly explainable and effective for inference tasks but not for prediction.

Machine learning approaches offer to the community an evidence validated non-deterministic way to classify free-text which adapts to the non-consistent writing of health professionals. In contrast with human classification, our method outperforms the speed of classification.

With the usage of machine learning along with neural word embedding we were able to extract semantic information from the data, and not only syntax information as in the case of rule-based methods and we did not need experts to engineer features from our text, which lowers the cost and time for developing text classification systems in healthcare.

We treated the general task into 2 subtasks, GES classification and urgency classification, because both classes are not mutually exclusive, so a referral can be GES and urgency at the same time. We also had 2 separate datasets for each of the subtasks and the features for each model are different, for GES classification we use diagnostic and age and for urgency classification we use only the referral diagnostic.

The metrics achieved by our methods outperform some previous works on detecting findings for apendicitis [35], detecting critical results [36] and detecting acute diseases [37,38], where the F1-Scores are around 0.85 and we achieved F1-Scores larger than 0.90.

The discussion of this work is going to be divided by the different types of results we retrieved in the last sections.

8.1 Data

Labelled data comes from the healthcare service which performs the largest number of medical consultations in the country [39], for that reason is that we have such a large number of referrals. When a patient goes to a medical consultation is often referred to other healthcare professionals for care continuity, normally from primary to secondary care, but the remaining combinations are also possible.

Unlabeled data was requested through transparency law, a government-driven solution to empower the country transparency [32]. Through this solution, every person has the right to request data directly from the institution, which has 20 days to answer. Institutions could deny the request based on their specific arguments, commonly lack of time. Nevertheless, the applicant can appeal to negative answers. In this work, we had a 79% positive answer to our requests, which is lower than the overall positive answer rate of 89% in the transparency law [40]. This government initiative is a relatively useful approach to empower open data and encourage data science research in Chile.

Age in GES dataset is not normally distributed and this bimodal distribution could be explained because of the increased necessity of care of children and elderly people. GES initiative also specifies some ages for a large number of health problems, these age-specific health problems also follows the distribution of referrals, benefiting children and elderly people [41].

The change from paper-based patient documentation to electronic health records in the health service is reflected in the high slope of the distribution of referrals between 2012 and 2013. The shift to electronic records is a reflection of the digital transformation the world is having and in Latin America, as the adoption of electronic health records solves health management problems and high demands in healthcare institutions [42]. The availability of interoperable electronic health records empowers the implementation of artificial intelligence solutions applied to medicine because of the easier connections for data retrieval [43].

To explore our text data we would have to analyze the corpus with classic NLP tools such as a frequency analysis of words in the different classes and extract the most informative words with the use of Pointwise Mutual Information Technique. For a visual representation of our corpus, word clouds could have been used to represent the frequencies of the words present in our corpus and communicate in a better way the composition of our corpus. An initial implementation of this method is described in the Appendix.

8.2 Word Embeddings

Word embeddings were calculated from the corpus subset to retrieve dense vectors that extract some semantic information and linguistic relationships from the corpus. The dense vectors were used as inputs for the prediction model to contextualize the text data from the referral. This approach has been used previously in the biomedical field [44,45] and the results outperform methods that do not use vector representations.

On the bidimensional projection of the embedding space, we were able to distinguish word clusters that correspond to certain medical concepts. As an example, our case study was the term tooth (*diente* in Spanish) wherein the embedding space appears closely related to its synonyms, namely abbreviations, misspellings, and linguistic synonyms. With this experiment, we exploited the capacity of word embedding to extract similarities between words. Word embeddings are one of the most used methods in text-similarity tasks [46]. We analyzed medical terms from different categories and word embeddings accomplished the task to retrieve semantically related terms.

