Supporting the Classification of Patients in Public Hospitals in Chile by Designing, Deploying and Validating a System Based on Natural Language Processing

Fabián Villena
University of Chile: Universidad de Chile
Jorge Perez
Universidad de Chile Facultad de Ciencias Físicas y Matemáticas: Universidad de Chile Facultad de Ciencias Fisicas y Matematicas
René Lagos
South East Metropolitan Health Service: Servicio de Salud Metropolitano Sur Oriente
Jocelyn Dunstan (jdunstan@uchile.cl)
Universidad de Chile https://orcid.org/0000-0001-6726-7242

Research article

Keywords: Decision Support Systems, Waiting Lists, Natural Language Processing, Machine Learning, Neural Networks (Computer)

Posted Date: November 19th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-108491/v1

Abstract

Background

In Chile, a patient needing a specialty consultation or surgery has to first be referred by a general practitioner, then placed on a waiting list. The Explicit Health Guarantees (GES in Spanish) ensure, by law, the maximum time to solve an important set of health problems. Usually, a health professional manually verifies if each referral, written in natural language, corresponds or not to a GES-covered disease. An error in this classification is catastrophic for patients, as it puts them on a non-prioritized waiting list, characterized by prolonged waiting times.

Methods

To support the manual process, we developed and deployed a system that automatically classifies referrals as GES-covered or not using historical data. Our system is based on word embeddings specially trained for clinical text produced in Chile. We used a vector representation of the reason for referral and patient's age as features for training machine learning models using human-labeled historical data. We constructed a ground truth dataset combining classifications made by three healthcare experts, which was used to validate our results.

Results

The best performing model over ground truth reached an AUC score of 0.94. During seven months of continuous and voluntary use, the system has amended 87 patient misclassifications.

Conclusion

This system is a result of a collaboration between technical and clinical experts, and the design of the classifier was custom-tailored for a hospital's clinical workflow, which encouraged the voluntary use of the platform. Our solution can be easily expanded across other hospitals since the registry is uniform in Chile.

Background

The analysis of the clinical free-text is challenging due to the generally non-standardized use of abbreviations and acronyms, the presence of negation, speculation and temporal expressions, or the limited availability of training corpora due to privacy concerns, among others (1). The challenge is even steeper when working on languages other than English because of the limited availability of tools and training corpora (2). Even though Spanish is one of the most spoken languages in the world, the number of language resources is still deficient, especially beyond Spain and the United States (3).

In Chile, the Public Health Fund FONASA covers 74% of the population (4). In contrast to the private sector, where a patient can go directly to a specialist, patients in the public healthcare system need a

referral from a general practitioner in primary care, which puts them in the national registry for the waiting list (WL) in the area where they are seeking medical attention. As a way to address health inequalities and health deterioration due to prolonged waiting times, in 2006 the Chilean government implemented 2006 the Explicit Health Guarantees (known as GES for its acronym in Spanish). GES prioritizes 80 health conditions that are prioritized, and their maximum amount of time, starting from the initial referral time, that patients may have to wait for treatment. It also offers economic coverage (5).

Public hospitals receive incentives to efficiently deal with GES-covered referrals, leading to dramatic differences in waiting times and volume of WL when one compares GES to non-GES waiting lists (6). Prolonged waiting times in the non-GES WL have been studied using hierarchical multivariate survival models applied to nearly a million patients in this WL, finding a statistically significant association between waiting time and death (7). In fact, 22,459 patients died in 2016 while waiting for their first consultation with a specialist, and 2,358 died before the surgery. These numbers are high compared with the 993 deaths in the GES group during the same year (8).

In 2018, the Chilean Healthcare Administration estimated that around 10% of patients with GES diagnoses were not receiving the prioritized treatment. Moreover, that same year, it sanctioned 83 healthcare institutions for the incorrect handling of GES cases (9). All of these statements imply that proper classification of GES and non-GES referrals are crucial, not only for patients but also for hospitals.

To assess if GES covers a referral, a healthcare professional needs to check if the free-text reason for referral, stated by the general practitioner, corresponds to one of 80 specific health conditions. This assertion is not always easy as each health problem has a subset of different pathologies covered for a given age range. For example, assume that there is a referral with the diagnosis *colelitiasis* (cholelithiasis) for a 36-year-old patient. This diagnosis should be marked as GES since (1) there is a GES health condition called *Colecistectomía preventiva del cáncer de vesícula en personas de 35 a 49 años*, which covers some diseases of the gallbladder for the age range 35 to 49 years old, and (2) this condition specifies a pathology called *Cálculo de la vesícula biliar sin colecistitis* which, although not explicitly, matches the referred diagnosis in the given age of the patient. These difficulties, together with the absence of standardized ways of defining pathologies, the heavy use of abbreviations, and spelling mistakes, among other reasons, require a health professional dedicated to the GES / non-GES classification in most hospitals across Chile. This professional, typically a nurse, reviews the WL manually, and uploads separated GES and non-GES databases to the National Repository.

The work presented here reports the design and performance of an automatic classifier of referrals trained over Chilean clinical text. We achieved this goal by collecting a clinical corpus and computing neural word embeddings. These embeddings were used in a variety of machine learning models, which were deployed as a web service in one of the biggest hospitals in Chile. Our system achieved a ROC AUC of 0.94 and continually used for seven months. During this time, the system analyzed 4,472 referrals and helped to re-classify 87 cases. Since the WL must be uploaded to a National Repository with a uniform

format, this classifier has the potential for use in every hospital in Chile. The word embeddings as well as the ground truth dataset are shared with the research community.

