Identifying and Extracting Patient Smoking Status information from Clinical Narrative Texts in Spanish.

Rosa L. Figueroa, D.Sc. IEEE-EMBS Member, Diego A. Soto, and Esteban J. Pino, D.Sc. IEEE-EMBS Member.

Abstract— **In this work we present a system to identify and extract patient's smoking status from clinical narrative text in Spanish. The clinical narrative text was processed using natural language processing techniques, and annotated by four people with a biomedical background. The dataset used for classification had 2,465 documents, each one annotated with one of the four smoking status categories. We used two feature representations: single word token and bigrams. The classification problem was divided in two levels. First recognizing between smoker (S) and non-smoker (NS); second recognizing between current smoker (CS) and past smoker (PS). For each feature representation and classification level, we used two classifiers: Support Vector Machines (SVM) and Bayesian Networks (BN). We split our dataset as follows: a training set containing 66% of the available documents that was used to build classifiers and a test set containing the remaining 34% of the documents that was used to test and evaluate the model. Our results show that SVM together with the bigram representation performed better in both classification levels. For S vs NS classification level performance measures were: ACC=85%, Precision=85%, and Recall=90%. For CS vs PS classification level performance measures were: ACC=87%, Precision=91%, and Recall=94%.**

I. Introduction

In 2010, the Chilean government reported that 40.6% of the population is engaged in smoking and tobacco use. Moreover, everyday 41 people die because of smoking related diseases. This percentage ranks among the highest of South America, making this issue one of the main concerns of the public health department in our country [1].

Electronic Medical Records (EMR) introduction in health care systems has enabled researchers to extract information such as patient's habits using text mining techniques. During the last decade in Chile the use EMR systems was not massive. However, this situation is changing rapidly due to the government certification process that hospitals must undertake. This certification process recommends and promotes the use of EMR systems to facilitate documentation and access of patient's information among health care professionals [2].

Identifying smokers within a health care institution is a key factor to obtain statistics and design preventive care plans. EMRs can be a valuable source to extract information about risk factors and patient's habits when the quality of the documentation is appropriate [3]–[6]. Usually, the patient's data in EMR systems is stored in two ways; using structured fields and narrative fields. These systems usually have small search features that allow to get statistics on some indicators using the structured data stored, but this capacity does not always includes information stored in narrative fields [3], [7].

Identification and extraction of smoking status information from narrative text is a challenging task for Natural Language Processing (NLP) researchers for three main reasons identified from previous literature [8]–[11]. First, confidentiality issues make the access to the data very limited. Second, the texts usually have a lack of grammatical structure and may also include misspelled and ambiguous terms. And third, the use of EMR data depends on the proper coding and entry of the information by practitioners. Some studies have shown that practitioners don't always record information in all the available fields and sometimes they misuse the fields categorizing some recorded information in the wrong fields [10], [12]. Despite those limitations, in the past there have been successful attempts to automatically extract symptoms and habits from narrative EMR fields. In 2006, the authors of the Informatics for Integrating Biology to the Bedside (i2b2) project, released a set of de-identified medical discharge records from Partners HealthCare System and announced the smoking classification challenge. The records were annotated by pulmonologists providing the following classes: Past Smoker, Current Smoker, Smoker, Non-Smoker, and Unknown. As a result, the organizers report in [13] that most of the participants were successful in classify the Unknown and Non-Smokers categories but they had some difficulties classifying the Past Smoker, current Smoker and Smoker categories. Another attempt on identifying smoking status from narrative texts from EMR was presented in [3]. The purpose of this study was to evaluate the availability and quality of the documented data in EMRs. The authors extracted smoking status from EMRs using NLP systems, and combining this information with the one provided by the structured fields of the records to measure coverage and correlation with the actual prevalence of smokers in the sample. They found that the availability of the smoking status increased from 11.6% to 64% when supplementing the information from the structured fields with the extracted information from free-text.

