Deep Bidirectional Recurrent Neural Networks as End-To-End Models for Smoking Status Extraction from Clinical Notes in Spanish.

Santiago Esteban, MD MPH[1,2*], Manuel Rodríguez Tablado, MD[1], Francisco E. Peper, MD[1], Sergio A. Terrasa, MD MSc[1,3], Karin S. Kopitowski, MD[1,2]

[1]Family and Community Medicine Division, Hospital Italiano de Buenos Aires, Buenos Aires, Argentina
[2]Research Department, Instituto Universitario Hospital Italiano, Buenos Aires, Argentina
[3]Public Health Department, Instituto Universitario Hospital Italiano, Buenos Aires, Argentina

Abstract

Although natural language processing (NLP) tools have been available in English for quite some time, it is not the case for many other languages, particularly for context specific texts like clinical notes. This poses a challenge for tasks like text classification in languages other than English. In the absence of basic NLP tools, manually engineering features that capture semantic information of the documents is a potential solution. Nevertheless, it is very time consuming. Deep neural networks, particularly deep recurrent neural networks (RNN), have been proposed as End-to-End models that learn both features and parameters jointly, thus avoiding the need to manually encode the features. We compared the performance of two classifiers for labeling 14718 clinical notes in Spanish according to the patients' smoking status: a bag-of-words model involving heavy manual feature engineering and a bidirectional long-short-term-memory (LSTM) deep recurrent neural network (RNN) with GloVe word embeddings. The RNN slightly outperforms the bag-of-words model, but with 80% less overall development time. Such algorithms can facilitate the exploitation of clinical notes in languages in which NLP tools are not as developed as in English.

Introduction

Globally, tobacco is one of the main risk factors for premature death[1] being responsible for 11.5% of the worldwide annual deaths (6.5 million)[2]. It is also one of the five main risk factors of disability[2]. In 2015, the global prevalence of daily tobacco use was estimated at 25% (uncertainty interval(UI) 24.2-25.7) in men and 5.4% (UI 5.1-5.7) in women with a marked decrease since 1990[2]. However, despite these general trends, there is a high level of heterogeneity among countries and this decline was not as marked in countries with low and middle socio-demographic levels, especially in women[2]. This is evident in that of the 933 million daily smokers, 80% reside in such countries[3]. However, clinical and public health research on risk factors such as smoking is often restricted, among other things, by the scarcity of trained human resources and economic resources[4, 5].

In this context, electronic health records (EHR) are presented as data sources that facilitate clinical research, especially taking into account the increasing adoption rate in both developed and developing countries[6].

Information regarding tobacco consumption is usually recorded in the EHRs in different formats, both as structured problems but also as free text, usually complementing each other[7]. However, there is a clear tendency to register a large part of the relevant information on smoking in the free text due, among other reasons, to the lack of flexibility of the more structured recording systems[8]. Thus, many authors have explored methods based on natural language processing (NLP) for the extraction of information regarding the patients' smoking status from clinical notes[9-17].

The task of classifying texts by assigning them a category or label (for example: smoker, non-smoker, ex-smoker) has traditionally been solved using techniques based on hard-coded rules and techniques based on statistical or machine-learning[18]. One of the most disseminated and also simplest approaches is the bag-of-words model. In this model each n-gram (term or token) of the corpus is treated as a feature and a value is assigned to each document based on weighted counts of the terms in each document. Once this document-term-matrix is created, a classifier is iteratively trained to minimize a cost function in order to find the best separation plane between the categories and then select those variables that best help to accurately label the observations. The performance of this method can be improved by developing variables that better capture the linguistic structure of the text. Many NLP tools have been designed for such subtasks like sentence parsing, dependency parsing, negation detection, part-of-speech (POS) tagging or entity name recognition.

[*] Correspondece to: Santiago Esteban <santiago.esteban@hospitalitaliano.org.ar>

However, these tools are, for the most part, language-specific and, in many cases, context-specific. Which means that, in order to apply them in a language or context different from the one they were developed in, it is necessary to retrain or even redesign them. Clinical notes from EHRs are clearly different from usual texts: specific jargon, incomplete or grammatically incorrect sentences and a large number of abbreviations are used. This implies that, in order to develop the beforementioned NLP tools for specific languages and contexts, it is necessary to develop a training corpus, usually annotated by linguistics experts, which is an expensive and slow process that in many cases ends up being prohibitive. This aspect is rather important considering that most of the research and tools in NLP have been developed in English and most of the developing countries, where it would be essential to promote research on smoking, do not have English as their primary language and therefore, these tools are impractical.