Methods

Data

We considered two datasets, one specific GES/non-GES *dataset* and one *general dataset* of referral narratives. The *GES dataset* was obtained from a collaboration with the Digital Health Unit at the South East Metropolitan Health Service (SEMHS) in Santiago, which provides healthcare service to 8.3% of the Chilean population. The SEMHS provided de-identified historical data of both GES and non-GES cases between 2005 and 2018. This dataset contains 2,105,129 cases, from which 375,969 were tagged as GES referrals. The access to SEMHS data, as well as the possibility of piloting the classifier at the Hospital Sótero del Rio, was possible due to a data agreement signed between SEMHS and the institution of the authors.

We obtained the second dataset, that we call the *general dataset*, via Chile's Transparency Law (10). It is composed of non-GES referrals from 23 of the 29 health services in the country for years in the range 2008-2018. That resulted in nearly 11 million referrals. We used this dataset as training corpora for the word embeddings, detailed in the next section.

Unsupervised learning: Word Embeddings

A good choice to deal with unstructured narratives are models based on artificial neural networks, which have reached state-of-the-art in several tasks (11). One of the techniques is word embeddings, which map each word to a real vector in D dimensions, with D much smaller than the vocabulary size (12,13). The idea of a word embedding is so useful because it allows us to assign a dense vector to each word in the vocabulary, and within this smaller dimension space, perform operations on these vectors and also test the quality of the representation (14–16).

Word embeddings are obtained by training a single-layer neural network over vast text corpora. The task that fine-tunes the weights in the network can either predict a word from the context words in sentences of the corpus (*continuous bag of words* method) or predict the context words from a central word (*skip-gram* method). We refer the reader to the work by Mikolov *et al.*(12) for details.

In this work, we computed word embeddings using Word2Vec with the *skip-gram* method (12), with a vector dimension of 300, and all the remaining options as default. Before vectorizing, the text was lowercased and tokenized using the NLTK package (17). In addition, characters other than alphabetical and punctuation were deleted through regular expressions. Finally, we dropped 156,948 sentences because of duplicated data or blank attributes.

To obtain a single vector for each referral, we took the average of the vector assigned to each word in the referral, which is a standard practice in Natural Language Processing (18,19). Furthermore, we weighted

this average using Term Frequency-Inverse Document Frequency score (TF-IDF). This score is proportional to the frequency of appearance of a given word within the document (in this case a referral), but it offsets this value by the number of documents in which this word appears in the *corpus* (all the referrals) (16). As it is constructed, stop words (such as *the, a, in*) score low, and semantically richer words (*cancer, pain*) receive a high score. This process gave us a 300-dimension vector for each referral.

The choice of training corpora for the word embedding construction is not trivial. In particular, for the clinical domain, authors have explored combinations of general language, biomedical literature, and clinical corpora (20,21). For the clinical Spanish language there is a lack of language resources, with the few corpora coming predominantly from Spain (22–24). In the work presented here, we extrinsically tested the Spanish Billion Word Corpus (25), the Chilean Biomedical corpus (26), and the general dataset described earlier, with the latest showing the best classification performance (see Results section).

Supervised learning: GES Classifier

The vector representation of the reason for referral, as well as the patient's age (transformed using min-max scaler), were the inputs to train supervised machine learning models. We tested Support Vector Machines, Random Forest, Logistic Regression, and Multi-Layer Perceptron using the scikit-learn package in Python.

Datasets were split into training and testing, with the training subset balanced by downsampling the majority class. The best hyperparameters were selected via grid search and choosing those that maximized the area under the curve of the receiver operating characteristic (ROC AUC) using 3-fold cross-validation.

The optimized models were compared based on their ROC AUC using 10-fold cross-validation. We assessed the statistical significance of the difference in performance between the averages calculating the paired t-student test with a Bonferroni p-value correction.

Ground truth construction and validation

For the creation of a ground truth testing dataset, three experts labeled a random subset of 942 referrals as GES or non-GES. In the case of discrepancies, the first author decided on the label based on the information contained in the official documents of the Healthcare Superintendence.

This ground truth was used to assess the performance of the best model and to compare the level of agreement between humans.

Deployment

For the implementation of the classification models in the hospital we designed a program that receives referrals and responds with the corresponding predicted class. The backend was designed in Python using the Flask web framework. The service receives a JavaScript Object Notation (JSON) encoded

message containing the referral information (diagnostic suspicion and patient's age). The message is then parsed, and the information is used as input in the model. In this process, the text data is preprocessed and vectorized, and the patient's age scaled. The model's predicted result is then compiled into another JSON message containing the Boolean result as a value of the GES key.

The frontend side of the deployment was developed in PHP, JavaScript, CSS, and HTML. A tailored web-based portal received an Excel spreadsheet of referrals containing the features needed to decide between GES and non-GES classes. The portal composes, parses, sends and receives JSON messages to display the results of the predictions in a user-friendly way.

The application checks every entry by sending requests to the Web service, and then displays, in a customized way, the human-machine discrepancies. Our application allows the user to discard or accept the classifier's suggestions. After managing discrepancies, the user can download rectified spreadsheet files.

