applied sciences

MDPI

Article

Classification of Severe Maternal Morbidity from Electronic Health Records Written in Spanish Using Natural Language Processing

Ever A. Torres-Silva [1,†], Santiago Rúa [2,†], Andrés F. Giraldo-Forero [1,†], Maria C. Durango [3], José F. Flórez-Arango [4] and Andrés Orozco-Duque [3,*]

[1] Faculty of Engineering, Instituto Tecnológico Metropolitano, Medellín 050034, Colombia; evertorres254547@correo.itm.edu.co (E.A.T.-S.); felipegiraldo@itm.edu.co (A.F.G.-F.);
[2] School of Basic Sciences, Technologies and Engineering, Universidad Nacional Abierta y a Distancia, Bogota 111321, Colombia; santiago.rua@unad.edu.co
[3] Department of Applied Sciences, Instituto Tecnológico Metropolitano, Medellín 050034, Colombia; mariadurango254547@correo.itm.edu.co
[4] Population Health Sciences, Weill Cornell Medicine, New York, NY 10065, USA; jff4001@med.cornell.edu
* Correspondence: andresorozco4302@correo.itm.edu.co
† These authors contributed equally to this work.

Abstract: One stepping stone for reducing the maternal mortality is to identify severe maternal morbidity (SMM) using Electronic Health Records (EHRs). We aim to develop a pipeline to represent and classify the unstructured text of maternal progress notes in eight classes according to the silver labels defined by the ICD-10 codes associated with SMM. We preprocessed the text, removing protected health information (PHI) and reducing stop words. We built different pipelines to classify the SMM by the combination of six word-embeddings schemes, three different approaches for the representation of the documents (average, clustering, and principal component analysis), and five well-known machine learning classifiers. Additionally, we implemented an algorithm for typos and misspelling adjustment based on the Levenshtein distance to the Spanish Billion Word Corpus dictionary. We analyzed 43,529 documents constructed by an average of 4.15 progress notes from 22,937 patients. The pipeline with the best performance was the one that included Word2Vec, typos and spelling adjustment, document representation by PCA, and an SVM classifier. We found that it is possible to identify conditions such as miscarriage complication or hypertensive disorders from clinical notes written in Spanish, with a true positive rate higher than 0.85. This is the first approach to classify SMM from the unstructured text contained in the maternal EHRs, which can contribute to the solution of one of the most important public health problems in the world. Future works must test other representation and classification approaches to detect the risk of SMM.

Keywords: electronic health records; machine learning; maternal health; pregnancy complications; natural language processing; word-embedding

Citation: Torres-Silva, E.A.; Rúa, S.; Giraldo-Forero, A.F.; Durango, M.C.; Flórez-Arango, J.F.; Orozco-Duque, A. Classification of Severe Maternal Morbidity from Electronic Health Records Written in Spanish Using Natural Language Processing. *Appl. Sci.* **2023**, *13*, 10725. https://doi.org/10.3390/app131910725

Academic Editors: Selen Bozkurt and Suzanne Tamang

Received: 13 July 2023
Revised: 12 September 2023
Accepted: 15 September 2023
Published: 27 September 2023

1. Introduction

Reduction in th maternal mortality ratio (MMR) has been a priority issue on the global health agenda for decades. For instance, the World Health Organization (WHO) has proposed several strategies for ending preventable maternal mortality (EPMM) [1]. The strategies called for improvements along the continuum of care approach for pregnant women and focus on high-risk obstetrics diseases.

The WHO strategies for EPMM include a call for action to improve metrics, measurement systems, and data quality that account for all maternal and newborn deaths. Also, another objective laid out in these strategies is to address all causes of maternal mortality, reproductive and maternal morbidities, and related disabilities. One important stepping

stone for reducing the preventable maternal mortality is to identify and reduce severe maternal morbidity (SMM). Conditions such as obstetric hemorrhage, preeclampsia, sepsis, cardiovascular disorders, and placental disorders, among others [2], fall within the ambit of SMM. Frequently, identification and reporting of SMM are performed retrospectively, so corrective actions may be delayed. Therefore, there is an impending need to have a near real-time, or just-in-time approach to identify SMM cases.

In recent years, the adoption of electronic health records (EHRs) has increased exponentially in the world. In the EHRs, large numbers of registers are saved chronologically. Their database contains structured elements such as vital signs, drugs, etc. Such records also incorporate narratives written by health personnel including outpatient notes, nursing notes, discharge notes, and progress notes, among others, which usually contain unstructured text. It is estimated that 80% of the health data remain unstructured [3]. However, nowadays most of the clinical decisions are based on structured EHR fields. Furthermore, social, behavioral, and other determinants of health are commonly captured in narrative fields. Therefore, there is an opening alternative for natural language processing (NLP) to contribute to improve the causal and predictive models of diseases.

Although advances in NLP in maternal health has been reported, these are mostly made on applications and analyses of social networks [4–7]. A smaller proportion has been identified on Electronic Medical Records, but for other health conditions. English is the predominant language of corpus, and those existing studies are focused mainly on the mental health domain [8–10]. Both in English and in Spanish, there is a dearth of information on the role of NLP in the prediction of maternal morbidity and mortality. In this study, we hypothesize that NLP and Machine Learning (ML) may help identify SMM from clinical notes, which would increase the likelihood of a timely diagnosis and the subsequent clinical intervention, as well as simplify administrative tasks such as reporting and surveillance. Thus, our main goal is to develop a novel pipeline testing different word-embedding schemes, algorithms for document representation, and machine learning classifiers to identify SMM of pregnant women on progress notes in Spanish extracted from electronic health records.

To achieve our objective, we examine various word-embedding approaches, including Word2Vec, GloVe, and FastText, employing different preprocessing techniques and classification methods. Furthermore, we assess two pretrained language models, fine-tuning them to determine the relevance of employing word-embedding techniques in today's context. The main contributions of our paper are as follows:

- The first attempt to classify Severe Maternal Morbidity from unstructured notes.
- The use of pretrained word-embeddings schemes improved the classification performance.
- Typos and misspelling correction increased the performance of the pipelines in general.
- Hypertensive disorders and miscarriage complications show true positive rates over 83%.

2. Related Work

Word-embeddings are an important part of NLP and have become an inescapable choice for text representation for NLP tasks. The representations based on word frequency and others usually are high dimensional, ignore the order of the text, are sparse, etc.

To deal with the problems of word frequency representation, different approaches were tested. In a previous study, authors used a bag of words with word weighting by TF-IDF scores. They tested different classifiers including support vector machines (SVM), random forest (RF), and extreme gradient boosting (XGB). They tested on 2-, 3-, or 7-stage cancer labels with an F1-score from 0.80 to 0.99 [11]. Similarly, another group used skip-gram and paragraph vectors-distributed bag of words (PV-DBOW) with multiple discriminant analysis (MDA) to generate document embeddings. Regarding multiclass classification (five classes), they obtained an F1-score ranging from 0.68 to 0.97 using an extreme learning machine (ELM) [12].

The use of word-embeddings for classification is common in the health sector to classify different diseases and other challenges. For instance, in the mental field [13], researchers

have used weighted document vectors as a combination of TF-IDF with Word2Vec for challenging behavior classification. They report an accuracy of 84.3–98.5% in a binary class between challenging behaviors in Autism Spectrum Disorder (ASD) using SVM. In [14], researchers focused on comparing different word representations, including TF-IDF, Word2Vec, batch-Word2Vec, and Doc2Vec. The objective was to classify cardiovascular diseases according to eight ICD codes. They obtained a better result in representation based on word frequency. In [15], the authors compared three different representations of the EHR notes, with a convolutional neural network classifier to predict a visual prognosis. They obtained an F1-score on this model of 67%.

Most of the previous works focus on EHRs in English. Nevertheless, there are some reports using NLP on EHRs in other languages. For instance, in [16], researchers developed different pipelines for the identification of adverse drug reactions (ADRs) in Dutch clinical notes. Likewise, there are a few studies using Spanish that use EHR notes to classify different health conditions. For example, in [17], they classified cancer diagnoses using different word-embedding techniques and machine learning. The results for the binary classification was an F1-score of 98%. In [18], they developed a predictive model for early and late progression to first-line treatment of HR+/HER2-negative metastatic breast cancer. According to their results, the best NLP-based model achieved an AUC of 0.752. In addition, we identified works focused on other problems such as the ICD-10 coding in Spanish, which included word-embeddings like the one presented in [19], or using a contextualized language model (BioBERT) with rule-based approaches [20].