The completeness of the reported information in EMRs is also an issue that automatic extraction methods for risk factors or habits should face [5], [8], [10], [12],[13]. One example of this is a study we conducted on documentation and form of entry practices, presented in [12], where we found that only a 6.4% of the analyzed medical records had information about patient's smoking status.

Research supported by Fondecyt Proyect n° 11121463.

R.L.F., D.A.S. and E.J.P are with the University of Concepción, Concepción, Chile. (D.A.S. is the corresponding author. e-mail: dsotoc@udec.cl Phone: +56-41-2204780 fax: +56-41-2246999; R.L.F. e-mail: rosa.figueroa@udec.cl; E.J.P. e-mail: estebanpino@udec.cl;).

This study was conducted at Guillermo Grant Benavente hospital (HGGB) located in Concepción, Chile. The EMR system used in this institution has structured fields to report risk factors, habits, some vital signs such as blood pressure, and primary diagnosis and narrative fields to report patient's medical history, physical examination, notes, and indications.

The aim of this work is to extract patient smoking status using text mining techniques from outpatient reports obtained from the HGGB EMR system. Our work will face not only the intrinsic challenges of extracting risk factors from narrative texts of medical systems, but also the challenge of processing medical reports that are completely written in Spanish.

II. Materials and Methods

A. Dataset

The dataset used in this study was obtained from the EMR system at the HGGB. We used a total of 255,260 outpatient notes from general medicine, cardiology, endocrinology, pulmonology, and otolaryngology medical specialties. This dataset corresponds to the information registered during the years 2011 and 2012. For the purpose of this study, we only used information included on the narrative text fields.

B. Preprocessing

The texts used in this study were preprocessed before conducting expert annotation. The goal of the preprocessing stage is first to eliminate records that do not contain information about smoking status and second to normalize the texts in order to facilitate the feature extraction process. Each report was normalized. Words were all changed to lower case, non alphanumeric characters were removed except spaces, tabs, and hyphen, and so were stops words (e.g., a, the, on, etc). The stop words list was tailored to this study to preserve words used in negations.

A quick inspection to our dataset revealed that not all the reports had information about smoking habits in narrative text fields. To address this issue we used a custom dictionary of smoking related keywords to filter out records that were not relevant to this study. This last step reduced our dataset to 2,028 reports.

C. Annotations

We defined four possible smoking categories: Smoker (S), Non-Smoker (NS), Current Smoker (CS), and Past Smoker (PS). To define the classes we used as a guideline the annotation process carried out in [13].

Four students with biomedical engineering backgrounds revised and annotated each record with one of the four possible designed smoking categories. They were trained in the use of the annotation tool developed for this purpose by working with 20 examples before beginning with the actual annotations.

We asked the annotators to use only the information present on the narrative fields of the records to assign categories. They were trained to first assign S and NS categories. If the S category was assigned to a record, the annotator had to recognize is there was temporal information about the habit to categorize the record as PS or CS. In order to label a document as PS, the annotator should look for temporal information that would indicate if the patient quit smoking at least 6 month ago.

At the end of the annotation process, 2,028 documents were annotated. We evaluated an inter-annotator agreement using *Fleiss's Kappa* (*k*) index [14]. *Fleiss's Kappa* is a statistical index to measure inter-annotator agreement when raters are more than two. When the inter-rater agreement is poor *k* values are expected to be less than zero; when the inter-rater agreement is almost perfect *k* values are expected to be in the range 0.81-1.

To generate our Gold Standard for classification, we asked to a fifth annotator to validate the assigned classes and to resolve any disagreement in the textual judgment.

D. Feature Extraction

The annotated records were split in sections using section headers such as diagnosis, patient's medical history, etc. Then, the sections that did not contain information about smoking status were filtered out. Now, our final dataset for classification had 2,455 documents, where each document corresponded to a section of the annotated records. Table I shows a description of the datasets in terms of the number of records and the class distribution.

TABLE I. SUMARY OF DATASETS USED FOR CLASSIFICATION.