Word embeddings have been used for text classification in a large number of models, one of the most relevant tasks in text classification is sentiment analysis [47–49], and word embeddings have achieved satisfactory results. In medicine, automatic text classification is not used extensively in Chile, and we aim that this successful case of use can encourage the implementation of similar systems. Traditional text classification in medicine are codification or referrals classification, but it can be extended to areas such as mining social media [50]. Moreover, we suggest the use of sentiment analysis to analyze patient response to medical interactions, such as the analysis of twitter posts about medicine to detect possible systematic problems in the Chilean healthcare network.

To increase the validity of our proposed system we would need to compare our method with different word embedding computing algorithms along with other text classification methods such as Naïve Bayes or by the usage of a more simple vector representation like a bag of words. Initial results for this method are reported in the Appendix.

8.3 Prediction Models

Our models achieved a Weighted Average F1-Score metric of 0.85 and 0.90, on GES and Urgency tasks, respectively, over a totally independent ground truth dataset labelled by 3 medical experts in the field of waiting lists.

In both tasks, humans achieved a substantial agreement, this is a reflection of the expertise of the professionals selected.

The results achieved by the machine did not outperform the results of the human experts, which obtained an average F1-Score of 0.95 and 0.94, GES and Urgency tasks, respectively. Overall negative classes had the best performance metrics both by machine and humans, this behaviour could be explained by the unbalance of the datasets and the problem itself because there are significantly more negative class cases.

The lowest metric in our models is the recall of the positive class in the GES task, and this could lead to not detect the entire number of GES cases. On the other hand, we could detect very precisely the GES cases, lowering the number of false positives. For the Urgency task, the lowest metric was the precision of the positive class, which could lead us to a larger amount of false positives, but we can detect most of the urgency cases.

Differences between the reported performance in developing and testing phases using the ground truth dataset can be explained by moderate overfitting in the training dataset. To lower the overfitting we could try to get more training data or use another balancing method for the training subset such as upsampling the minority class using Synthetic Minority Over-sampling Technique [51].

Even if we could not outperform human performance in this task, our method is significantly faster than human labelling. Our code for prediction over a collection of referrals took 10 minutes in a daily-usage laptop to predict over the ground truth while each human took around 120 minutes to label the ground truth dataset, more than 10 times more than the machine.

For being the first deployed automatic system to perform classification over referral data in Chile, the result we obtained is satisfactory, with a moderate room for further improvement.

For the enhancement of the performance of the model, we would need a multi-expert labelled training dataset because the results of our model could have been biased by the noisiness of the training dataset in our work. Using more advanced machine learning methods to solve this task could dramatically improve the performance, such as the usage of Recurrent Neural Networks with attention mechanisms [52], because they are the state of the art in predicting over sequence data.

8.4 Usability

The use of the platform by the healthcare professional in charge of uploading the cleaned waiting list was more frequent at the beginning of the project and it went down over time. The explanation for this phenomenon can be the non-existent active incentive to use the platform. To overcome this issue, frequent emails to the user can be sent to remember to use the platform or by the strategic adoption of the platform by the hospital administration, the latter is the one that probably would work better to increase the usage level.

We did not have complaints about the user experience, therefore a web-based solution can be a good option to deploy intelligent systems in healthcare institutions.

The platform was custom-tailored to be adapted specifically to the institution workflow, which maximized user experience, and initial user adoption to use the platform. To implement the frontend interface in a different healthcare institution, an initial current workflow analysis is mandatory.

A qualitative analysis was performed by master's in journalism about the adoption of intelligent systems in the Chilean healthcare network [53]. The results showed a positive reaction to the platform, emphasizing that humans can make mistakes and the system can be used to correct them. The professional stated that the platform is not making her work easier but is improving its quality as platform is a second filter before uploading the waiting list. About usability, the professional commented that the user interface was quite friendly and she had direct contact with the author if an issue was detected.

8.5 Clinical and Public Health Importance

Patient misclassification can lead to tremendous damage to a patient's fate. Waiting times in the non-GES waiting list are much longer than the GES waiting list and a misclassification could lead even to the patient's death [6,7]. An increased waiting time directly affects the quality of life and social and psychological health of the patient [54].