To our best knowledge, there are no studies to classify SMM using NLP neither in English nor Spanish.

3. Methods

In this study, we adopted a novel approach, exploring whether it is possible to classify SMM by building different pipelines of NLP methods combined with machine learning classifiers applied on the progress notes of the episodes before the mother is discharged from the hospital. For the purpose of this work, a clinical episode starts with the admission of a pregnant woman into the hospital by outpatient or emergency consultation. In this time frame, we can find both ambulatory encounters and inpatient encounters. The episode concludes when the women have been discharged from the hospital or have a delivery.

3.1. Severe Maternal Morbidity (SMM)

There is not a single, comprehensive definition of SMM; it is also recognized as severe acute maternal morbidity [21] and also known as "near miss" [2]. But, in general, it is considered a complication that puts the life of the pregnant woman at risk and requires urgent medical intervention [22]. Colombia's National Epidemiological Surveillance System (SIVIGILA) uses ICD-10 codes to report SMM. They use eight groups with the purpose of following the conditions that require more action from the government [23]. Table 1 presents the definitions of the groups according to the ICD-10 aggrupation.

These events are reported to the government according to the data recorded on the EHRs. For our purposes, all ICD-10 classifications in discharge notes that do not fit in the grouping system of SIVIGILA were considered to be without SMM (WS).

Table 1. Definition of SMM Groups and number of ICD-10 codes per group

Acronym	SMM Group	Number of ICD-10 Codes	Example of ICD-10 Codes
WS	Without SMM	–	–
OC	Other causes	590	O24.4, O25, O26.1, O34.0, O43.0, T36, V09...
HC	Hemorrhagic complications	46	O08.1, O20, O44.1, O45, O46, O72, N93.9...
MC	Miscarriage complications	65	O02.1, O03, O04, O05, O06, O07, O08.2...

Table 1. *Cont.*

Acronym	SMM Group	Number of ICD-10 Codes	Example of ICD-10 Codes
HD	Hypertensive disorders	30	O10, O11, O12, O13, O14, O15, O16, I10, I13.9...
OS	Obstetric sepsis	28	O08.0, O02.3, O85, O91.0, A54.2, A56.0, ...
PI	Complication of pre-existing illness	147	O24.0, O98, O99, C14, D59, E02, G40, I25, N03
NS	Non-Obstetrics sepsis	94 + 48 (Pulmonary sepsis)	A02.9, A03.0, A04.0, A05.0, A06.0, G00.8, K67.0...

3.2. Data Source and Data Quality

Clinica Universitaria Bolivariana (CUB) is a university general hospital with an important focus in obstetrics. CUB has provided more than 80,000 deliveries in the city of Medellín, with a mean of 5200 deliveries per year which was approximately 13.5% of the city's total in 2020. The Hospital Information System and EHR system used by the clinic is Servinte Clinical Suite (Carvajal SA, Medellín, Colombia, v1.3, 2019). According with de Data Dictionary, we queried the EHR system to extract clinical records of patients who gave birth in the institution during the period 2015 to 2019. CUB did not provide demographic data; due to privacy constraints, the dataset was limited to extracted unstructured patient notes. The data were obtained using Oracle SQL Developer and stored in binary format for processing and analysis. We used tables containing information about progress notes corresponding on SOAP (Subjective-Objective-Analysis-Plan) forms performed by clinicians in each encounter. Patients' data were assigned an internal identification number to match the progress notes and their corresponding ICD-10 code at discharge.

To assess the document quality, we conducted an Exploratory Data Analysis (EDA) on the progress notes. Different aspects were examined, including note length, vocabulary size, word frequency distribution, structural patterns, sparsity, punctuation usage, and formatting conventions. Subsequently, we decide the steps for preprocessing.

The content of the progress notes is separated into various fields according to the design of the form in the EHR system. As a consequence, we found duplicate information in the fields of the same progress notes. Then, before processing the text we eliminated the duplicated data and joined all the content of each field into a single document. Lastly, episodes without progress notes or notes with fewer than 10 words were discarded.

To assess the performance of our model, we divided the dataset into two independent subsets: a training set and a test set. Specifically, we allocated 20% of the total dataset as the test set. To ensure that both sets maintained a proportional representation of classes, we employed a stratified sampling. Stratification helps preserve the distribution of target labels in each subset; this methodology ensures that our model is subjected to a representative test set.

3.3. Preprocessing

To comply with Colombia's rules (Law 1581) about Protected Health Information (PHI) [24], we applied a basic ruled-based approach to identify first and last names in the dataset. To accomplish this, we used a library in Python developed by Colombia's National Planning Department (DNP), named *Contexto* [25], that contains lists of Colombian first and last names. This list includes some names that could have an important meaning in our context, for instance: *bueno* (good), *bien* (well), *rojas* (reds), *cama* (bed), *rojo* (red), *cesárea* (c-section), *rosa* (pink), *dolor* (pain), *olores* (pains), and *blanco* (white), etc. Those words were excluded from the dictionary.

We continued the preprocessing of the text with typical sequential steps that contained the following methods: *tokenization, remove numbers, lower case, remove punctuation, remove stop words, and remove accent marks*. We stored these tokens in a list and used a matching dictionary to change the tokens that corresponded to PHI instances using the label $<NOMBRE>$ or $<APELLIDO>$ as appropriate. Finally, we reconstructed the text with all the tokens. In our rules-based approach to de-identification, we did not take into account the numbers of IDs, telephone numbers, and addresses, because they are removed in the sequential steps above.

For the stop words removal step, we implemented two schemes. First, we used the set of Spanish stop words from NLTK for default. Second, we customized a dictionary of stop words including the most common words detected in our dataset without an important meaning in the medical context by a data exploration analysis. A total of 2000 tokens were stored in an .xls file. To build the final customized dictionary of customized stop words, a health informatics physician analyzed the file and selected 168 tokens to create the customized stop words list. Some of the tokens considered as customized stop words were *paciente (patient), años (years), embarazo (pregnant), horas (hours), refiere (refers), eco (echo), cada (each), fetal (fetal), sem (contraction of weeks), si (yes), mg (mg), extremidades (extrimities), semana (week), momento (the time), dia (day), madre (mother), instrucciones (instructions), eps (insurance company), materna (maternal), ecografía (echography), consulta (consultation), gestacional (gestational), bebé (baby), am (at morning), minutos (minutes), residente (resident), clínica (clinic), encuentra (find), materno (maternal), derecha (righ), pone (to put), debe (should), entiende (understand), sexo (gender), pediatria (pediatrician), médico (doctor), tarde (late), gestación (gestation), xmin (per minute)*. Note that, for instance, the maternal or medical words, appear many times in the documents; however, they were not important to differentiate between patients with or without SMM. The repository can be found at the following link: https://github.com/sruap1214/SMM_NLP, (accessed on 14 August 2023).

3.4. Word-Embedding Schemes

After the text preprocessing stages, we used word-embedding to generate a vector representation of each word. We explored six different word-embedding schemes with the aim of evaluating which was the best scheme to represent the episodes in the EHRs.

First, we trained a Word2Vec model using a continuous bag of words (CBOW) to obtain a vector representation for each word in the dictionary constructed by all the words contained in our EHR database. We trained the model from scratch using a neural network with 300 neurons in the hidden layer. Once the data were preprocessed, we created our own word-embeddings representation using the training data with 300 dimensions and a context windows of 5 tokens. To differentiate if there is a change in the sense of the paragraph, we added the special token <SEP> after each period or before a new line according to [26]. At the end of the training stage, we obtained 147,272 vectors with 300 dimensions.