A more accessible potential solution is the extension of the models described so far by developing simple linguistic models and manually engineering features specific to the classification task. This is usually done by rigorous exploration of the text and with specific knowledge of the classification problem itself. Examples of this would be the identification of key topic words, the development of specific tools to detect key term negations and modifiers. This, although feasible, requires the investment of many hours in the process of manual feature engineering and model optimization. On the other hand, these models are not usually exportable to other classification problems.

End-to-End models for text classification

The models described so far are usually implemented in two phases, where the features and the parameters of the model are developed separately and by different actors: the features by humans, the parameters by the algorithm. The philosophy of 'End-to-End' models seeks that both, features and parameters are learned jointly and by the same actor: the algorithm. In recent years ‘End-to-End' models have been excelling at tasks such as image classification[19], machine translation[20], autonomous driving[21] and speech recognition[22]. Most of these examples have used deep neural networks but the 'End-to-End' philosophy exceeds this type of models. Basically, End-to-End models remove intermediate steps by freeing the human from the task of manually engineering features and can potentially help find patterns in the data that would be difficult for a human to discover. In the specific case of NLP, they can avoid having to develop many of the most widely used tools (parsers, POS taggers, etc.). This is particularly advantageous for languages or contexts for which these tools are not available or don’t perform very well.

Bidirectional Long Short Term Memory (LSTM) Recurrent Neural Networks (RNN) as End-to-End models for text classification

Recurrent neural networks are a type of artificial neural network architecture that allows to process series of data sequentially. Basically, the output of a neuron t-1 is part of the input of the neuron t. Thus, the output of the neuron t depends on the output of the neuron t-1. This ability to 'remember' the states of previous neurons allows modeling of data sequences by capturing sequential patterns. Modifications to the structure of basic RNNs (Long Short Term Memory[23]; Gated Recurrent Units[24]) have allowed these algorithms to learn even more distant dependencies within the series. This characteristic makes them natural candidates for the classification of texts, since they can represent key aspects of linguistic structures without human aid[25]. On the other hand, the bidirectional processing of the sequences allows the output of the neuron t to be also influenced by data later in the sequence. In turn, the representation of words by means of pretrained word embeddings[26-28] further increases the representational capacity of these models.

Background

In recent years, End-to-End models implemented through artificial neural networks (ANN) have become more relevant for NLP tasks in general[29-32] as well as in clinical texts. In 2016, Jagannatha et al. [33] compared the performance of a bidirectional RNN vs conditional random fields (CRF) [34] for detecting medical events in notes from an EHR. They found improvements in accuracy, recall and F1 score vs. the base system. Other applications of RNNs in medical texts have explored their performance in tasks such as deidentification of clinical notes[25], information extraction[35, 36], entity name recognition[37], relation classification[38] and text classification[39] with superior results to the reference models. However, we find did not find any articles that assessed the capacity of these models to classify clinical notes in Spanish.

Aim

The main objective of our research is the comparison of two models developed for the classification of clinical notes in Spanish according to the patients' smoking status: a feed-forward neural network with manually engineered features (baseline model) and a deep bidirectional LSTM RNN as an End-to-End model.

Dataset

Clinical notes were obtained from the EHR database at Hospital Italiano de Buenos Aires, Buenos Aires, Argentina. From a random sample of 6000 notes, we established that only 3.65% (95% CI 3.14 - 4.1) contained some type of information related to smoking. To optimize the manual labeling process given the low proportion of notes with information on tobacco use, we defined a filter based on four stems ('tab', 'cig', 'fum', 'tbq') to achieve a sample that had a greater proportion of notes with information on tobacco. We evaluated this filter on another sample of 6000 notes sampled at random (also manually labeled), resulting in a negative predictive value of 0.99948 (95% CI 0.9985 - 0.9999) and a positive predictive value of 0.83077 (95% CI 0.7796 - 0.8743). Using this filter, we selected a sample of notes that contained any of the four stems from a random sample of 3054 patients 18 years of age or older, between 2005 and 2015. This yielded a total of 14718 clinical notes. The notes were labeled by three domain experts (family doctors) as 'current smoker', 'current non-smoker' and 'clinical note without information on tobacco' (inter-annotator agreement: kappa 0.97, 95% CI 0.9 - 1). The classification 'previous smoker' was avoided since we found that in many cases these ex-smokers were described as non-smokers in the notes and that patients were almost never described as 'never smokers'. So, the 'current non-smoker' category includes 'never smokers' and 'ex-smokers'. Table 1 shows the characteristics of the dataset.