Dataset	Category I	Category II	Total
SNS	1,443 (S)	1,012 (NS)	2,455
PSCS	1,394 (CS)	402 (PS)	1,796

Normalized documents were "tokenized" using regular expressions to produce single word tokens (N1) and bigrams[1] (N2) representations.

$$tf - idf_{t,d} = tf_{t,d} \times idf_{t,D} \quad (1)$$

$$idf_{t,D} = log \frac{N}{df(t_k)}$$

tf : term frequency
idf: inverse document frequency
N:number of documents
$df(t_k)$: number of documents that contain t_k

For each of the representations, we built two bag of words models using term frequency-inverse document frequency (TF-IDF) (see Eq. 1) score to represent the occurrences of the words in the document [15]. In the final model, each document is represented as a word/bigram vector where the jth element of the ith column is the TF-IDF score of the occurrences of the jth word/bigram in the ith document.

E. Classification and Evaluation

For the classification process, we use the implementations of Support Vector Machines (SVM) and Bayesian Networks (BN) available from Weka [16] to build models. SVM was set to use a linear kernel and normalization/standardization was turned off. BN was set to use k2 algorithm that uses a

[1] A bigram can be thought as sliding window place over the text that shows two words at a time.

heuristic search strategy to find the most probable belief network structure given the set of examples [17].

We divided the classification problem in two binary problems: recognizing between S vs NS and PS vs CS.

Using the Train/Test percentage split feature of Weka, we split our dataset as follows: a training set containing 66% of the available documents that was used to build classifiers and a test set containing the remaining 34% of the document that was used to test and evaluate the model.

The performance measures used to evaluate the classifier were: Accuracy (ACC), Precision or Positive predicted value, Recall or True positive rate (TPR), and F-score. In order to perform a more detailed analysis about the behavior of the methods, we also calculated the False Positive rate (FPR), False Negative rate (FNR), and True Negative rate (TNR). Equations 2, 3, 4, and 5 show how each performance measure was calculated. Each experiment was repeated ten times and performance measures were averaged over the ten runs. To compare performance measures we used a paired t-test with a significance level of 0.05.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (2)$$

$$Precision = \frac{TP}{TP+FP} \quad (3)$$

$$Recall = \frac{TP}{TP+FN} \quad (4)$$

$$F - Score = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (5)$$

TP: True Positives
FP: False Positives
TN: True Negatives
FN: False Negatives

III. RESULTS

A. Inter-rater agreement.

As we mentioned on section II-C, we used *Fleiss's Kappa index* (*k*) to obtain the level of agreement between the annotators. For the first annotation level which decided between S or NS category, we obtained a value of k=0.86. For the second level of annotation which decided between CS or PS categories, we obtained a value of k=0.81. Both values of k can be considered indicators of almost perfect agreement between annotators [18]. This result indicates that our Gold Standard is reliable and can be used for classification.

B. Classification Results

Table II and III show the performance measures for both classification problems. In table II, we can observe for most of the cases that FPR is slightly higher than the FNR. This result indicates that our system had some difficulties to handle negative categories NS and PS. From table III, we can observe that in most of the cases SVM lead to better ACC, Precision, Recall and F-score than BN. In terms of ACC and Recall, we can observe that, in most of the cases, SVM performed statistically better than BN (significance level of 0.05, except PS-CS using bigrams).

TABLE II. EVALUATION MEASURES FOR CLASSIFIER PERFORMANCE I

		FPR	FNR	TNR
S vs NS				
SVM				
	N1	0,24	0,16	0,76
	N2	0,23	0,10	0,77
BN				
	N1	0,29	0,27	0,71
	N2	0,31	0,11	0,69
PS vs CS				
SVM				
	N1	0,30	0,08	0,7
	N2	0,34	0,06	0,66
BN				
	N1	0,40	0,11	0,6
	N2	0,41	0,04	0,59