Second, we used a pretrained Word2Vec model in the Spanish Billion Word Corpus (SBWC) [27]. The pretrained word embedding models employed in this study were chosen due to their pretraining being carried out in Spanish. In particular, this model contains 1,000,653 vectors with $m = 300$, where m is the vector size. Taking into account that the progress notes analyzed from EHRs contain a huge amount of words with typos and misspellings, there is a high probability that those words will not be found in the dictionary of the SBWC. We implemented a method to be able to assign a vector value to words with typos or misspellings based on searching the closest word in the dictionary. This adjustment was implemented by computing the Levenshtein distance between the misspelling or typo words and the vocabulary of the SBWC. The minimum Levenshtein distance was found using a searching technique based on a tree data structure, which is faster than a traditional search. The Word2Vec model was implemented with and without the misspelling and typos adjustment.

To compare different word-embedding approaches, we implemented three additional embeddings schemes. First, the GloVe Vectors (GloVe) embeddings were used with a model pretrained from the SBWC. Second, the GloVe was used with and without the misspelling and typos adjustment (same algorithm as in Word2Vec). And third, we implemented FastText using a model pretrained in Common Crawl and Wikipedia. The FastText model was trained using CBOW, with dimension 300, a character n-grams of length 5, a window of size 5, and 10 negatives. We did not use the misspelling adjustment in FastText because it works with subword information.

The six embedding schemes used in this work are listed below:

- Word2Vec from scratch trained in our dataset.

- Pretrained Word2Vec.
- Pretrained Word2Vec with typos and spelling correction.
- Pretrained GloVe.
- Pretrained GloVe with typos and spelling correction.
- FastText.

3.5. Features Extraction

Word-embeddings can be faster to implement and provide good results alternatively to feature engineering for classification and prediction tasks in the clinical domain [28]. The use of word-embeddings assigns a vector for each word. However, all the documents have different length. Therefore, it is necessary to have a mechanism that allows one to have a combination of the representations of each document. In this section, we describe three different mechanisms to obtain only one feature vector that represents all the information of each document under analysis. Each document representation was tested with the six word-embeddings schemes described previously in Section 3.4.

- *Average*: In this approach, we represent each document by averaging vectors of all the words contain in each episode, similar to [29].
- *Clustering*: Unlike the average approach, in the present technique, the idea is to represent episodes using word clusters. For that purpose, we followed the bag-of-centroids approach presented in [30]. First, we defined the number of clusters to use due to the proportion of word vectors in the documents. We selected the number of clusters k in such a way that on average the clusters had 10 words. As a consequence, the value of k must be set to 1155; this value is the result of dividing the total number of words between the number of words per cluster. Second, we applied the k-means algorithm to obtain the centroids of each cluster. Third, we represented each word using an embedding model and obtained the cluster to which its word vector belongs, increasing its count by one. Finally, we will obtain a k-dimensional vector, where each element represents one of the clusters. The corresponding value of each element is the number of words in the record belonging to this cluster.
- *Principal Component Analysis (PCA)*: As an alternative approach, we proposed using the principal component analysis (PCA) method, which seeks to rotate the space in the direction where the highest variance is presented. We represented each document as the first eigenvector resulting from applying the PCA method on the set of words that make up a document from an episode. In this way, regardless of the number of words in the document, we can guarantee that the resulting representation vector will have a fixed dimension of 300 features.

3.6. Classification Schemes

We defined different pipelines by the combination of the six different word-embedding approaches described in Section 3.4, the three representation methods defined in Section 3.5, and the following five well-known classifiers: logistic regression (LR), multilayer perceptron (MLP), support vector machine (SVM), random forest (RF), and k nearest neighbor (kNN). The different pipelines were tested using macro average F1-score (F1-macro) with a 5-fold cross-validation scheme.

We used the implementation from the Python library `scikit-learn` [31] for training and testing each pipeline. The parameters of the classifier were instantiated following a grid search scheme. In particular, for the MLP various combinations of neurons organized in 1 to 3 hidden layers were tested using the Adam solver and the values $10^{-4}, 5 \times 10^{-2}$ for the α. For the SVM, the polynomial, radial basis, and sigmoid kernels were tested with their default values together with an exponential growth sequence $10^{-1}, 10^{0}, \ldots, 10^{3}$ for the penalty. For the logistic regression, we varied the penalty term C $(100, 10, 1, 0.01)$. In kNN we assessed the values $1, 3, \ldots, 9$ for the number of nearest neighbors k. Finally, in RF we varied the number of estimators $(100, 200, 300)$, the minimum number of samples in a leaf node $(1, 2, 4)$, and the minimum number of samples required to split a node $(2, 10, 20)$.

Our dataset is imbalanced, mainly associated with the class without SMM (WS). To limit the class imbalance impact, we undersampled the majority class (WS) over the training set. Instead of using undersampling in the rest of the dataset in a one-vs-rest scheme, we randomly removed samples from the majority class so that the size of the majority class had the same number of samples as the class with the second-highest number of samples. We chose this method because the classifiers do not always use the one-vs-rest strategy to deal with multiclass classification problems as is the case for MLP and kNN.

For surveillance of SMM, both precision and recall are important. A high precision allows the system to assign the correct group to SMM; meanwhile, a high recall is intended to not miss a correct classification when the episode corresponds to a real condition associated with SMM. For that reason, we used the F1-macro as an evaluation measure, because it includes the precision and recall metrics and considers each class in the model.

3.7. Language Model Schemes

In addition to the representation schemes and classification models discussed in Sections 3.5 and 3.6, we assessed the performance of two language models for the purpose of comparison with the proposed classification schemes using word-embedding.

- RoBERTa [32] is a transformer-based masked language model for the Spanish language. It has been pretrained using a Spanish corpus with a total of 570 GB of data.
- Longformer-es [33] is the Longformer version of the RoBERTa. Longformer-es employs a blend of sliding window (local) attention and global attention mechanisms that scales linearly with sequence length that allows it to process documents with thousands of tokens.

Our experiments are performed on a Nvidia Quadro RTX5000 GPU with 16 GB, for which a pretraining process was carried out involving 15 epochs. We trained the model with a learning rate of 10^{-5}, dropout rate of 0.1, and AdamW optimizer for both models; the other parameters are set by default. The models are available at hugging face https://huggingface.co/PlanTL-GOB-ES/roberta-base-bne, (accessed on 20 August 2023) https://huggingface.co/PlanTL-GOB-ES/longformer-base-4096-bne-es, (accessed on 20 August 2023). Given the length of the clinical notes, we chose to use Longformer-es, which supports a maximum input length of up to 4096 tokens. For batch processing, we utilized the *padding longest* option, guaranteeing a uniform sequence length within each batch, aligning with the longest sequence in that batch. The batch size was configured as 6 for both models, and the remaining parameters retained their default values.

We fine-tuned the language models using the same preprocessing steps described in Section 3.3. We performed the pretraining and evaluation of the transformer models using the documents without typos or misspelling adjustment because, like in FastText, those models works with subword information. In addition, the ability of BERT-based models to consider context from both directions helps it capture the meaning of words, even in the presence of typos or misspellings.

4. Results

4.1. Dataset Characterization

We identified 22,937 patients from the EHRs, most of them with more than one progress note. We built a dataset by grouping all the progress notes per episode into one document. A total of 43,529 documents were built. We assigned the ICD-10 discharge code to each document. Figure A1 shows the distribution of ICD-10 codes in our dataset by its chapters.

These documents have a mean of 4.15 progress notes. Table 2 presents the number of documents per class and the mean progress notes in such documents. Due to the low frequency of pulmonary sepsis after validating with physicians, we merged this class into non-obstetrics sepsis for analysis.

Despite the query being targeted for women that have a delivery in the study timeframe, we found ICD-10 codes associated to the episodes in almost all chapters of ICD-10. As expected, the most frequent chapter was the XV, corresponding to a pregnancy, child-

birth, and puerperium. In this chapter, the most frequent code found was the single spontaneous delivery followed by single delivery by cesarean section and false labor. The second most frequent chapter was XXI, which is related to the factors influencing health status and contact with health services. In this chapter, the most frequent code is related to the supervision of normal pregnancy. The third most frequent chapter was XVIII, which is related to symptoms, signs, and abnormal clinical and laboratory findings; the most frequent codes in this chapter were pelvic pain and headache. Lastly, the fourth most frequent chapter was the XIV, associated to diseases of the genitourinary system, where the most frequent code were other disorders of the urinary system.

Table 2. Number of documents per SMM class and mean progress notes.