A critical aspect of our application is that when conflicts are found, the platform retrieves the human-corrected class of each referral so the data can be later used to retrain the model. A general overview of the deployed platform is shown in Figure 1.

Results

Word embeddings

Word embeddings can be tested in two ways: intrinsically and extrinsically (27). In the first case, two common tasks are the semantic analogy and the semantic similarity. For the extrinsic evaluation, on the other hand, we measure the performance on a downstream task that makes use of the embeddings (21,28). Since we were interested in choosing the best embedding for the GES/non-GES classification task, we selected between three different embeddings using this classification as extrinsic evaluation. We assessed the classification on the ground truth created by human experts.

The three embeddings were calculated using Word2vec with identical hyperparameters (12), but with different training corpora: the general dataset described previously (using non-GES referrals), the biomedical corpus (26), and the Spanish Billion Word Corpus (25). As shown in Table 1, the general dataset, constructed with non-GES referrals, showed the best performance in the classification task. This result is in agreement with the work by Chen *et al.* (21) for the English language.

Training corpus	Vocabulary size (tokens)	ROC AUC
General dataset	57,112	0.94
Biomedical literature	183,766	0.90
General Spanish language	1,000,653	0.90

Table 1: Extrinsic evaluation of Word2vec embeddings with identical hyperparameters, but different training corpora.

Development Performance

The embedding with the best performance was used to vectorize the diagnostic suspicion, which was then used, along with patient's age, as input in machine learning classifiers. Table 2 summarized the performance of each of the models. The statistical significance of the difference between the mean performance of each of the combinations of models was significant, with a p-value < 0.01.

Model	ROC AUC (SD)
Logistic Regression	0.91 (7.8 e-4)
Support Vector Machine	0.95 (5.4 e-4)
Random Forest	0.96 (5.2 e-4)
Multilayer Perceptron	0.95 (5.9 e-4)

Table 2: Performance of machine learning models.

Random Forest showed the best performance, reaching a ROC AUC of 0.96. Table 2 shows other metrics for this model for the GES, non-GES, and weighted by frequency classes for the testing dataset and the ground truth.

Testing dataset				
Class	**Precision**	**Recall**	**F1**	**Support**
GES	0.67	0.90	0.77	37,502
non-GES	0.98	0.91	0.94	173,011
Weighted Average	0.92	0.92	0.91	210,513
Ground truth dataset				

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 2: Performance of the Random Forest Classifier over the testing dataset and the ground truth constructed from human classifications.

Upon closer inspection of Table 2, one notes that the precision of the GES case in the training dataset is not very high (67%). Nevertheless, the recall is significantly better (90%). Possibly, the most critical metric in our case is the recall of the GES class, as we want to retrieve as many misclassified GES cases as possible. Having a high performance for the non-GES case (0.94 F1 in our model) is also essential if one wants to use our system as a support for the clinical decision for a set of manually classified referrals.

Human-machine comparison

In order to compare the performance of the best method with a consolidated ground truth, we asked three health professionals related to WL classification to label 941 diagnostic suspicions as GES or non-GES. In 829 diagnoses, there were no discrepancies between the experts.

The experts' agreement was further quantified using the Fleiss-Kappa coefficient, which is a statistical coefficient similar to Cohen's kappa but for more than two raters (29). The three experts achieved 0.80 in the Fleiss-Kappa coefficient, which is considered a substantial agreement.

As shown in Table 3, the individual performance of humans is excellent. These raters were chosen to participate in this validation from their experience in the GES / non-GES classification, which is evidenced in these metrics.

Expert	Weighted Average		
	Precision	Recall	F1-Score
1	0.96	0.96	0.96
2	0.95	0.94	0.94
3	0.95	0.95	0.95
Average	0.95	0.95	0.95

Table 3: Expert performance over ground truth.

Finally, the best machine learning classifier was tested on this ground truth dataset. Table 2 describes the results and Figure 2 displays the ROC curve.

Deployment

The Web application is in agreement with the workflow of the healthcare professional at Hospital Sótero del Rio in charge of checking and uploading the non-GES WL to the National Repository. Additionally, in the deployed application, a chat with the development team was embedded so healthcare professionals can communicate with one another if they have questions or comments regarding the platform. After seven months of consecutive work, the platform analyzed 4,472 referrals. Human-machine discrepancies were 129 cases, wherein 87 cases the machine was right.

Discussion

The classifier and Web application reported in this paper is a result of nearly two years of collaboration between the SEMHS and the University of Chile. We wanted to identify a problem with clinical relevance, with enough data to train models, and where supporting the decision making could make a clear difference.

An automatic referral system that detects GES cases in the non-GES WL is beneficial at least in the following aspects: (a) it facilitates the job of the health professional in charge of the WL classification by supporting his/her decisions; (b) patients increase their chances of receiving prioritized attention when corresponds, and (c) hospitals avoid fines due to misclassifications. In summary, it provides support for better decision making, improving safety for patients and hospitals.

Patient misclassification can be detrimental to a patient's outcome. Waiting times in the non-GES WL are much longer than the GES WL, and a misclassification could even lead to a patient's death [6,7]. Furthermore, an increased waiting time directly affects the quality of life and social and psychological health of patients [54].