We helped to save 87 patients from the aforementioned fate by helping the healthcare professional to check a second time their classification. Human-machine discrepancies were caused because the healthcare professional classified the patient as a non-prioritized case and the patient was classified by the machine as a GES and/or Urgency case.

The usage of our intelligent system is helping the country to achieve the healthcare objectives of the decade [55] because (1) we are improving the quality of the health information systems by erasing human error in their records, (2) empowering cross-sector research by implementing computer science elements into the public healthcare sector, (3) improving the quality of sanitary technologies by applying cutting-edge methods to their information infrastructure and (4) improving patient satisfaction by decreasing misclassification and waiting times for GES patients.

Institutions receive fines for unsatisfactory management of GES patients [5] therefore by avoiding such fines using our system, that money could be invested in more critical necessities of the institution.

9 Conclusion

We were able to deploy a production-ready intelligent system to automatically classify referrals into GES and Urgency categories faster than human classification validating the work hypothesis. The performance of our platform is moderately comparable to human classification and outperforms a classical Bag of Words and Naïve Bayes classifier by 0.08 in F1-Score. The usage of neural methods for free-text systematization, along with machine learning classification algorithms, achieved the objective to classify automatically referrals with free-text narratives achieving better results than using non-neural vectorization.

Neural word embeddings trained over clinical text data were able to work as an input to machine learning algorithms to train models that classify diagnostics into GES and Urgency categories.

The usage of human labellers to construct a ground truth dataset achieved substantial results because of the high level of agreement achieved by the experts.

The platform was tailored to be adapted to the current data-cleaning workflow of the healthcare professional with quantitative and qualitative adequate results.

More work is needed in the creation of different methods for text classification to compare our metrics with other works. Baseline results with classical NLP methods such as bag-of-words are needed to really know if our method is better and there is also a necessity to compare our method to end-to-end deep learning approaches to reach the same goal.

Appendix

In order to compare the classifier proposed in this thesis , we made some additional baseline experiments using the Bag of Words model (BOW) and the Pointwise Mutual Information (PMI). Moreover, we made a descriptive analysis to detect the most informative words for each class. For simplicity we just focused on the GES classification, but Urgency classification is similar.

Baseline

We performed the task of classifying a text into GES class using a classic NLP text vectorization method named Bag of Words, which count the occurrence of each word in the vocabulary in each of the diagnostics. This vectorization method returns a very sparse matrix because each column corresponds to a word in the vocabulary. To train a classification model we used a Naïve Bayes classifier that uses the probability of each word to correspond to each class to make the prediction [11].

The performance metrics of this baseline are described in Table 16 and Figure 20.

Class	**Precision**	**Recall**	**F1-score**	**Support**
no-GES	0.97	0.81	0.88	173011
GES	0.50	0.87	0.64	37502
Weighted Average	0.88	0.82	0.84	210513

Table 16: Model performance of the Bag of Words and Naïve Bayes Method.