Acronym	SMM Group	Number of Documents	Proportion of Documents	Mean of Progress Notes
WS	Without SMM	34,117	0.784	4.01 ± 6.28
OC	Other causes	2345	0.054	5.54 ± 9.39
HC	Hemorrhagic complications	1922	0.044	3.19 ± 4.90
MC	Miscarriage complications	1919	0.044	2.39 ± 3.96
HD	Hypertensive disorders	1538	0.035	6.52 ± 9.37
OS	Obstetric sepsis	1187	0.027	5.99 ± 7.38
PI	Complication of pre-existing illness	249	0.006	6.97 ± 8.91
NS	Non-Obstetrics sepsis	228	0.005	4.69 ± 14.50

We analyzed the number of tokens for each document at two moments. First, after removing the common stop words, and second, after the elimination of customized stop words identified by the health informatician of the research team. Additionally, MC and NS were the classes most affected by the removal of stop words. Without common stop words the average number of tokens per document is 431. After removing the customized stop words, the average tokens per documents decreases to 64. Figure 1 shows the distribution of the number of tokens differentiated by SMM classes. Importantly, the documents left empty after removing stop words were discarded.

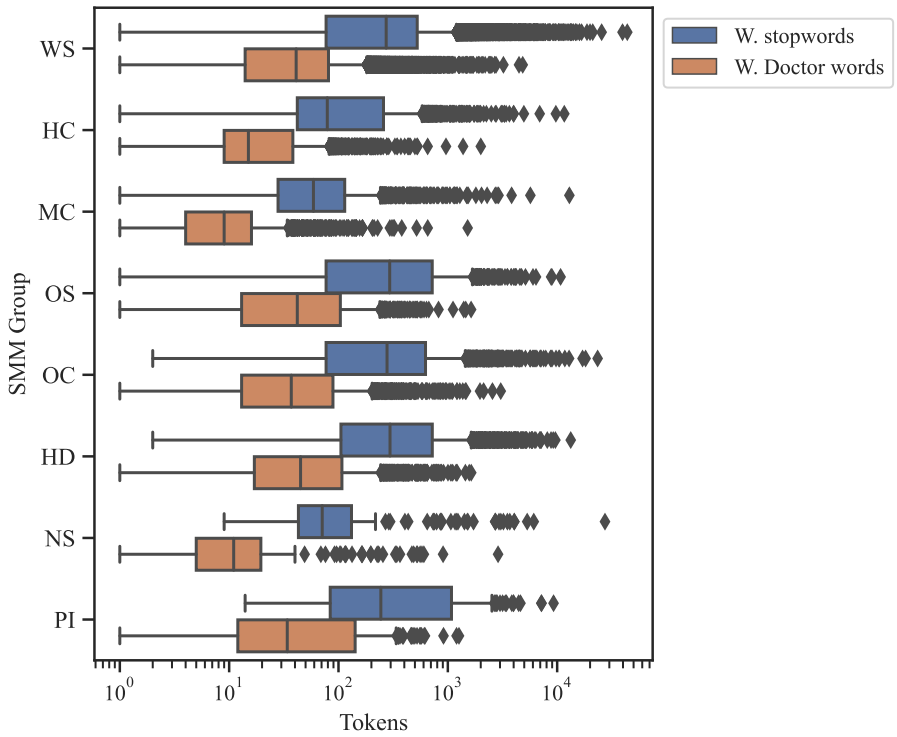


Figure 1. Tokens per document with and without stop words.

When compared to the GloVe and Word2Vec vocabularies, the percentage of typos and misspelled words in our entire dataset was between 10.85% and 14.13%. Also, the average percentage of missing words per document was 10.58% in GloVe and 13.77% in Word2Vec. We consider that these spelling errors occur due to the time assigned to physicians to fill out the EHR and the lack of an autocorrection functionality in the EHR system. Our numbers were similar with the results of Ruch [34] in follow-up notes.

4.2. Classification

Table 3 presents the detailed classification results of SMM groups in the validation sets defined in the cross-validation step, comparing the different pipelines proposed. In general, all the pipelines have similar performances in the validation set. The best performance in the validation set was achieved using clustering representation followed by average. The best performance was found by MLP with the pretrained Word2Vec with the typos and spelling adjustment. The poorest performance was obtained by the LR classifier using the Word2Vec trained in our dataset with the average representation. In general, the worst classifier using a different representation and characterization approach was kNN.

Table 3. Mean and standard deviation for different models in k-fold cross-validation

Algorithm	**Word-Embedding**	**F1-Macro**				
		LR [%]	**MLP [%]**	**SVM [%]**	**RF [%]**	**k-NN [%]**
Average	FastText	65.71 ± 0.93	65.72 ± 1.33	66.36 ± 0.94	50.57 ± 0.78	50.91 ± 1.28
	GloVe No Correction	66.15 ± 0.67	66.04 ± 1.24	67.42 ± 1.41	51.80 ± 1.02	57.13 ± 1.44
	W2V No Correction	64.27 ± 0.93	66.08 ± 0.89	66.60 ± 0.99	50.69 ± 0.60	56.37 ± 0.63
	GloVe Correction	68.12 ± 1.38	69.13 ± 0.77	69.46 ± 0.87	54.01 ± 0.69	59.69 ± 1.05
	W2V Correction	62.70 ± 1.97	67.43 ± 0.89	67.39 ± 0.72	51.55 ± 0.70	57.29 ± 1.49
	W2V Trained	24.69 ± 0.54	51.49 ± 0.82	65.91 ± 0.82	51.23 ± 0.67	54.81 ± 1.08
Clustering	FastText	64.59 ± 0.82	69.37 ± 1.32	65.69 ± 1.65	60.46 ± 2.08	54.86 ± 1.50
	GloVe No Correction	62.81 ± 0.68	67.11 ± 1.05	63.99 ± 0.48	59.21 ± 1.59	51.71 ± 1.78
	W2V No Correction	63.66 ± 1.03	68.48 ± 0.87	65.49 ± 0.90	62.43 ± 1.54	52.10 ± 1.08
	GloVe Correction	66.59 ± 0.95	70.36 ± 0.97	67.73 ± 1.69	62.44 ± 1.66	54.78 ± 1.80
	W2V Correction	66.55 ± 0.80	70.54 ± 1.92	67.14 ± 0.99	62.31 ± 1.42	55.67 ± 2.05
	W2V Trained	64.19 ± 1.30	68.52 ± 0.63	65.74 ± 0.70	59.63 ± 0.62	54.65 ± 0.32
PCA	FastText	47.64 ± 0.53	52.30 ± 1.26	50.44 ± 2.32	37.96 ± 1.03	34.11 ± 1.59
	GloVe No Correction	55.86 ± 1.12	61.83 ± 2.45	63.18 ± 1.87	48.07 ± 0.95	50.00 ± 1.51
	W2V No Correction	52.44 ± 1.73	61.25 ± 1.63	63.25 ± 1.13	50.17 ± 1.08	52.56 ± 1.29
	GloVe Correction	55.71 ± 1.40	61.70 ± 2.23	63.32 ± 1.62	49.80 ± 1.41	53.11 ± 1.41
	W2V Correction	64.26 ± 1.03	66.88 ± 0.76	67.54 ± 0.93	55.63 ± 0.97	59.89 ± 2.56
	W2V Trained	38.62 ± 0.62	52.88 ± 2.31	53.31 ± 1.36	45.79 ± 0.42	48.08 ± 0.67

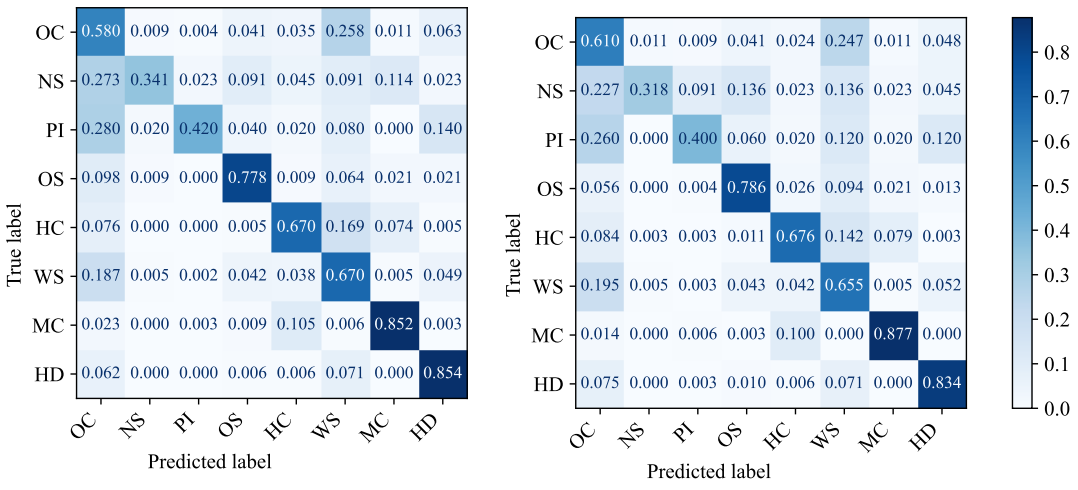
Table 4 shows the F1-macro of the same pipelines used in the training set applied to the test set. The test set was not under-sampled and reflects the reality of the studied population. Consequently, we experienced considerable reduction in the performance in all pipelines as expected because the groups were not balanced in the test set (in comparison with the training set where the groups were balanced). However, due to there being a multi-class problem with eight classes, these results are promising, and each class could be distinguished from the rest.