To classify free-text referrals, we first created a vector representation for each reason for referral and then used this vector in machine learning models. We used word embeddings to vectorize free-text narratives, which are based on neural networks. For the work presented here, we did not use pre-defined embeddings for the Spanish language. We instead collected 11 million non-GES referrals for specialty consultations in the public healthcare system, obtaining vector representations tailored for this specific type of clinical narrative. The embeddings trained with this corpus showed the best performance in the classification task.

Word embeddings have been used previously in the biomedical and clinical fields. Examples of applications include the quantification of relatedness of biomedical terms (30,31), the identification of entities in clinical narratives (32), the use of embeddings to expand abbreviations (33), or the automatic codification of diseases (34), to mention some.

For the classification of referrals in GES and non-GES classes, we used a variety of machine learning models. They received as input the vectorization of the referral as well as the age of the patient. The use of machine learning in medicine has been slower than in other disciplines, but its extensive use is auspicious (35). We can roughly group machine learning applications in those using *classical* machine learning methods and those that use *deep* learning. A key aspect of choosing between them is the amount of training data and if interpretation of the models is a must (36). In our case, from the amount of data we had, classical machine learning methods were the most suitable option.

For our application, the method with the best performance was Random Forest, which has been widely used since its publication in 2001 (37). Its use in medicine has been natural as they can be explained as an arrangement of decision trees, a concept rooted in medicine. In terms of medical applications, we find

a variety of successful cases of the application of tree-based methods. Examples include the detection of hospital-acquired infections (38), the prediction of obesity rates from food sales (39), or detecting suicidal behavior in emergency consultations (40).

Our model achieved a weighted average F1-Score metric of 0.85, calculated over an independent ground truth dataset labeled by three medical experts in the field of waiting lists. In the classification task, humans achieved a substantial agreement, which reflects the expertise of the professionals selected.

Differences between the reported performance in testing and validating phases can be explained by moderate overfitting in the training dataset. To lower the overfitting, we can get more training data by using the platform in another hospital or using another balancing method for the training subset, such as upsampling the minority class using the Synthetic Minority Oversampling Technique (41).

In terms of time used by the machine vs. humans, the automatic classification takes 10 minutes in a daily-usage laptop to predict the class of 1,000 referrals. In contrast, each human took around 120 minutes to label the same referrals. Therefore, even if the automatic classifier does not outperform human experts, our method is significantly faster than human labeling. On top of that, the amount of highly qualified health professionals is heterogenous along the country.

In order to enhance the performance of the classifier, we hope to deploy the platform in other hospitals in or even better, automatically check the non-GES WL in the National Repository in the Ministry of Health. A second way to improve our work is to retrain the model by taking into account the mistakes of the machine when compared to the health professional in charge of WL as well as the comparison with the ground truth. Due to the large number of examples used in the training process, we could create synthetic examples of these mistakes in order to enforce the learning over these cases. Finally, if we manage to get significantly more data, we could use more advanced machine learning methods to solve this task, such as Recurrent Neural Networks with attention mechanisms (42), which are state of the art in predicting over free-text inputs.

Our solution does not replace human decision; instead, it provides a second opinion based on historical information. The voluntary and continuous use for seven months demonstrates its usefulness for the healthcare professional that used the platform. This person did not need considerable extra time or specialized training to use the classifier, which was a crucial factor in the success of this project. Detailed understanding of the WL reporting process and an agile development approach were crucial for deploying the application in the hospital procedures correctly. Besides, in this project, we verified what the literature states: successful machine learning projects in healthcare require both clinical, and technical participants with solutions that can be used following the clinical workflow (43).

Conclusions

We were able to deploy a production-ready system to automatically classify referrals into GES and no-GES in a public hospital in Chile. The performance of the platform was compared with a ground truth

made from the classification of three waiting list experts, and the automatic system is moderately worse than human classification, but more than ten times faster than the experts.

In order to use the information contained in the reason for referrals, we used neural word embeddings specifically trained over Chilean clinical text. These vectors were the input of machine learning algorithms that classify diagnoses into GES and non-GES categories, with Random Forest showing the best performance. The platform was tailored to be adapted to the current data-cleaning workflow of the healthcare professional, used continuously and voluntarily in the system for seven months.

The use of our intelligent system is helping Chile achieve the healthcare objectives of the decade (44) 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.

Abbreviations

GES: Explicit Health Guarantees (acronym in Spanish); WL: Waiting List; SEMHS: South East Metropolitan Health Service; TF-IDF: Term Frequency-Inverse Document Frequency; ROC AUC: area under the curve of the receiver operating characteristic; JSON: JavaScript Object Notation; SD: standard deviation.

Declarations

Ethics approval

The ethics committee at the South East Metropolitan Health Service has waived the need for approval.

Availability of data and materials

The word embeddings trained over 11 million free text diagnostics from all over Chile are publicly shared within this publication doi.org/10.5281/zenodo.3924799.

Competing interests

Nothing to declare.

Funding

JD acknowledges U-INICIA VID 2019 UI-004/19. JD & FV are supported by the Center for Mathematical Modeling ANID AFB 170001 and the Center for Medical Informatics and Telemedicine expense center 570111. This research was partially supported by the supercomputing infrastructure of the NLHPC (ECM-02).

Author Contribution

Study concept and design: FV, JD, RL. Acquisition of data: JD, RL. Deployment: FV. Compute models: FV, JP. Drafting of the manuscript: FV, JD. Supervision of the overall study: JD. All authors read and approved the final manuscript.