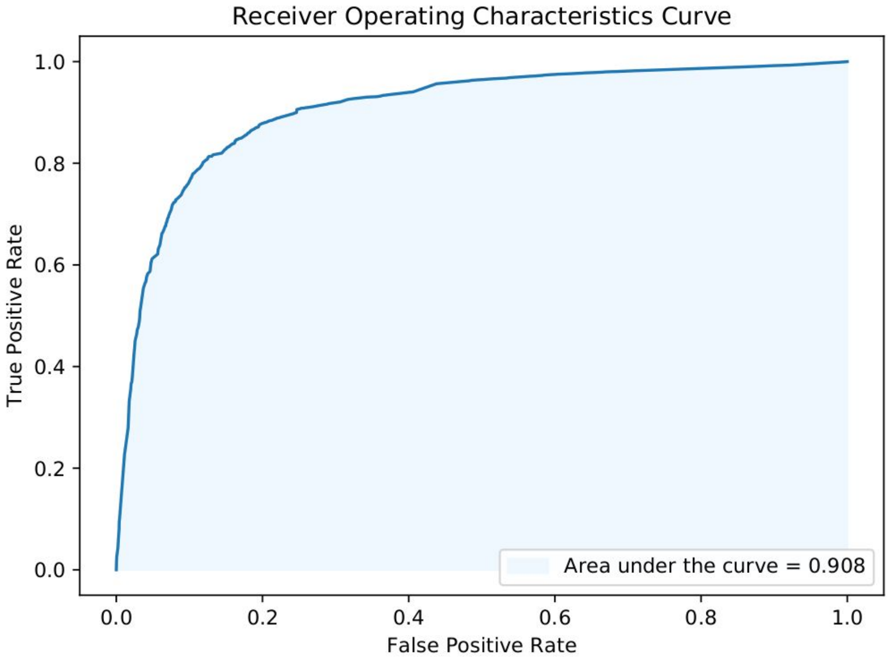


Figure 20: ROC curve of the Bag of Words and Naïve Bayes Method.

The classifier described in this thesis outperforme the baseline method in every metric, nevertheless the results shown by this simple approach are not bad. In addition, we can get the coefficients for each word in the vocabulary and these coefficients are directly related to the probability to belong to the GES class. The words with the largest coefficients are reported in Table 17.

Word	**Coefficient**	**Word**	**Coefficient**
refraccion	-3.40	diabetes	-4.65
especificado	-3.83	enfermedad	-4.67
trastorno	-3.88	hipoacusia	-4.72
ambos	-4.19	diabetica	-4.73
as	-4.21	retinopatia	-4.73
especificada	-4.34	mellitus	-4.77
vicio	-4.54	tumor	-4.88

catarata	-4.55	etapa	-4.91
cronica	-4.58	artrosis	-4.93
renal	-4.62	cadera	-4.96

Table 17: Words and their Naïve Bayes Coefficients

The words with the highest coefficients are closely related to GES problems. In fact, some GES problems that appeared in the previous table are:

- *Vicios de refracción en personas de 65 años y más.*
- *Tratamiento quirúrgico de cataratas.*
- *Diabetes Mellitus tipo I.*
- *Diabetes Mellitus tipo II.*
- *Hipoacusia bilateral en personas de 65 años y más que requieren uso de audífono.*
- *Retinopatía diabética.*
- *Enfermedad renal crónica etapa 4 y 5.*
- *Tratamiento médico en personas de 55 años y más con artrosis de cadera y/o rodilla, leve o moderada.*

Pointwise Mutual Information

To make some inference over the corpus and to explore which words are the most informative for each class, we computed the Pointwise Mutual Information (PMI) for each word in the corpus. PMI is a measure of how often two events, x and y, co-occur, compared with what we would expect if they were independent in our case x is the word itself and y is the label where the word appeared (GES or non-GES) [56]. In Table 18, we show the words with the highest PMI for each GES class.

GES		non-GES	
Word	**PMI**	**Word**	**PMI**
neurolues	1.72	arcos	0.19
saaf	1.72	seguim	0.19
coronari	1.72	iento	0.19
exudados	1.72	eo	0.19
cuadrantes	1.72	sicca	0.19

Table 18: Highest PMI values for both GES classes.

As we can see the words with the highest PMI are basically mistyped words, therefore it does not help us to extract valuable information. To check if in some way the results of the Naïve Bayes coefficients and PMI values are consistent we are going to compare some words extracted from Table 16 and check if they have a highest PMI value for the GES class.

Word	GES PMI	non-GES PMI
refraccion	1.26	-0.80
catarata	1.50	-1.42
diabetes	0.37	-0.10
hipoacusia	1.18	-0.67

Table 19: PMI values of selected words.

The values for the words extracted which had the highest Naïve Bayes coefficient are consistent with the PMI values, where each word is more important for the GES class than for the non-GES class.