The pipeline with better performance was the one that included Word2Vec, typos and misspelling adjustment, PCA, and the SVM classifier. The second was the same pipeline but with an MLP classifier, which achieved a similar F1-macro. The third was the one that included GloVe, typos and misspelling adjustment, average representation , and the LR classifier. The most effective algorithm for obtaining a single feature vector was the clustering, showing a possible overfitting. This effect was evidenced in the k-NN classifier by [35] showing a reduction in the performance on imbalanced data.

In addition, thetwo confusion matrices with the best performances on the algorithm of PCA and average representation of the document in the test set are shown in Figure 2. These confusion matrices provide a breakdown of the actual groups, which are listed vertically, and the predicted groups, which are listed horizontally. The normalized outcome of the correct prediction is listed diagonally. We noticed the best performances in the groups of MC and HD, and the worst performances were found in the categories with the least number of documents, such as NS and PI. In addition, most of miss-classification occurs in the groups of OC and WS.

Table 4. Comparison of different models applied in the test set (20% of documents)

Algorithm	Word-Embedding	F1-Score Macro				
		LR [%]	MLP [%]	SVM [%]	RF [%]	k-NN [%]
Average	FastText	49.01	44.32	48.71	37.42	38.08
	GloVe No Correction	48.68	47.21	50.03	39.81	42.41
	W2V No Correction	49.08	49.36	49.23	38.75	41.30
	GloVe Correction	51.18	48.88	50.76	42.98	43.42
	W2V Correction	47.97	50.37	50.19	39.81	42.16
	W2V Trained	26.10	40.17	47.84	40.17	42.14
Clustering	FastText	44.89	49.10	48.26	43.59	40.54
	GloVe No Correction	45.06	47.94	48.36	43.09	37.20
	W2V No Correction	46.32	47.46	47.06	47.07	38.05
	GloVe Correction	46.93	50.06	49.29	46.44	39.95
	W2V Correction	48.16	49.63	50.02	45.03	41.66
	W2V Trained	46.75	49.79	48.61	44.51	40.31
PCA	FastText	35.14	34.92	36.38	28.40	23.59
	GloVe No Correction	40.44	44.92	48.32	37.35	37.42
	W2V No Correction	36.73	48.44	47.73	40.38	39.03
	GloVe Correction	42.93	48.31	50.03	38.86	41.25
	W2V Correction	48.48	51.47	52.54	43.39	42.88
	W2V Trained	27.99	38.91	40.40	35.85	36.33



(**a**) (**b**)

Figure 2. Confusion matrix normalized for the best models in the test set. (**a**) SVM—Word2Vec Correction—PCA. (**b**) LogReg—GloVe Correction—Average.

To offer a more comprehensive assessment of the model's clinical utility [36], Table 5 presents the calculated positive predictive values (PPVs) and true positive rates (TPRs) for the best-performing models on the test set. Our emphasis on these metrics is due to the model's role as a potential screening solution. We have chosen not to report metrics dependent on true negative values. This decision is based on the substantial sample imbalance,

which results in markedly higher true negative values compared to false negative values. We can highlight that, further to WS, the classes MC and HD have the best performance in terms of TPR.

Table 5. True negative (TN), false positive (FP), false negative (FN), true positive (TP), true positive rate (TPR), and positive predictive value (PPV) for the best models in the test set.

SMM Group	SVM—Word2Vec Correction—PCA						LogReg—GloVe Correction—Average					
	TN	FP	FN	TP	Sens.	PPV	TN	FP	FN	TP	Sens.	PPV
WS	1588	228	2216	4507	0.670	0.952	1594	222	2322	4401	0.654	0.952
OC	6715	1362	194	268	0.580	0.164	6672	1405	180	282	0.610	0.167
HC	7856	316	121	246	0.670	0.438	7834	338	119	248	0.676	0.423
MC	8111	77	52	299	0.852	0.795	8115	73	43	308	0.877	0.808
HD	7856	375	45	263	0.854	0.412	7846	385	51	257	0.834	0.400
OS	7988	317	52	182	0.777	0.365	7981	324	50	184	0.786	0.363
PI	8469	20	29	21	0.420	0.512	8453	36	30	20	0.400	0.357
NS	8452	43	29	15	0.341	0.259	8453	42	30	14	0.318	0.250

4.3. Language Model Results

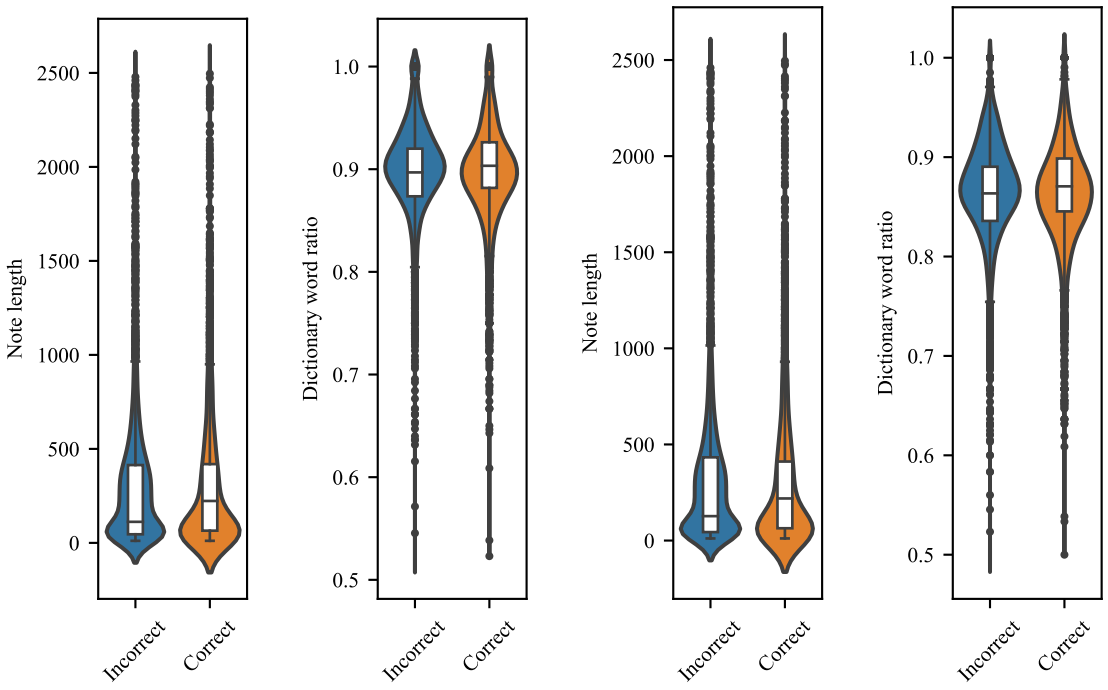
We assessed the RoBERTa and LONGFORMER-ES models applyied to the test set with 20% of the documents. For comparative purposes, we used the identical test database employed for evaluating the different pipelines proposed in the previous sections. Taking into account the distribution of the number of tokens for each clinical note presented in Figure 1, it can be seen that most clinical records have fewer than 1000 tokens. We configured the maximum input length to be 512 tokens for RoBERTa and 1024 tokens for LONGFORMER-ES.