Table 1. Dataset characteristics.

Manual label	Number of documents (%)
Current Non-Smoker	5652 (38.4)
Current Smoker	4933 (33.5)
Clinical note without information on tobacco	4133 (28.1)

Methods

Once the filtered sample of 14718 clinical notes was obtained, it was randomly split into a training set (11775 notes, 80%) and a test set (2943 notes, 20%). The test set was set aside and only used for the final evaluation of the final models. The training set was used to train both models, using 10-fold cross-validation for the hyperparameter tuning process. The final models' performance in the test set was assessed using per class and macro-averaged precision (positive predictive value), recall (sensitivity) and F1-score. All analyses were performed in R 3.4.1[40] using the dplyr, ggplot, tidyr, text2vec, keras, tidytext, tensorflow and caret packages. All models were trained through the Keras API[41] with TensorFlow[42] as backend. Figure 1 shows an overview of the models' development and validation process.

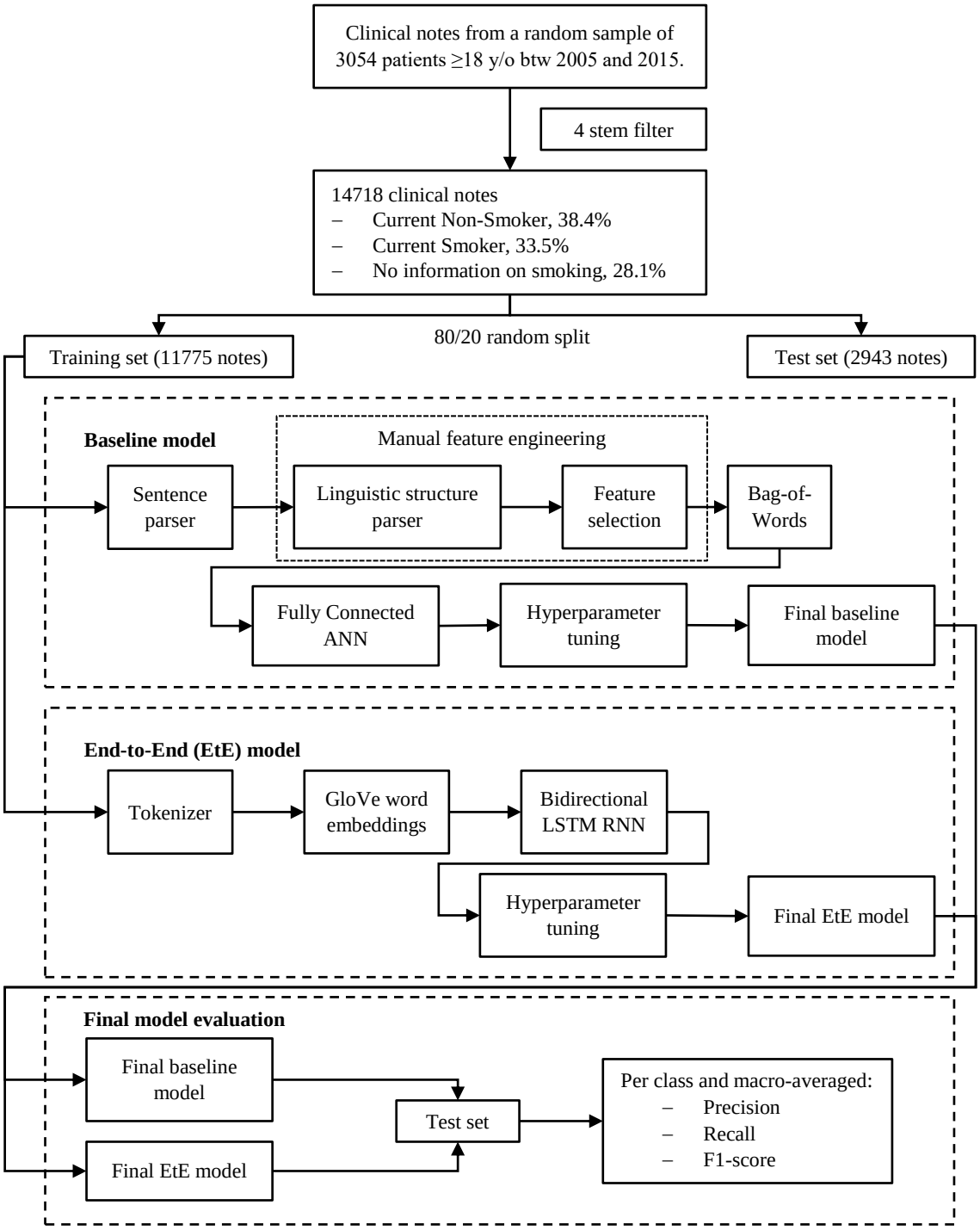

Figure 1. Overview of the models' development and validation process.
ANN: artificial neural network; EtE: End-to-End model.

Baseline model

The development of the model was divided into two stages: 1. Development of a simple linguistic model and manually engineering the features, 2. Training of the classifier through a feed-forward neural network.

Linguistic parser development (feature engineering and selection)

Initially the training set was explored. Following this, twelve central terms related to smoking were extracted. Then a module was developed that detects negations of those terms implemented using regular expressions. This module detects negation terms 25 characters before and 10 after the central terms, unless some punctuation sign is detected such as commas, periods, question marks, semicolons and parentheses. Then, modifier phrases (positive and negative) of the central terms were detected. Using a logic similar to that of the negation module, these modifiers were only