10 References

1 Ministerio de Salud. Análisis de situación nuevas proyecciones de población ine, comparación con las proyecciones basadas en el censo 2002 y recomendaciones para su utilización. 2015.

2 Fondo Nacional de Salud. Población Inscrita en FONASA. 2013.

3 Organización de los establecimientos de salud en Chile. Wikipedia Encicl. Libre. 2019.https://es.wikipedia.org/w/index.php?title=Organizaci%C3%B3n_de_los_establecimientos_de_salud_en_Chile&oldid=116511556 (accessed 2 Jul 2019).

4 Ministerio de Salud de Chile. Ley 19.966. 2004.https://www.leychile.cl/Navegar?idNorma=229834 (accessed 2 Jul 2019).

5 Superintendencia de Salud de Chile. Sanciones a Prestadores. Supt. Salud Gob. Chile. 2019.http://www.supersalud.gob.cl/664/w3-propertyvalue-6249.html (accessed 2 Jul 2019).

6 Martinez DA, Zhang H, Bastias M, *et al.* Prolonged wait time is associated with increased mortality for Chilean waiting list patients with non-prioritized conditions. *BMC Public Health* 2019;**19**:1–11. doi:https://doi.org/10.1186/s12889-019-6526-6

7 Comisión Asesora del Ministerio de Salud. Estado de Situación Personas Fallecidas en Listas de Espera no GES y Garantías Retrasadas GES. 2017.

8 Superintendencia detecta fallas del Auge en tres cánceres "críticos." La Tercera. 2018.https://www.latercera.com/nacional/noticia/superintendencia-detecta-fallas-del-auge-tres-canceres-criticos/371560/ (accessed 12 Aug 2019).

9 Weber GM, Mandl KD, Kohane IS. Finding the Missing Link for Big Biomedical Data. *JAMA* Published Online First: 22 May 2014. doi:10.1001/jama.2014.4228

10 Kong H-J. Managing Unstructured Big Data in Healthcare System. *Healthc Inform Res* 2019;**25**:1–2. doi:10.4258/hir.2019.25.1.1

11 Dan Jurafsky, James H. Martin. *Speech and Language Processing*. 3rd ed. draft.https://web.stanford.edu/~jurafsky/slp3/

12 Pons E, Braun LMM, Hunink MGM, *et al.* Natural Language Processing in Radiology: A Systematic Review. *Radiology* 2016;**279**:329–43. doi:10.1148/radiol.16142770

13 Yim W, Yetisgen M, Harris WP, *et al.* Natural Language Processing in Oncology: A Review. *JAMA Oncol* 2016;**2**:797. doi:10.1001/jamaoncol.2016.0213

14 Topaz M, Lai K, Dowding D, *et al.* Automated identification of wound information in clinical notes of patients with heart diseases: Developing and validating a natural language processing application. *Int J Nurs Stud* 2016;**64**:25–31. doi:10.1016/j.ijnurstu.2016.09.013

15 Kreimeyer K, Foster M, Pandey A, *et al.* Natural language processing systems for capturing and standardizing unstructured clinical information: A systematic review. *J Biomed Inform* 2017;**73**:14–29. doi:10.1016/j.jbi.2017.07.012

16 Dalianis H. *Clinical Text Mining*. Cham: : Springer International Publishing 2018. doi:10.1007/978-3-319-78503-5

17 Urbanowicz RJ, Moore JH. Learning Classifier Systems: A Complete Introduction, Review, and Roadmap. *J Artif Evol Appl* 2009;**2009**:1–25. doi:10.1155/2009/736398

18 Mikolov T, Chen K, Corrado G, *et al.* Efficient Estimation of Word Representations in Vector Space. *ArXiv13013781 Cs* Published Online First: 16 January

2013.http://arxiv.org/abs/1301.3781 (accessed 10 Jul 2019).