We conducted evaluations for the initial 15 epochs in both models. However, we specifically present findings from the epoch with the lowest training loss value, along with its corresponding F1-macro performance. Here, epoch 10 in the LONGFORMER-ES model obtained 0.485, while epoch 12 in the RoBERTa model obtained 0.461.

4.4. Error Analysis

In order to identify the dependencies of the documents classification with their length and their ratio of words present in the GloVe's and Wor2Vec's dictionaries, we applied a set of statistical tests over two groups of data, the well-classified documents and the miss-classified documents. For calculating the ratio of words present in the dictionary for each document, we count words belonging to the dictionary and divide them by the length of the note. The average word ratio in the dictionary for the whole dataset was $89.41 \pm 4.85\%$. We used a Mann–Whitney–Wilcoxon test, given the nature of the data, where each electronic note is independent one each other. The *t*-test was discarded because the length of the documents and words ratio in dictionary did not have a normal distribution. The Mann–Whitney–Wilcoxon (MWW) test is a non-parametric test that contrasts whether two samples come from equidistributed populations.

For the error analysis, we used the results given by the two pipelines with top performance: (1) the one that included the LR classifier and the average of GloVe vector for document representation, and (2) the one that included the SVM classifier and PCA computed on Wor2Vec characterization. Figure 3a shows that the distribution of documents that were correctly classified present higher medians than incorrect ones; the MWW test confirms that there are significant differences between both groups with a p-value of 2.47×10^{-10}. Additionally, the dictionary word ratio presents a significant difference between documents classified as correct and incorrect (p-value = 2.725×10^{-16}). Similar behavior was observed in pipeline (2) where the documents of a correctly classified document tended to contain more words (p-value = 4.25×10^{-7}) and tended to have a higher dictionary word ratio (p-value = 2.2×10^{-16}).



(**a**) LogReg-GloVe-Average (**b**) SVM-W2VCor-PCA

Figure 3. Boxplot and violin plots that compare the length and ratio of words present in the word-embedding dictionary of two group of electronic notes: blue group corresponds to notes correctly predicted and orange group to documents incorrectly predicted (**a**) for pipeline (1), and (**b**) for pipeline (2).

5. Discussion

Despite the fact that SMM is a major public health issue worldwide, there is scarce evidence on the use of NLP to assist in SMM classification. In fact, to the best of our knowledge, this is the first attempt to classify SMM using NLP on Spanish EHRs.

Since the number of studies related to NLP of medical records in Spanish is limited, we tested different approaches of feature representations of clinical notes with different conventional classifiers. This is a necessary step to establish a baseline that allows us to propose new strategies in the future.

Some classes in our dataset have a smaller amount of tokens, such as MC and NS. However, it is not possible to differentiate any class by the number of tokens in the documents. There is a great variability in the number of tokens as in the number of notes per episode, since no treatment was carried out to eliminate outliers.

Some authors claim that the used of a word-embedding training in a corpus on the specific domain can achieve better performance in domain-specific tasks. However, our results showed that the Word2Vec model trained in our dataset from scratch exhibited the worst performance in comparison with the used of pretrained word-embedding models. This result can be due to the insufficient number of documents in our dataset, taking in account that the pretrained models used datasets with billions of words. Our experiments are in line with [37], indicating that a biomedical domain corpus does not necessarily have better performance than a general domain corpus. However, there is a need for more word-embedding models applied to low-resource languages and domains[38], including clinical Spanish.

The preprocessing of the text can play an important role in the performance of the classifiers; it is a little-explored field and it depends on the domain [39]. The high frequency of typos and misspellings in the EHRs could be improved with corrections; however, there is a risk of changing the meaning of some words, due to the abundant lexical variants. Our

results show that adjusting of typography and spelling, in general, increased the models' performances. Testing new manners of correction of typos and spellings could have a better improvement in the classification task [40]

Regarding the document representation techniques, PCA exhibited the best performance and clustering representation the worst performance in the test set. Although average representation had a performance similar to that of PCA in the test set, it performed better than PCA in the validation set. PCA and average give a representation by a vector that summarized the whole document, while clustering gives a representation based on word counting, which gives a vector with many zeros. This behavior can affect the performance of the classifiers. This supports the conclusions made by [41], which indicate that the text classification algorithms are more efficient with a better understanding of feature extraction methods and that document cleaning could help the accuracy and robustness.

As for the performance of the classifiers, the best classifier achieves a macro F1-score of 70.54% in the validation set (which is balanced) and the macro F1-score falls to 52.54 in the best pipeline in the test set (which is imbalanced). The imbalance of the test set has an important role and affects the performance; however, in a real context, we have to deal with imbalanced classes because most of the encounters are with patients with some alarm signs but not with a severe outcome.

We implemented two classifiers based on pretrained language models to have a comparison with the classifiers based on features extracted from word-embedding schemes. The results show that the Longformer model achieves better results than RoBERTa, which could be due to its handling of a larger number of tokens, causing RoBERTa to truncate some notes. However, it is evident that neither of the two fine-tuned models surpassed the results obtained with the word-embedding schemes. This outcome is support by [42], who found that the performance of binary classification for ICD-9 codes exhibited similarity when employing word-embedding schemes compared to transformer-based models. These results might stem from the fact that for particular tasks that do not require extensive context, such as text classification, the complexity of language models might not be necessary. In such cases, word-embeddings can provide more efficient solutions. Future work should be aimed at testing different large language models with a larger dataset in the fitting process, especially in those classes with fewer samples.

According with the confusion matrix, the most false positives are present in the other causes (OC) class. The reason for this is that this group incorporates a large number of ICD-10 codes distributed in the different chapters. A similar situation occurs in the WS category, since all the ICD-10s that did not fit into the SMM grouping were aggregated here. Note that WS and OC are classes that comprise a large amount of ICD-10 codes. In contrast, more defined categories such as hypertensive disorders (HD), miscarriage complications (MC), and obstetric sepsis (OS) show a performance of around 80% of true positive rates. Taking in account that the objective is to detect groups with a high risk of SMM, this is a promising result.

As shown in Table 5, the results suggest that our premise and model proposed can be used as screening tool for classes such as HD and MC. These classes exhibit the highest TPR. Nevertheless, it is worth noting that only the MC class displays higher values for both TPR and PPV. For instance, the performance of class HD, despite its strong TPR, has a low PPV. Even so, the model was effective at correctly identifying instances with performance above 85% in this class. This means that while it correctly identifies instances of the HD class, it also generates a notable number of false positives for this class, primarily from the without-SMM class (WS). This disparity arises from the substantial class imbalance, where a considerable number of samples belong to WS. As a result, even though the model's performance in these classes is suitable, there is a relatively higher count of false negatives in comparison to true positives. Regardless, in a potential future application of these models as screening tools for generating alerts related to maternal mortality risk, emphasizing TPR remains crucial. The issue of potential false positives can be addressed through a secondary evaluation conducted by a specialist.

Limitations

The use of proprietary databases makes it difficult to carry out external validations and requires more time to understand the infrastructure and structure of the records. Furthermore, the schemes of the health record databases in a Colombian context focus the most on administrative topics and usually lack schemes for research. For that reason, we needed a long time to understand the native structure and extract, transform, and load the data required for this study. This could be more efficient with the standardization of conditions and the analysis of data through common data models such as OMOP.

There was a high frequency of typos and misspellings in the narratives that were treated compared with the vocabulary of a corpus not specialized in the biomedical or clinical domain. This required a better approach to treat these special tokens and be careful with the impact that removing words could have, or changing the meaning of the sentence [43].

The choice of the ICD-10 code can vary in different institutions, is influenced by the school where the doctors were trained, and it depends on their level of knowledge and other multiple elements, as described in [44]. In addition to these factors, the different electronic medical record software can influence the quality of classification depending on the usability and functionality [45]. For this reason, using a supervised validation system for the medical records labeling, the quality of the data can be increased and, therefore, the performance of the model.