detected if they were in the proximity of the central terms. Then the detection of negations of these modifiers was added. Finally, a list of terms that would detect other subjects related to the central terms (mother, father, brother, spouse, etc.) was developed (figure 2). Once this model was defined, the documents were parsed into sentences and only those that had any of the central terms were kept. A table was constructed in which each row was a sentence and each column contained the count of each of the terms detected in the document. Thus, initially, 9021 variables were defined. This number was later reduced to 561 after dropping those that had zero values in all observations, those with a correlation greater than 0.85, and those whose entropy-based importance was zero.

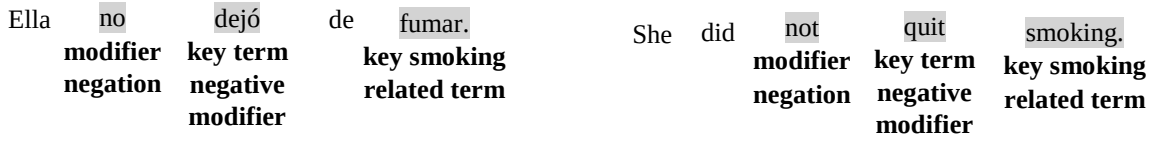

Figure 2. Example of the simple linguistic structure parser (Spanish and English).

Development of the classifier

The final classifier resulted in a feed-forward fully connected artificial neural network with three hidden dense layers (483, 377 and 358 hidden nodes respectively) and a Softmax layer as the output layer (figure 3). It was optimized for 150 epochs in batches of 128 observations by mini-batch gradient descent with momentum of 0.9, learning rate of 0.2 and learning rate decay of 0.004. We used Rectified Linear Units (ReLU) as activation functions between layers and categorical cross entropy as the objective function. All hyperparameters were tuned using 10-fold cross validation using a random number search approach. L2 regularization and dropout[43] were evaluated to avoid overfitting, but they did not have a significant impact on the model's performance in the validation set so they were not incorporated into the model. The total development time of the model was 320 hours, using an average of four hours per day five days a week for 16 weeks. The first 280 hours were used in the development and selection of variables; the last 40 were assigned to training the classifier.