19 Mikolov T, Sutskever I, Chen K, *et al.* Distributed Representations of Words and Phrases and their Compositionality.

20 Bojanowski P, Grave E, Joulin A, *et al.* Enriching Word Vectors with Subword Information. *ArXiv160704606 Cs* Published Online First: 15 July 2016.http://arxiv.org/abs/1607.04606 (accessed 12 Jul 2019).

21 Pennington J, Socher R, Manning C. Glove: Global Vectors for Word Representation. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: : Association for Computational Linguistics 2014. 1532–43. doi:10.3115/v1/D14-1162

22 Allen C, Hospedales T. Analogies Explained: Towards Understanding Word Embeddings. *ArXiv190109813 Cs Stat* Published Online First: 11 May 2019.http://arxiv.org/abs/1901.09813 (accessed 3 Dec 2019).

23 Bengio Y, Courville A, Vincent P. Representation Learning: A Review and New Perspectives. *ArXiv12065538 Cs* Published Online First: 24 June 2012.http://arxiv.org/abs/1206.5538 (accessed 17 Oct 2019).

24 Pouransari H, Ghili S. Deep learning for sentiment analysis of movie reviews.

25 Jin P, Zhang Y, Chen X, *et al.* Bag-of-Embeddings for Text Classification. *Proc Twenty-Fifth Int Jt Conf Artif Intell*

26 Rudkowsky E, Haselmayer M, Wastian M, *et al.* More than Bags of Words: Sentiment Analysis with Word Embeddings. *Commun Methods Meas* 2018;**12**:140–57. doi:10.1080/19312458.2018.1455817

27 Rajkomar A, Dean J, Kohane I. Machine Learning in Medicine. *N Engl J Med* 2019;**380**:1347–58. doi:10.1056/NEJMra1814259

28 James G, Witten D, Hastie T, *et al. An Introduction to Statistical Learning: with Applications in R*. Springer Science & Business Media 2013.

29 Leo Breiman. Random Forests. *Mach Learn* 2001;**45**:5–32.

30 Urgent care: Definition. https://publichealth.gwu.edu/departments/healthpolicy/CHPR/nnhs4/GSA/Subheads/gsa114.html (accessed 11 Mar 2020).

31 Wang Y, Liu S, Afzal N, *et al.* A comparison of word embeddings for the biomedical natural language processing. *J Biomed Inform* 2018;**87**:12–20. doi:10.1016/j.jbi.2018.09.008

32 Ministerio Secretaría General de la Presidencia. Ley 20.285. 2008.https://www.leychile.cl/Navegar?idNorma=276363&idParte= (accessed 10 Sep 2019).

33 Powers D. Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation.

34 Horsky J, Drucker EA, Ramelson HZ. Accuracy and Completeness of Clinical Coding Using ICD-10 for Ambulatory Visits. *AMIA Annu Symp Proc* 2018;**2017**:912–20.

35 Rink B, Roberts K, Harabagiu S, *et al.* Extracting Actionable Findings of Appendicitis from Radiology Reports Using Natural Language Processing. *AMIA Jt Summits Transl Sci Proc*;**2013**:221.

36 Lakhani P, Langlotz CP. Automated Detection of Radiology Reports that Document

Non-routine Communication of Critical or Significant Results. *J Digit Imaging* 2010;**23**:647–57. doi:10.1007/s10278-009-9237-1

37 Fiszman M, Chapman WW, Aronsky D, *et al.* Automatic Detection of Acute Bacterial Pneumonia from Chest X-ray Reports. *J Am Med Inform Assoc JAMIA* 2000;**7**:593–604.

38 Solti I, Cooke CR, Xia F, *et al.* Automated Classification of Radiology Reports for Acute Lung Injury: Comparison of Keyword and Machine Learning Based Natural Language Processing Approaches. *Proc IEEE Int Conf Bioinforma Biomed* 2009;**2009**:314–9. doi:10.1109/BIBMW.2009.5332081

39 Ministerio de Salud de Chile. Reporte REM Consultas Médicas. 2018.http://webdeis.minsal.cl/rem2018/?serie=1&rem=25&seccion_id=239&tipo=3&tipoReload=3&regiones=0&regionesReload=0&servicios=-1&serviciosReload=-1&periodo=2018&mes_inicio=1&mes_final=12 (accessed 3 Mar 2020).