Concerning pretrained deep learning models, we applied identical preprocessing steps as those used with the word-embedding schemes, including the removal of stop words. Although eliminating stop words could potentially result in the loss of specific contextual nuances, we expect that, due to the nature of the classification task, the impact on outcomes will not be substantial. However, future works can address this concern by exploring various preprocessing pipelines.

6. Conclusions

This is the first case aimed to contribute to the solution of one of the most important public health problems in the world. Pregnancy is a complex process, whose outcome prediction could be fed with a different series of data. Structured data are commonly used for this task; however, this is the first approach with unstructured data. In the future, it will be an interesting research topic to use other sources of data as proposed by [46]. We give evidence that it is possible to identify SMM from progress notes in Spanish with an F1-score macro of 52.54%. Hypertensive disorders and miscarriage complications show true positive rates over 83%. We need to continue research to improve the pipelines in order to have clinical utility.

Author Contributions: Conceptualization, E.A.T.-S., S.R., A.F.G.-F., J.F.F.-A., and A.O.-D.; Data curation, E.A.T.-S.; Formal analysis, E.A.T.-S., S.R., A.F.G.-F., M.C.D., J.F.F.-A., and A.O.-D.; Funding acquisition, A.O.-D.; Methodology, E.A.T.-S., S.R., A.F.G.-F., M.C.D., J.F.F.-A., and A.O.-D.; Project administration, A.O.-D.; Software, E.A.T.-S., S.R., A.F.G.-F., M.C.D., and A.O.-D.; Validation, E.A.T.-S. and J.F.F.-A.; Writing—original draft, E.A.T.-S., S.R., and M.C.D.; Writing—review and editing, A.F.G.-F., J.F.F.-A., and A.O.-D. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by the Instituto Tecnológico Metropolitano through the project P20242. Also, this project received funds from the Agencia de Educación Superior de Medellín (Sapiencia) and Universidad Nacional Abierta y a Distancia. This work was also supported by the Clinica Universitaria Bolivariana, Medellín, Colombia, by granting access to anonymized data for this project.

Institutional Review Board Statement: Thid study was conducted according to the guidelines of the Declaration of Helsinki, and approved by the Ethics Committee of Universidad Pontificia Bolivariana (Act 02, 10 February 2020)

Informed Consent Statement: Patient consent was waived because the hospital where the data were collected is a university hospital. All admissions to the hospital imply the signature of the acceptance of data capture through the clinical history based on Colombian regulations. Law of habeas data. In this case, users authorize the processing of data for research purposes.

Data Availability Statement: Due to the sensitive nature of the research, supporting data (raw text) is not available. Only embedded data presented in this study are available on request from the corresponding author.

Acknowledgments: We thank to Clinica Universitaria Bolivariana for allowing access to EHRs.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

To gain a comprehensive clinical perspective, we systematically categorized the document dataset based on ICD-10 chapters (Figure A1). This categorization facilitates a deeper understanding of the specific health conditions afflicting patients. In particular, it provides insights into the distribution of maternal morbidity and mortality attributed to both diseases and external factors. Notably, we found that Chapter XV, which encompasses conditions related to pregnancy, childbirth, and the puerperium was the principal as expected. However, the dataset has an important proportion in chapter XXI (Factors influencing health status and contact with health services), chapter XVIII (XVIII Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified) and XIV (Diseases of the genitourinary system)

Figure A1. Treemap proportion of ICD-10 codes across ICD chapters.

References

1. WHO; UNFPA. *Ending Preventable Maternal Mortality (EPMM): A Renewed Focus for Improving Maternal and Newborn Health and Well-Being*; World Health Organization: Geneva, Switzerland, 2021.
2. Kilpatrick, S.K.; Ecker, J.L.; American College of Obstetricians and Gynecologists. Severe maternal morbidity: Screening and review. *Am. J. Obstet. Gynecol.* **2016**, *215*, B17–B22. [CrossRef] [PubMed]
3. Murdoch, T.B.; Detsky, A.S. The Inevitable Application of Big Data to Health Care. *JAMA* **2013**, *309*, 1351–1352. [CrossRef] [PubMed]
4. Sarker, A.; Chandrashekar, P.; Magge, A.; Cai, H.; Klein, A.; Gonzalez, G. Discovering Cohorts of Pregnant Women from Social Media for Safety Surveillance and Analysis. *J. Med. Internet Res.* **2017**, *19*, e361. [CrossRef] [PubMed]

5. Klein, A.Z.; Cai, H.; Weissenbacher, D.; Levine, L.D.; Gonzalez-Hernandez, G. A natural language processing pipeline to advance the use of Twitter data for digital epidemiology of adverse pregnancy outcomes. *J. Biomed. Inform.* **2020**, *112*, 100076. [CrossRef]
6. Jin, W.; Zhao, B.; Yu, H.; Tao, X.; Yin, R.; Liu, G. Improving embedded knowledge graph multi-hop question answering by introducing relational chain reasoning. *Data Min. Knowl. Discov.* **2022**, *37*, 255–288. [CrossRef]
7. Jin, W.; Zhao, B.; Zhang, L.; Liu, C.; Yu, H. Back to common sense: Oxford dictionary descriptive knowledge augmentation for aspect-based sentiment analysis. *Inf. Process. Manag.* **2023**, *60*, 103260. [CrossRef]
8. Zhong, Q.Y.; Karlson, E.W.; Gelaye, B.; Finan, S.; Avillach, P.; Smoller, J.W.; Cai, T.; Williams, M.A. Screening pregnant women for suicidal behavior in electronic medical records: Diagnostic codes vs. clinical notes processed by natural language processing. *BMC Med. Inform. Decis. Mak.* **2018**, *18*, 104678. [CrossRef]
9. Qiu-Yue, Z.; Mittal, L.P.; Nathan, M.D.; Brown, K.M.; González, D.K.; Cai, T.; Sean Finan, B.G.; Avillach, P.; Smoller, J.W.; Karlson, E.W.; et al. Use of natural language processing in electronic medical records to identify pregnant women with suicidal behavior: Towards a solution to the complex classification problem. *Eur. J. Epidemiol.* **2019**, *34*, 153–162. [CrossRef]
10. Ayre, K.; Bittar, A.; Kam, J.; Verma, S.; Howard, L.M.; Dutta, R. Developing a Natural Language Processing tool to identify perinatal self-harm in electronic healthcare records. *PLoS ONE* **2021**, *16*, e0253809. . [CrossRef]
11. Lenain, R.; Seneviratne, M.G.; Bozkurt, S.; Blayney, D.W.; Brooks, J.D.; Hernandez-Boussard, T. Machine learning approaches for extracting stage from pathology reports in prostate cancer. *Stud. Health Technol. Inform.* **2019**, *264*, 1522.
12. Lauren, P.; Qu, G.; Zhang, F.; Lendasse, A. Discriminant document embeddings with an extreme learning machine for classifying clinical narratives. *Neurocomputing* **2018**, *277*, 129–138. [CrossRef]
13. Atchison, A.; Pinto, G.; Woodward, A.; Stevens, E.; Dixon, D.; Linstead, E. Classifying Challenging Behaviors in Autism Spectrum Disorder with Word Embeddings. In Proceedings of the 2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA), Virtual, 13–16 December 2021; pp. 1325–1332.
14. Zhan, X.; Humbert-Droz, M.; Mukherjee, P.; Gevaert, O. Structuring clinical text with AI: Old versus new natural language processing techniques evaluated on eight common cardiovascular diseases. *Patterns* **2021**, *2*, 100289. [CrossRef] [PubMed]
15. Gui, H.; Tseng, B.; Hu, W.; Wang, S.Y. Looking for low vision: Predicting visual prognosis by fusing structured and free-text data from electronic health records. *Int. J. Med. Inform.* **2022**, *159*, 104678. [CrossRef] [PubMed]
16. Siegersma, K.R.; Evers, M.; Bots, S.H.; Groepenhoff, F.; Appelman, Y.; Hofstra, L.; Tulevski, I.I.; Somsen, G.A.; den Ruijter, H.M.; Spruit, M.; et al. Development of a Pipeline for Adverse Drug Reaction Identification in Clinical Notes: Word Embedding Models and String Matching. *JMIR Med. Inform.* **2022**, *10*, e31063. [CrossRef]
17. Magna, A.A.R.; Allende-Cid, H.; Taramasco, C.; Becerra, C.; Figueroa, R.L. Application of machine learning and word embeddings in the classification of cancer diagnosis using patient anamnesis. *IEEE Access* **2020**, *8*, 106198–106213. [CrossRef]
18. Ribelles, N.; Jerez, J.M.; Rodriguez-Brazzarola, P.; Jimenez, B.; Diaz-Redondo, T.; Mesa, H.; Marquez, A.; Sanchez-Muñoz, A.; Pajares, B.; Carabantes, F.; et al. Machine learning and natural language processing (NLP) approach to predict early progression to first-line treatment in real-world hormone receptor-positive (HR+)/HER2-negative advanced breast cancer patients. *Eur. J. Cancer* **2021**, *144*, 224–231. [CrossRef]
19. Almagro, M.; Unanue, R.M.; Fresno, V.; Montalvo, S. ICD-10 coding of Spanish electronic discharge summaries: An extreme classification problem. *IEEE Access* **2020**, *8*, 100073–100083. [CrossRef]
20. Chen, P.F.; Chen, K.C.; Liao, W.C.; Lai, F.; He, T.L.; Lin, S.C.; Chen, W.J.; Yang, C.Y.; Lin, Y.C.; Tsai, I.C.; et al. Automatic International Classification of Diseases coding system: Deep contextualized language model with rule-based approaches. *JMIR Med. Inform.* **2022**, *10*, e37557. [CrossRef]
21. Mantel, G.D.; Buchmann, E.; Rees, H.; Pattinson, R.C. Severe acute maternal morbidity: A pilot study of a definition for a near-miss. *BJOG Int. J. Obstet. Gynaecol.* **1998**, *105*, 985–990. [CrossRef]
22. Waterstone, M.; Murphy, J.D.; Bewley, S.; Wolfe, C. Incidence and predictors of severe obstetric morbidity: Case-control study. *BMJ* **2001**, *322*, 1089–1094. [CrossRef]
23. De Salud INS, I.N. Protocolo de Vigilancia de Salud Pública—Morbilidad Materna Extrema. *Boletin SIVIGILA* **2023**, *743*. Available online: https://www.ins.gov.co/buscador-eventos/Paginas/Vista-Boletin-Epidemilogico.aspx (accessed on 2 May 2023).
24. Gobierno Nacional Republica de Colombia Ley Estatutaria 1581 De 2012. 2012, Available online: https://www.funcionpublica.gov.co/eva/gestornormativo/norma.php?i=49981 (accessed on 10 June 2022).
25. De Científicos de Datos (UCD) Departamento Nacional de Planeación. ConTexto—Librería de Procesamiento y Análisis de Textos v0.2.0. 2021. Available online: https://github.com/ucd-dnp/ConTexto (accessed on 11 November 2022).
26. Al-Rfou, R.; Perozzi, B.; Skiena, S. Polyglot: Distributed word representations for multilingual NLP. In Proceedings of the CoNLL 2013—17th Conference on Computational Natural Language Learning, Sofia, Bulgaria, 8–9 August 2013; pp. 183–192.
27. Cardellino, C. Spanish Billion Words Corpus and Embeddings, 2019. Available online: https://crscardellino.ar/SBWCE/ (accessed on 6 February 2022).
28. Khattak, F.K.; Jeblee, S.; Pou-Prom, C.; Abdalla, M.; Meaney, C.; Rudzicz, F. A survey of word embeddings for clinical text. *J. Biomed. Inform.* **2019**, *100*, 100057. [CrossRef]
29. Lauren, P.; Qu, G.; Zhang, F.; Lendasse, A. Clinical narrative classification using discriminant word embeddings with elm. In Proceedings of the 2016 International Joint Conference on Neural Networks (IJCNN), Vancouver, BC, Canada, 24–29 July 2016; pp. 2931–2938.