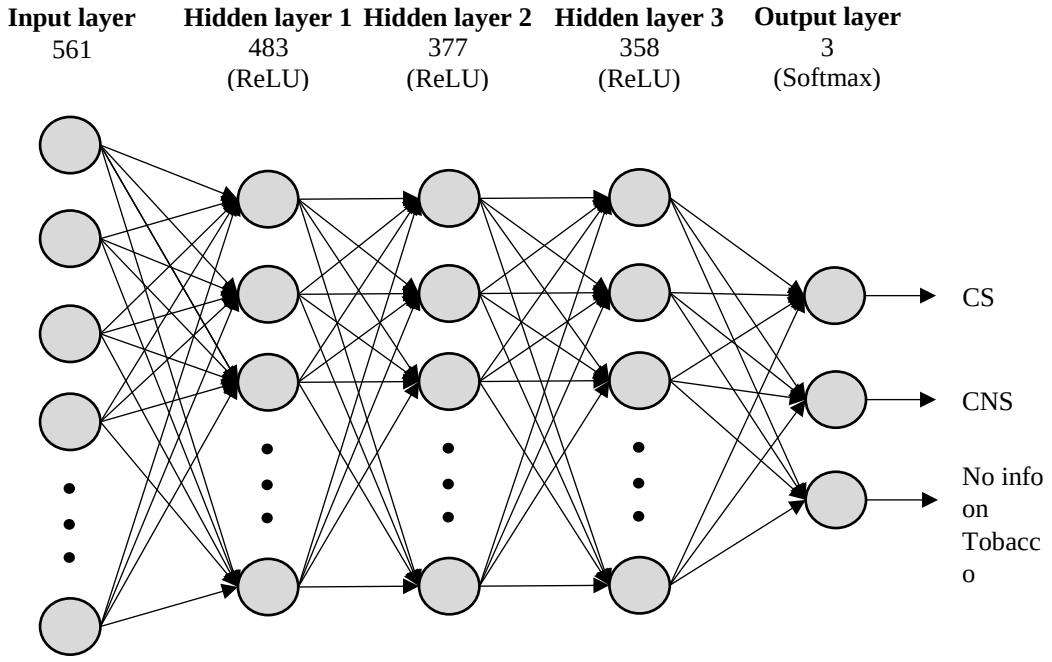

Figure 3. Baseline model: Fully connected feed forward artificial neural network.
CS: Current smoker; CNS: Current non-smoker; ReLU: Rectified linear unit.

End-to-End model

Model development

We trained a RNN at the document level with an embedding layer, three hidden bidirectional LSTM layers (100, 100, 100 hidden nodes per layer per direction), three hidden dense layers (100, 100, 50 hidden nodes) and a Softmax layer

to output the probabilities for each category (figure 4). We used recurrent dropout[44] (dropout probability = 0.2) to avoid overfitting in the LSTM layers and regular dropout (dropout probability = 0.2) in the dense layers. The activation function between the LSTM layers is tanh and ReLU between the dense layers. As optimization algorithm we used RMSprop[45]. We also used categorical cross entropy as the objective function. All hyperparameters were tuned using 10-fold cross validation. The total time of development of the model was 56 hours, using an average of 8 hours per days for 7 days. The first 10 hours were used for the training and exploration of word embeddings; the last 46 were assigned to the training of the RNN.

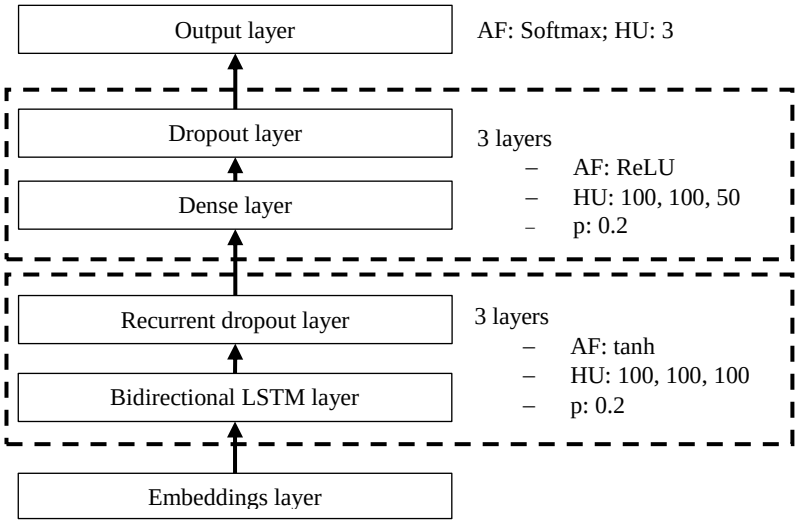

Figure 4. Structure of the bidirectional long-short-term-memory recurrent neural network.
AF: activation function; HU: hidden units; p: dropout probability.