TABLE III. EVALUATION MEASURES FOR CLASSIFIER PERFORMANCE II

		ACC	Precision	Recall(TPR)	F-Score
S vs NS					
SVM					
	N1	0,81(*)	0,83(*)	0,84(*)	0.84(*)
	N2	0,85(*)	0,85(*)	0,90(*)	0.87(*)
BN					
	N1	0,72	0,78	0,73	0.75
	N2	0,81	0,81	0,89	0.85
PS vs CS					
SVM					
	N1	0,87(*)	0,91(*)	0,92(*)	0.92(*)
	N2	0,87	0,91(*)	0,94	0.92
BN					
	N1	0,83	0,89	0,89	0.89
	N2	0,87	0,89	0,96(*)	0.92

(*) indicate that the corresponding classifier was found statistically better.

In terms of feature representation, bigram or N2 representation leads to better performance than single word tokens or N1 representation. We believe that using a N2 representation, our model was better at recognizing positive classes S and CS.

IV. DISCUSSION AND CONCLUSION

In this work, we extracted patient's smoking status from outpatient records in Spanish. Our available dataset got drastically reduced after we filtered out records with no information about the smoking habit. Findings from our previous study on documentation and entry practices about patient smoking status told us that practitioners don't report smoking habit very often. Although, one could believe that practitioners would rather report this kind of information on a narrative field, it turns out that the amount of records with

information on those field is still very low. Consequently, it is essential to educate practitioners on the importance of documenting habits for research and also to design preventive health care plans.

To extract patient smoking status, we used NLP and machine learning techniques. Specifically, we used two feature representations: one based on single word tokens and another one based on bigram words. Using this representations we built SVM and BN classification models. The results indicate that word bigrams representation (N2) together with SVM classifier performed better in both classification levels. For S vs NS classification level performance measures were: ACC=85%, Precision=85%, Recall=90%, and F-score = 87%. For CS vs PS classification level performance measures were: ACC=87%, Precision=91%, Recall=94%, and F-score=92%. We believe our results can be explained by the robustness of SVM to handle high dimensionality problems together with the ability of bigram models to capture negated verbs and temporal features.

From table III, we can observe that both classifiers performed better when recognizing between CS vs PS. This result can be explained by the amount of features with temporal information about the smoking status; somehow practitioners were more informative in terms of documentation to explain how long a patient has been smoking or since when the patient have quit smoking than to report a non smoker patient.

We found that our extraction system had difficulties handling expressions of negations. This can be appreciated on table II results, where FPR is slightly higher than FNR. When using bigrams, our system obtained a lower FNR for both SVM and NB. Using two consecutive words as features, somehow helped our system to better identify examples that contained negation tokens but were not necessarily grouped with smoking keywords. To our knowledge, most of the tools to recognize negation in Spanish need syntactic annotations on the text. Medical records in Spanish, as well as in other languages, don't have a well-defined syntactic structure thus applying a general domain part of the speech tagger won't work on these kinds of texts.

In terms of single word token representation, a factor that could have contributed to render more difficult negation extraction is the length of the documents. A longer document should be more difficult to classify since it may contain more features but no necessarily relevant.

Overall, we believe our system performed reasonably well on both classification levels given the low number of records containing smoking status information and the nature of the EMR sublanguage. For future work, we expect to test our system using more data and also to improve negation extraction using syntactic annotations on the records.

ACKNOWLEDGMENT

The authors want to thank CONICYT (Chilean National Council for Science and Technology Research), Universidad de Concepcion, and FONDECYT (grant N°11121463) for their support.

REFERENCES

[1] G. Valdivia Cabrera, "Encuesta nacional de salud 2009-2010: Enseñanzas y desafíos", *Rev. Chil. Enfermedades Respir.*, vol. 27, nº 1, pp. 5–6, mar. 2011.

[2] "Derechos y deberes de los pacientes en salud - Ley fácil - Biblioteca del Congreso Nacional de Chile". [En línea]. Disponible en: http://www.bcn.cl/leyfacil/recurso/derechos-y-deberes-de-los-pacientes-en-salud. [Accedido: 06-abr-2014].