Acknowledgments

The authors are grateful to Nury González for using the platform and her feedback on the Web service, Ignacio Castro and Maricella Reyes for manual annotation of data. We also thank Juan Cristóbal Morales for overall support, and Ren Cerro and Pablo Báez for comments on this manuscript.

References

1. Dalianis H. Clinical Text Mining: Secondary Use of Electronic Patient Records. Springer; 2018.
2. Névéol A, Dalianis HK, Savova G, Zweigenbaum P. Clinical Natural Language Processing in Languages Other Than English: opportunities and challenges. J Biomed Semantics. 2018;9:12:1–13.
3. Moreno A, Torre D, Valverde A, Campillos L. Estudio sobre documentos reutilizables como recursos lingüísticos en el marco del desarrollo del Plan de Impulso de las Tecnologías del Lenguaje. Proces del Leng Nat. 2019;63:167–70.
4. Fondo Nacional de Salud. Población Inscrita en FONASA [Internet]. [citado 26 de marzo de 2020]. Disponible en: https://www.fonasa.cl/sites/fonasa/documentos
5. Ministerio de Salud de Chile. Ley GES 19.966. 2004.
6. Ministerio de Salud de Chile. Estado de situación personas fallecidas en listas de espera no GES y garantías retrasadas GES. Informe de la comisión médica asesora ministerial [Internet]. 2017. Disponible en: https://www.minsal.cl/wp-content/uploads/2018/01/Informe-Final-Comision-Asesora-LE-y-Garantias-Retrasadas-GES-17082017_.pdf
7. Martinez DA, Zhang H, Bastias M, Feijoo F, Hinson J, Martinez R, et al. Prolonged wait time is associated with increased mortality for Chilean waiting list patients with non-prioritized conditions. BMC Public Health. 2019;1–11.
8. Cuadrado C, Crispi F, Estay R, González F, Alvarado F, Cabrera N. Desde el Conflicto de las Listas de Espera, Hacia el Fortalecimiento de los Prestadores Públicos de Salud. Una Propuesta para Chile. Col Médico, Cuad médicos-sociales.
9. Sandoval G, Leiva L. Superintendencia detecta fallas del Auge en tres cánceres “críticos”. Tercera [Internet]. 22 de octubre de 2018; Disponible en: https://www.latercera.com/nacional/noticia/superintendencia-detecta-fallas-del-auge-tres-canceres-criticos/371560/
10. Portal de Transparencia Chile [Internet]. Disponible en: www.portaltransparencia.cl/PortalPdT/
11. NLP Progress [Internet]. Disponible en: https://nlpprogress.com/

12. Mikolov T, Sutskever I, Chen K, Corrado G, Dean J. Distributed Representations of Words and Phrases and their Compositionality. 2013;1–9. Disponible en: http://arxiv.org/abs/1310.4546
13. Pennington J, Socher R, Manning C. Glove: Global Vectors for Word Representation. En: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2015. p. 1532–43.
14. Mikolov T, Yih W, Zweig G. Linguistic Regularities in Continuous Space Word Representations. Hlt-Naacl. 2013;(June):746–51.
15. Levy O, Goldberg Y. Dependency-Based Word Embeddings. Proc 52nd Annu Meet Assoc Comput Linguist (Volume 2 Short Pap [Internet]. 2014;302–8. Disponible en: http://aclweb.org/anthology/P14-2050
16. Jurafsky D, Martin JH. Speech and language processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Third Edit. 2018.
17. Bird S, Klein E, Loper E. Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit. O'Reilly Media; 2009.
18. Pouransari H. Deep learning for sentiment analysis of movie reviews. CS224N [Internet]. 2014;1–8. Disponible en: http://web.stanford.edu/class/cs224d/reports/PouransariHadi.pdf
19. Afshar M, Phillips A, Karnik N, Mueller J, To D, Gonzalez R, et al. Natural language processing and machine learning to identify alcohol misuse from the electronic health record in trauma patients: development and internal validation. J Am Med Informatics Assoc. 2019;26(3):254–61.
20. Wang Y, Liu S, Afzal N, Rastegar-Mojarad M, Wang L, Shen F, et al. A comparison of word embeddings for the biomedical natural language processing. J Biomed Inform [Internet]. 2018;87(April):12–20. Disponible en: https://doi.org/10.1016/j.jbi.2018.09.008
21. Chen Z, He Z, Liu X, Bian J. Evaluating semantic relations in neural word embeddings with biomedical and general domain knowledge bases. BMC Med Inform Decis Mak [Internet]. 2018;18. Disponible en: http://dx.doi.org/10.1186/s12911-018-0630-x
22. Campillos-Llanos L. First Steps towards a Medical Lexicon for Spanish with Linguistic and Semantic Information. BioNLP 2019. 2019;152.
23. Soares F, Villegas M, Gonzalez-Agirre A, Krallinger M, Armengol-Estapé J. Medical Word Embeddings for Spanish: Development and Evaluation. Proc ofthe 2nd Clin Nat Lang Process Work [Internet]. 2019;124–33. Disponible en: http://doi.org/10.5281/zenodo.2542722
24. Krallinger M, Rabal O, Leitner F, Vazquez M, Salgado D, Lu Z, et al. The CHEMDNER corpus of chemicals and drugs and its annotation principles. J Cheminform. 2015;7(Suppl 1):1–17.
25. Cardellino C. Spanish Billion Word Corpus and Embeddings [Internet]. 2016. Disponible en: https://crscardellino.github.io/SBWCE/
26. Durán M, Villena F, Dunstan J. Corpus médico en español de revistas médicas de Chile [Internet]. 2019. Disponible en: http://corpusmedico.cimt.cl/%0A