GloVe Word embeddings

In order to initialize the embedding layer of the RNN, we pretrained word embeddings from a corpus formed by the training set documents plus the 12,000 documents of the samples described in the 'Dataset' section, using the model described by Pennington et. al[28]. A window of 10 words was selected. Of the 36,558 unique words, those that appeared less than 5 times throughout the corpus were discarded. The remaining 9,483 tokens were trained to obtain 200 dimensional vectors.

Results

Table 2 shows the performance of both models on the test set. The RNN was slightly superior to the base model in all metrics (recall: + 0.06, accuracy: + 0.09 and F1-score: + 0.09.) In particular, it correctly classified a greater number of 'current non-smokers' that the baseline model misclassified as 'current smokers'.

Table 2. Performance of both models in the test set.

	Confusion Matrix				**Recall**		**Precision**		**F1 Score**	
	Pred. x Reference				**Per class**	**M**	**Per class**	**M**	**Per class**	**M**
Baseline model		CNS	CS	NTI	CNS: 0.941 CS: 0.952 NTI: 0.976	0.956	CNS: 0.968 CS: 0.928 NTI: 0.967	0.954	CNS: 0.954 CS: 0.94 NTI: 0.971	0.955
	CNS	1070	33	2						
	CS	54	937	18						
	NTI	13	14	802						
Bidirectional RNN		CNS	CS	NTI	CNS: 0.96 CS: 0.955 NTI: 0.98	0.96	CNS: 0.971 CS: 0.96 NTI: 0.959	0.963	CNS: 0.966 CS: 0.957 NTI: 0.97	0.964
	CNS	1090	27	5						
	CS	28	940	11						
	NTI	17	17	806						

CNS: Current Non-Smoker; CS: Current Smoker; NTI: No Tobacco Information; M: Macro-averaged metric.

Discussion

We developed a classifier that detects smoking status from clinical notes is Spanish using an End-to-End approach by means of a bidirectional Deep LSTM RNN. It performed slightly better than traditional methods that involve a great investment of time in the manual development of features. To the best of our knowledge, this is the first paper that evaluates the performance of these models in clinical notes in Spanish.

This capability of the RNN to represent the information contained in the documents without any manual development of features is explained by several of its characteristics. Bidirectional sequential processing allows the incorporation of previous and subsequent information to each word and is very similar to the way in which humans understand texts. On the other hand, the different layers of the RNN allow a hierarchical representation of the information, learning more abstract representations of the texts in each layer. This, added to the representation of the words by word embeddings that capture the semantic relationships between words, place the RNN as an excellent choice of End-to-End models.

Not having to engineer features manually facilitates texts analysis in languages other than English. In addition, according to our experience, it saves time (in our case, the development of the RNN required less than 20% of the time it took to develop the baseline model) and allows the analyst to focus on the architecture of the network rather than on the exploration of the notes and extraction of variables. This in turn makes the workflow easily exportable to other classification problems, in various contexts and languages unlike the process of manual variable development. In the particular context of public health management based on EHR data, it is even more valuable since it allows to potentially design, in a short time, precise classifiers for creating electronic phenotypes than can be used in policy evaluation or research.

Our work has several limitations. Among them, we use a bag-of-words model as a baseline model since it is part of our usual workflow, however other alternatives such as conditional random fields or other state-of-the-art models could be excellent baseline models as well, as proposed by other authors[25, 33, 35, 36, 38, 39]. On the other hand, we only evaluated the models in a classification problem with only a few classification categories. However, other teams have had very good results even with more categories and fewer observations[33, 35].

Conclusion

We have shown that deep bidirectional LSTM RNNs as End-to-End models achieve at least the same level of performance in the classification of clinical texts in Spanish, as models with heavy manual feature engineering albeit in less than 20% of the development time. This makes them an important tool for streamlining text processing in languages in which the development of NLP tools has not advanced enough so far. This facilitates the exploitation of already available data sources such as EHRs for research or public health management purposes.