40 Consejo Para la Transparencia de Chile. Informe Mensual de Estadísticas. 2020.https://www.portaltransparencia.cl/PortalPdT/documents/10179/62801/202001+Informe+mensual+PdT.pdf_1581001483084/51c73313-e0ea-4f31-ad47-23440137f770

41 Ministerio de Salud de Chile. Patologías garantizadas GES. Supt. Salud Gob. Chile. http://www.supersalud.gob.cl/664/w3-propertyname-501.html (accessed 3 Mar 2020).

42 Rodrigues RJ, Risk A. eHealth in Latin America and the Caribbean: Development and Policy Issues. *J Med Internet Res* 2003;**5**:e4. doi:10.2196/jmir.5.1.e4

43 Gil E, Medinaceli Díaz K. Electronic Health Record in Bolivia and ICT: A Perspective for Latin America. *Int J Interact Multimed Artif Intell* 2017;**4**:96. doi:10.9781/ijimai.2017.4412

44 Liu S, Tang B, Chen Q, *et al.* Effects of Semantic Features on Machine Learning-Based Drug Name Recognition Systems: Word Embeddings vs. Manually Constructed Dictionaries. *Information* 2015;**6**:848–65. doi:10.3390/info6040848

45 Tang B, Cao H, Wang X, *et al.* Evaluating Word Representation Features in Biomedical Named Entity Recognition Tasks. *BioMed Res Int* 2014;**2014**:1–6. doi:10.1155/2014/240403

46 Zhu Y, Yan E, Wang F. Semantic relatedness and similarity of biomedical terms: examining the effects of recency, size, and section of biomedical publications on the performance of word2vec. *BMC Med Inform Decis Mak* 2017;**17**:95. doi:10.1186/s12911-017-0498-1

47 Ren Y, Wang R, Ji D. A topic-enhanced word embedding for Twitter sentiment classification. *Inf Sci* 2016;**369**:188–98. doi:10.1016/j.ins.2016.06.040

48 Tang D, Wei F, Yang N, *et al.* Learning Sentiment-Specific Word Embedding for Twitter Sentiment Classification. In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Baltimore, Maryland: : Association for Computational Linguistics 2014. 1555–65. doi:10.3115/v1/P14-1146

49 Yang X, Macdonald C, Ounis I. Using word embeddings in Twitter election classification. *Inf Retr J* 2018;**21**:183–207. doi:10.1007/s10791-017-9319-5

50 Horvitz E, Mulligan D. Data, privacy, and the greater good. *Science* 2015;**349**:253–5. doi:10.1126/science.aac4520

51 Chawla NV, Bowyer KW, Hall LO, *et al.* SMOTE: Synthetic Minority Over-sampling Technique. *J Artif Intell Res* 2002;**16**:321–57. doi:10.1613/jair.953

52 Vaswani A, Shazeer N, Parmar N, *et al.* Attention is all you need. In: *Advances in neural information processing systems*. 2017. 5998–6008.

53 Mardones C, Maureira C. *Un computador puede salvar tu vida. La Inteligencia Artificial al servicio de la salud pública*. 2020.

54 Oudhoff J, Timmermans D, Knol D, *et al.* Waiting for elective general surgery: impact on health related quality of life and psychosocial consequences. *BMC Public Health* 2007;**7**:164. doi:10.1186/1471-2458-7-164

55 Chile, Ministerio de Salud. *Estrategia nacional de salud para el cumplimiento de los objetivos sanitarios de la década 2011-2020*. Santiago, Chile: : MINSAL 2011.

56 Dan Jurafsky, James H. Martin. *Speech and Language Processing*. 3rd ed. draft. 2018. https://web.stanford.edu/~jurafsky/slp3/