27. Chiu B, Crichton G, Korhonen A, Pyysalo S. How to Train good Word Embeddings for Biomedical NLP. En: Proceedings of the 15th workshop on biomedical natural language processing. 2016. p. 166–74.
28. Wang Y, Liu S, Afzal N, Rastegar-Mojarad M, Wang L, Shen F, et al. A comparison of word embeddings for the biomedical natural language processing. J Biomed Inform. 2018;87:12–20.
29. Ide N, Pustejovsky J. Handbook of Linguistic Annotation. Springer; 2017.
30. 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(1):95.
31. Chen Z, He Z, Liu X, Bian J. Evaluating semantic relations in neural word embeddings with biomedical and general domain knowledge bases. BMC Med Inform Decis Mak [Internet]. 2018;18(Suppl 2). Disponible en: http://dx.doi.org/10.1186/s12911-018-0630-x
32. Habibi M, Weber L, Neves M, Wiegandt DL, Leser U. Deep learning with word embeddings improves biomedical named entity recognition. Bioinformatics. 2017;33(14):i37–48.
33. Henriksson A, Moen H, Skeppstedt M, Daudaravičius V, Duneld M. Synonym extraction and abbreviation expansion with ensembles of semantic spaces. J Biomed Semantics. 2014;5(1):6.
34. Kavuluru R, Rios A, Lu Y. An empirical evaluation of supervised learning approaches in assigning diagnosis codes to electronic medical records. Artif Intell Med. 2015;65(2):155–66.
35. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. N Engl J Med. 2019;380(14):1347–58.
36. Beam AL, Kohane IS. Big Data and Machine Learning in Health Care. JAMA [Internet]. 2018;02115. Disponible en: http://jama.jamanetwork.com/article.aspx?doi=10.1001/jama.2017.18391
37. Breiman L. Random forests. Mach Learn [Internet]. 2001;5–32. Disponible en: http://link.springer.com/article/10.1023/A:1010933404324
38. Ehrentraut C, Ekholm M, Tanushi H, Tiedemann J, Dalianis H. Detecting hospital-acquired infections: A document classification approach using support vector machines and gradient tree boosting. Health Informatics J [Internet]. 2018;24:24–42. Disponible en: http://jhi.sagepub.com/cgi/doi/10.1177/1460458216656471
39. Dunstan J, Aguirre M, Bastías M, Nau C, Glass TA, Tobar F. Predicting nationwide obesity from food sales using machine learning. Health Informatics J. 2020;26(1):652–63.
40. Metzger MH, Tvardik N, Gicquel Q, Bouvry C, Poulet E, Potinet-Pagliaroli V. Use of emergency department electronic medical records for automated epidemiological surveillance of suicide attempts: a French pilot study. Int J Methods Psychiatr Res. 2017;26(2).
41. Chawla N V, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res. 2002;16:321–57.
42. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. En: Advances in neural information processing systems. 2017. p. 5998–6008.
43. Cresswell KM, Bates DW, Wright A, Sheikh A. An Overview of Clinical Informatics [Internet]. Key Advances in Clinical Informatics: Transforming Health Care through Health Information Technology.

Elsevier Inc.; 2017. 3–11 p. Disponible en: http://dx.doi.org/10.1016/B978-0-12-809523-2.00001-7

44. Ministerio de Salud de Chile. Estrategia Nacional de Salud para el cumplimiento de los Objetivos Sanitarios de la Década 2010-2020. Santiago de Chile: Ministerio de Salud. 2011.

Figures

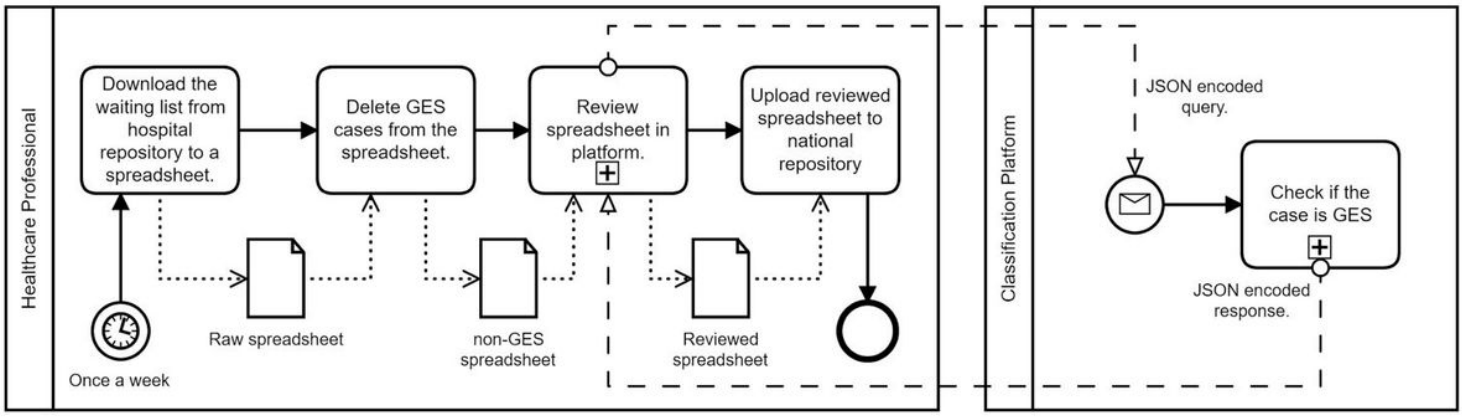

Figure 1

Diagram of the classification process. The non-GES WL is checked in the classification platform to make sure there are no GES cases in it, which should be prioritized by law.