References

1. *Global, regional, and national comparative risk assessment of 79 behavioural, environmental and occupational, and metabolic risks or clusters of risks, 1990-2015: a systematic analysis for the Global Burden of Disease Study 2015.* Lancet, 2016. **388**(10053): p. 1659-1724.
2. Reitsma, M.B., N. Fullman, M. Ng, et al., *Smoking prevalence and attributable disease burden in 195 countries and territories, 1990–2015: a systematic analysis from the Global Burden of Disease Study 2015*. The Lancet. **389**(10082): p. 1885-1906.
3. WHO. *Tobacco fact sheet*. 2016 [cited 2018 03/03/2018]; Available from: http://www.who.int/mediacentre/factsheets/fs339/en/.
4. *Strengthening clinical research capacity in low and middle-income countries. Workshop report.*, T.A.o.M. Sciences, Editor. 2017.
5. Käser, M., C. Maure, B.M.M. Halpaap, et al., *Research Capacity Strengthening in Low and Middle Income Countries – An Evaluation of the WHO/TDR Career Development Fellowship Programme.* PLoS Neglected Tropical Diseases, 2016. **10**(5): p. e0004631.
6. Stone, C.P., *A Glimpse at EHR Implementation Around the World: The Lessons the US Can Learn*. 2014, The Health Institute for E-Health Policy.
7. Wang, Y., E.S. Chen, S. Pakhomov, E. Lindemann, and G.B. Melton, *Investigating Longitudinal Tobacco Use Information from Social History and Clinical Notes in the Electronic Health Record.* AMIA Annu Symp Proc, 2016. **2016**: p. 1209-1218.

8. Chen, E.S., E.W. Carter, I.N. Sarkar, T.J. Winden, and G.B. Melton, *Examining the use, contents, and quality of free-text tobacco use documentation in the Electronic Health Record.* AMIA Annu Symp Proc, 2014. **2014**: p. 366-74.
9. Wu, C.Y., C.K. Chang, D. Robson, et al., *Evaluation of smoking status identification using electronic health records and open-text information in a large mental health case register.* PLoS One, 2013. **8**(9): p. e74262.
10. Wicentowski, R. and M.R. Sydes, *Using implicit information to identify smoking status in smoke-blind medical discharge summaries.* J Am Med Inform Assoc, 2008. **15**(1): p. 29-31.
11. Uzuner, O., I. Goldstein, Y. Luo, and I. Kohane, *Identifying patient smoking status from medical discharge records.* J Am Med Inform Assoc, 2008. **15**(1): p. 14-24.
12. Solberg, L.I., T.J. Flottemesch, S.S. Foldes, B.A. Molitor, P.F. Walker, and A.L. Crain, *Tobacco-use prevalence in special populations taking advantage of electronic medical records.* Am J Prev Med, 2008. **35**(6 Suppl): p. S501-7.
13. Savova, G.K., P.V. Ogren, P.H. Duffy, J.D. Buntrock, and C.G. Chute, *Mayo clinic NLP system for patient smoking status identification.* J Am Med Inform Assoc, 2008. **15**(1): p. 25-8.
14. McCormick, P.J., N. Elhadad, and P.D. Stetson, *Use of semantic features to classify patient smoking status.* AMIA Annu Symp Proc, 2008: p. 450-4.
15. Heinze, D.T., M.L. Morsch, B.C. Potter, and R.E. Sheffer, Jr., *Medical i2b2 NLP smoking challenge: the A-Life system architecture and methodology.* J Am Med Inform Assoc, 2008. **15**(1): p. 40-3.
16. Cohen, A.M., *Five-way smoking status classification using text hot-spot identification and error-correcting output codes.* J Am Med Inform Assoc, 2008. **15**(1): p. 32-5.
17. Clark, C., K. Good, L. Jezierny, M. Macpherson, B. Wilson, and U. Chajewska, *Identifying smokers with a medical extraction system.* J Am Med Inform Assoc, 2008. **15**(1): p. 36-9.
18. Manning CD, P.R., Schütze H, *Text classification & Naive Bayes*, in *Introduction to Information Retrieval*. 2008, Cambridge University Press: Cambridge, England. p. 253-289.
19. Alex Krizhevsky, S., Ilya, Hinton, Geoffrey E, *ImageNet Classification with Deep Convolutional Neural Networks*, in *Advances in Neural Information Processing Systems 25*, F.P.a.C.J.C.B.a.L.B.a.K.Q. Weinberger, Editor. 2012, Curran Associates, Inc. p. 1097-1105.
20. Yonghui Wu, M.S., Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, *Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation.* CoRR, 2016.
21. Mariusz Bojarski, D.D.T., Daniel Dworakowski, *End to End Learning for Self-Driving Cars.* CoRR, 2016.
22. Wayne Xiong, J.D., Xuedong Huang and *The Microsoft 2016 Conversational Speech Recognition System.* CoRR, 2016.
23. Hochreiter, S.S., Jürgen *Long Short-Term Memory.* Neural Computation, 1997. **9**(8): p. 1735–1780.
24. Junyoung Chung, C.G., KyungHyun Cho, Yoshua Bengio, *Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling.* 2014.
25. Dernoncourt, F., J.Y. Lee, O. Uzuner, and P. Szolovits, *De-identification of patient notes with recurrent neural networks.* J Am Med Inform Assoc, 2017. **24**(3): p. 596-606.
26. Mikolov T, S.I., Chen K. *Distributed representations of words and phrases and their compositionality*. in *Advances in Neural Information Processing Systems*. 2013.
27. Collobert R, W.J., Bottou L, *Natural language processing (almost) from scratch.* Journal of Machine Learning Research, 2011. **12**: p. 44.
28. Pennington J, S.R., Manning CD. . *GloVe: Global vectors for word representation*. in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 2014.
29. Mikolov T, K.a.M., Burget L, *Recurrent neural network based language model*, in *Interspeech*. 2013.
30. Socher R, P.A., Wu JY. *Recursive deep models for semantic compositionality over a sentiment treebank*. in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2013.
31. Y, K., *Convolutional neural networks for sentence classification.*, in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*. 2014. p. 1746–51.
32. Lee JY, D.F., *Sequential short-text classification with recurrent and convolutional neural networks*, in *Human Language Technologies 2016: The Conference of the North American Chapter of the Association for Computational Linguistics.* 2016. p. 515–20.
33. Jagannatha, A.N. and H. Yu, *Bidirectional RNN for Medical Event Detection in Electronic Health Records.* Proc Conf, 2016. **2016**: p. 473-482.
34. Lafferty, J.D.M., Andrew; Pereira, Fernando CN. *Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data*. in *Proceedings of the Eighteenth International Conference on Machine Learning, ICML '01.* 2001. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.
35. Gao, S., M.T. Young, J.X. Qiu, et al., *Hierarchical attention networks for information extraction from cancer pathology reports.* J Am Med Inform Assoc, 2017.
36. Zheng, W., H. Lin, L. Luo, et al., *An attention-based effective neural model for drug-drug interactions extraction.* BMC Bioinformatics, 2017. **18**(1): p. 445.