30. Segura-Bedmar, I.; Colón-Ruíz, C.; Tejedor-Alonso, M.Á.; Moro-Moro, M. Predicting of anaphylaxis in big data EMR by exploring machine learning approaches. *J. Biomed. Inform.* **2018**, *87*, 50–59. [CrossRef] [PubMed]
31. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
32. Gutiérrez-Fandiño, A.; Armengol-Estapé, J.; Pàmies, M.; Llop-Palao, J.; Silveira-Ocampo, J.; Carrino, C.P.; Gonzalez-Agirre, A.; Armentano-Oller, C.; Rodriguez-Penagos, C.; Villegas, M. Maria: Spanish language models. *arXiv* **2021**, arXiv:2107.07253.
33. Beltagy, I.; Peters, M.E.; Cohan, A. Longformer: The long-document transformer. *arXiv* **2020**, arXiv:2004.05150.
34. Ruch, P.; Baud, R.; Geissbühler, A. Using lexical disambiguation and named-entity recognition to improve spelling correction in the electronic patient record. *Artif. Intell. Med.* **2003**, *29*, 169–184. [CrossRef]
35. Xuan, L.; Zhigang, C.; Fan, Y. Exploring of clustering algorithm on class-imbalanced data. In Proceedings of the 2013 8th International Conference on Computer Science & Education, Colombo, Sri Lanka, 26–28 April 2013; pp. 89–93.
36. Norgeot, B.; Quer, G.; Beaulieu-Jones, B.K.; Torkamani, A.; Dias, R.; Gianfrancesco, M.; Arnaout, R.; Kohane, I.S.; Saria, S.; Topol, E.; et al. Minimum information about clinical artificial intelligence modeling: The MI-CLAIM checklist. *Nat. Med.* **2020**, *26*, 1320–1324. [CrossRef] [PubMed]
37. Wang, Y.; Liu, S.; Afzal, N.; Rastegar-Mojarad, M.; Wang, L.; Shen, F.; Kingsbury, P.; Liu, H. A comparison of word embeddings for the biomedical natural language processing. *J. Biomed. Inform.* **2018**, *87*, 12–20. [CrossRef]
38. Gladkova, A.; Drozd, A. Intrinsic evaluations of word embeddings: What can we do better? In Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP, Berlin, Germany, 7–12 August 2016; pp. 36–42.
39. Naseem, U.; Razzak, I.; Eklund, P.W. A survey of pre-processing techniques to improve short-text quality: A case study on hate speech detection on twitter. *Multimed. Tools Appl.* **2021**, *80*, 35239–35266. [CrossRef]
40. Workman, T.E.; Shao, Y.; Divita, G.; Zeng-Treitler, Q. An efficient prototype method to identify and correct misspellings in clinical text. *BMC Res. Notes* **2019**, *12*, 42. [CrossRef] [PubMed]
41. Kowsari, K.; Jafari Meimandi, K.; Heidarysafa, M.; Mendu, S.; Barnes, L.; Brown, D. Text classification algorithms: A survey. *Information* **2019**, *10*, 150. [CrossRef]
42. Yogarajan, V. Domain-Specific Language Models for Multi-Label Classification of Medical Text. Ph.D. Thesis, The University of Waikato, Hamilton, New Zealand, 2022.
43. Kumar, V.; Recupero, D.R.; Riboni, D.; Helaoui, R. Ensembling classical machine learning and deep learning approaches for morbidity identification from clinical notes. *IEEE Access* **2020**, *9*, 7107–7126. [CrossRef]
44. Carvalho, R.; Lobo, M.; Oliveira, M.; Oliveira, A.R.; Lopes, F.; Souza, J.; Ramalho, A.; Viana, J.; Alonso, V.; Caballero, I.; et al. Analysis of root causes of problems affecting the quality of hospital administrative data: A systematic review and Ishikawa diagram. *Int. J. Med. Inform.* **2021**, *156*, 104584. [CrossRef] [PubMed]
45. Horsky, J.; Drucker, E.A.; Ramelson, H.Z. Accuracy and Completeness of Clinical Coding Using ICD-10 for Ambulatory Visits. *AMIA Annu. Symp. Proc.* **2017**, *2017*, 912—920.
46. Espinosa, C.; Becker, M.; Marić, I.; Wong, R.J.; Shaw, G.M.; Gaudilliere, B.; Aghaeepour, N.; Stevenson, D.K.; Stelzer, I.A.; Peterson, L.S.; et al. Data-driven modeling of pregnancy-related complications. *Trends Mol. Med.* **2021**, *27*, 762–776. [CrossRef]