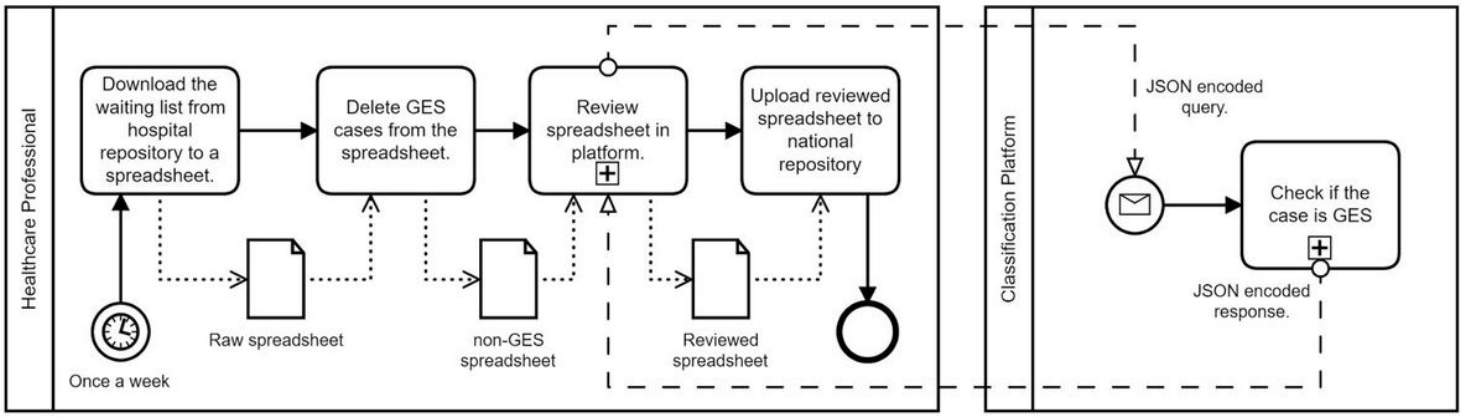

Figure 1

Diagram of the classification process. The non-GES WL is checked in the classification platform to make sure there are no GES cases in it, which should be prioritized by law.