37. Liu, Z., M. Yang, X. Wang, et al., *Entity recognition from clinical texts via recurrent neural network.* BMC Med Inform Decis Mak, 2017. **17**(Suppl 2): p. 67.
38. Luo, Y., *Recurrent neural networks for classifying relations in clinical notes.* J Biomed Inform, 2017. **72**: p. 85-95.
39. Weng, W.H., K.B. Wagholikar, A.T. McCray, P. Szolovits, and H.C. Chueh, *Medical subdomain classification of clinical notes using a machine learning-based natural language processing approach.* BMC Med Inform Decis Mak, 2017. **17**(1): p. 155.
40. Team., R.C., *R: A language and environment for statistical computing.* 2017, R Foundation for Statistical Computing: Vienna, Austria.
41. others, F.C.a., *Keras*. 2015.
42. others, M.A.a., *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. 2015.
43. Nitish Srivastava, G.H., Alex Krizhevsky, Ilya Sutskever, Ruslan Salakhutdinov, *Dropout: A Simple Way to Prevent Neural Networks from Overfitting.* Journal of Machine Learning Research, 2014. **15**: p. 1929-1958.
44. Yarin Gal, Z.G., *A Theoretically Grounded Applicationof Dropout in Recurrent Neural Networks*, in *Neural Information Processing Systems*. 2016.
45. Tijmen Tieleman, G.H. *Lecture 6.5-RMSprop: Divide the gradient by a running average of its recent magnitude.* Neural Networks for Machine Learning 2012 [cited 2018 8/3/2018]; Available from: http://www.cs.toronto.edu/~tijmen/csc321/slides/lecture_slides_lec6.pdf.