[3] C.-Y. Wu, C.-K. Chang, D. Robson, R. Jackson, S.-J. Chen, R. D. Hayes, y R. Stewart, "Evaluation of Smoking Status Identification Using Electronic Health Records and Open-Text Information in a Large Mental Health Case Register", *PloS One*, vol. 8, nº 9, p. e74262, 2013.

[4] P. J. McCormick, N. Elhadad, y P. D. Stetson, "Use of semantic features to classify patient smoking status", en *AMIA Annual Symposium Proceedings*, 2008, vol. 2008, p. 450.

[5] A. J. Fouwels, S. J. Bredie, H. Wollersheim, y G. M. Schippers, "A retrospective cohort study on lifestyle habits of cardiovascular patients: how informative are medical records?", *BMC Health Serv. Res.*, vol. 9, nº 1, p. 59, 2009.

[6] Q. T. Zeng, S. Goryachev, S. Weiss, M. Sordo, S. N. Murphy, y R. Lazarus, "Extracting principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system", *BMC Med. Inform. Decis. Mak.*, vol. 6, nº 1, p. 30, 2006.

[7] M. Martinell, J. St\a alhammar, y J. Hallqvist, "Automated data extraction-A feasible way to construct patient registers of primary care utilization", *Ups. J. Med. Sci.*, vol. 117, nº 1, pp. 52–56, 2012.

[8] A. N. Goudswaard, K. Lam, R. P. Stolk, y G. E. Rutten, "Quality of recording of data from patients with type 2 diabetes is not a valid indicator of quality of care. A cross-sectional study", *Fam. Pract.*, vol. 20, nº 2, pp. 173–177, 2003.

[9] P. Dhiman, J. Kai, L. Horsfall, K. Walters, y N. Qureshi, "Availability and Quality of Coronary Heart Disease Family History in Primary Care Medical Records: Implications for Cardiovascular Risk Assessment", *PloS One*, vol. 9, nº 1, p. e81998, 2014.

[10] S. Garies, D. Jackson, B. Aliarzadeh, K. Keshavjee, K. Martin, y T. Williamson, "Improving usability of smoking data in EMR systems", *Can. Fam. Physician*, vol. 59, nº 1, pp. 108–108, 2013.

[11] R. Koeling, J. Carroll, A. R. Tate, y A. Nicholson, "Annotating a corpus of clinical text records for learning to recognize symptoms automatically", en *Proceedings of the 3rd Louhi Workshop on Text and Data Mining of Health Documents*, 2011, pp. 43–50.

[12] R. L. Figueroa y D. A. Soto, "Medical Records Systems Usage in Developing Countries : A study of documentation and form entry practices". 2013.

[13] Ö. Uzuner, I. Goldstein, Y. Luo, y I. Kohane, "Identifying patient smoking status from medical discharge records", *J. Am. Med. Inform. Assoc.*, vol. 15, nº 1, pp. 14–24, 2008.

[14] J. L. Fleiss, "Measuring nominal scale agreement among many raters.", *Psychol. Bull.*, vol. 76, nº 5, p. 378, 1971.

[15] V. Yatsko, S. Dixit, A. J. Agrawal, S. T. Y. Myint, y M. M. Khin, "TF* IDF Revisited", *intelligence*, vol. 16, nº 4, p. 2, 2013.

[16] G. Holmes, A. Donkin, y I. H. Witten, "Weka: A machine learning workbench", en *Intelligent Information Systems, 1994. Proceedings of the 1994 Second Australian and New Zealand Conference on*, 1994, pp. 357–361.

[17] X.-W. Chen, G. Anantha, y X. Lin, "Improving Bayesian network structure learning with mutual information-based node ordering in the K2 algorithm", *Knowl. Data Eng. IEEE Trans. On*, vol. 20, nº 5, pp. 628–640, 2008.

[18] A. J. Viera y J. M. Garrett, "Understanding interobserver agreement: the kappa statistic", *Fam Med*, vol. 37, nº 5, pp. 360–363, 2005.