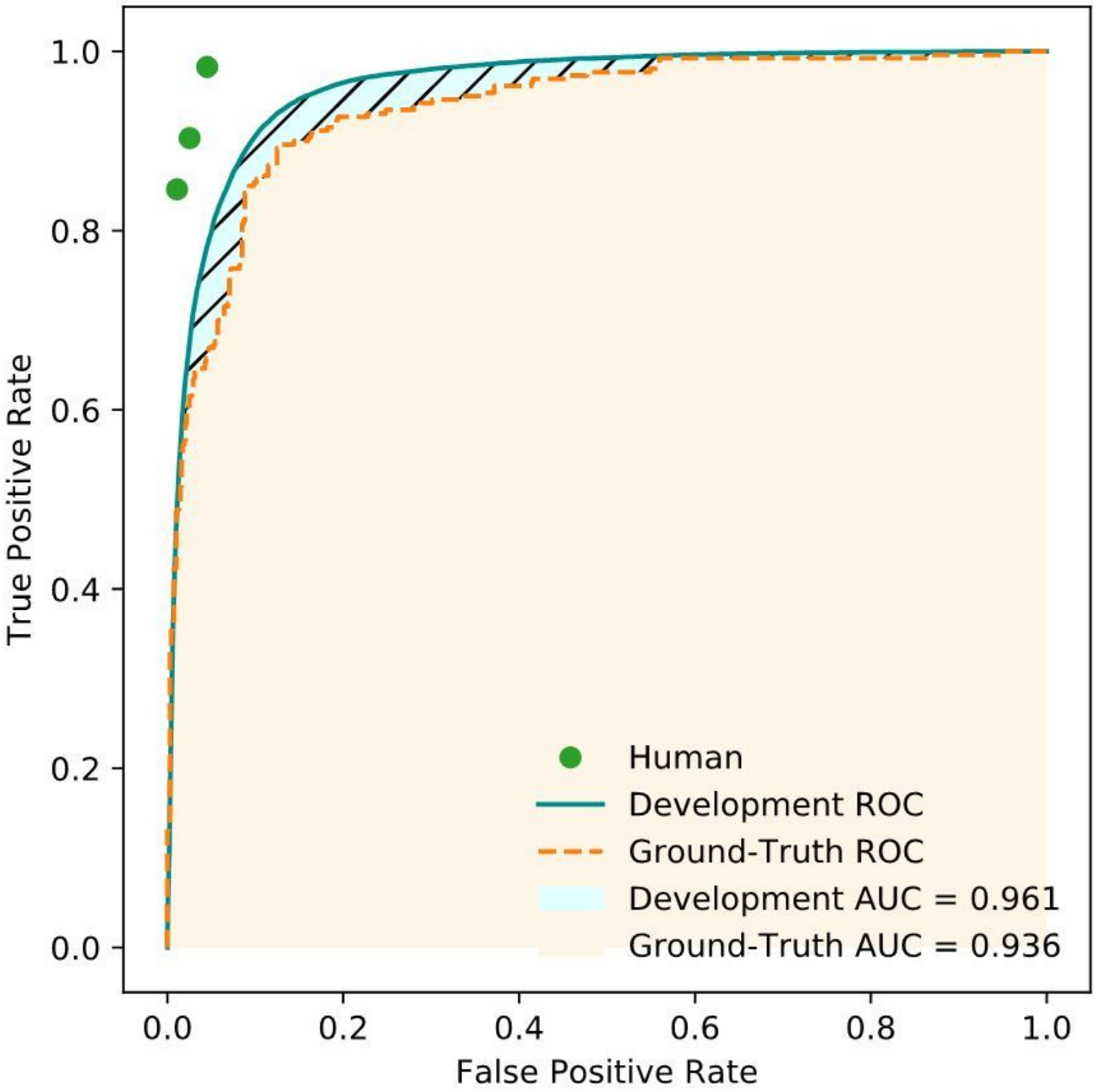

Figure 2

Extrinsic evaluation of Word2vec embeddings with identical hyperparameters, but different training corpora.

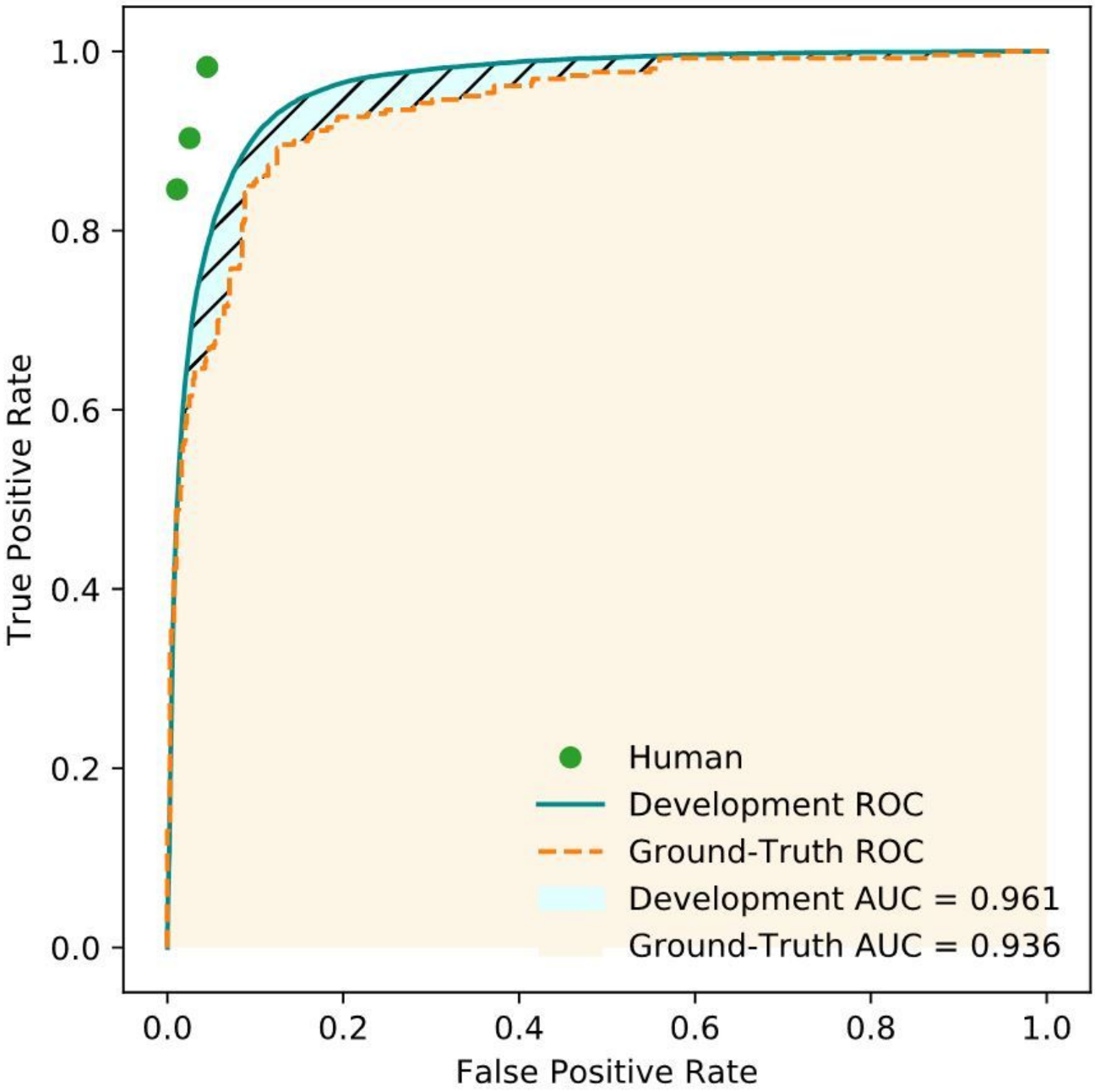

Figure 2

Extrinsic evaluation of Word2vec embeddings with identical hyperparameters, but different training corpora.