Procesamiento del Lenguaje Natural, Revista nº 69, septiembre de 2022, pp. 129-140 recibido 21-03-2022 revisado 17-05-2022 aceptado 30-05-2022

Detección de Indicios de Autolesiones No Suicidas en Informes Médicos de Psiquiatría Mediante el Análisis del Lenguaje

Detecting Signs of Non-suicidal Self-Injury in Psychiatric Medical Reports Using Language Analysis

Juan Martinez-Romo[1],[2] **Blanca Reneses**[1],[2],[3] **Ignacio Martínez-Capella**
Lourdes Araujo[1],[2] **J. Sevilla-Llewellyn-Jones**[1] **Germán Seara-Aguilar**
[1]NLP & IR Group
Universidad Nacional de Educación a Distancia (UNED)
[2]Instituto Mixto UNED-ISCIII IMIENS
juaner,lurdes@lsi.uned.es

[1]IdISSC
Hospital Clínico San Carlos
[2]CIBERSAM
[3]Universidad Complutense
blanca.reneses@salud.madrid.org
julia.sevilla@salud.madrid.org

Unidad de Innovación
IdISSC
Hospital Clínico San Carlos
Madrid
imcapella@salud.madrid.org
gseara@shealth.eu

Resumen: La autolesión no suicida, a menudo denominada autolesión, es el acto de dañarse deliberadamente el propio cuerpo, como cortarse o quemarse. Normalmente, no pretende ser un intento de suicidio. En este trabajo se presenta un sistema de detección de indicios de autolesiones no suicidas, basado en el análisis del lenguaje, sobre un conjunto anotado de informes médicos obtenidos del servicio de psiquiatría de un Hospital público madrileño. Tanto la explicabilidad como la precisión a la hora de predecir los casos positivos, son los dos principales objetivos de este trabajo. Para lograr este fin se han desarrollado dos sistemas supervisados de diferente naturaleza. Por un lado se ha llevado a cabo un proceso de extracción de diferentes rasgos centrados en el propio mundo de las autolesiones mediante técnicas de procesamiento del lenguaje natural para alimentar posteriormente un clasificador tradicional. Por otro lado, se ha implementado un sistema de aprendizaje profundo basado en varias capas de redes neuronales convolucionales, debido a su gran desempeño en tareas de clasificación de textos. El resultado es el funcionamiento de dos sistemas supervisados con un gran rendimiento, en donde destacamos el sistema basado en un clasificador tradicional debido a su mejor predicción de clases positivas y la mayor facilidad de cara a explicar sus resultados a los profesionales sanitarios.
Palabras clave: Detección de autolesiones no suicidas, análisis del lenguaje, aprendizaje automático, redes neuronales.

Abstract: Non-suicidal self-injury, often referred to as self-injury, is the act of deliberately harming one's own body, such as cutting or burning oneself. It is not usually intended as a suicide attempt. This paper presents a system for detecting signs of non-suicidal self-injury, based on language analysis, on an annotated set of medical reports obtained from the psychiatric service of a public hospital in Madrid. Both explainability and accuracy in predicting positive cases are the two main objectives of this work. In order to achieve this goal, two supervised systems of different natures have been developed. On the one hand, a process of extraction of different features focused on the world of self-injury itself has been carried out using natural language processing techniques to subsequently feed a traditional classifier. On the other hand, a deep learning system based on several layers of convolutional neural networks, due to its high performance in text classification tasks. The result are two supervised systems with high performance, where we highlight the system based on a traditional classifier due to its better prediction of positive classes and the greater ease to explain its results to health professionals.
Keywords: Non-suicidal Self-injury detection, language analysis, machine learning, neural networks.

ISSN 1135-5948. DOI 10.26342/2022-69-11

1 Introducción

Los trastornos de salud mental, como las autolesiones, son problemas cuya incidencia en la población aumenta de manera alarmante en los últimos años. Estas afecciones pueden pasar desapercibidas durante muchos años, lo que hace que las personas que las padecen no reciban la asistencia médica adecuada. Los problemas de salud mental sin tratar pueden acarrear graves consecuencias, como el deterioro personal o incluso el suicidio. Las autolesiones, también conocidas como autolesiones deliberadas o autoagresiones, son un tipo de problema de salud mental menos conocido que afecta principalmente a los jóvenes (Young et al., 2007). La autolesión se refiere al acto de causarse daño corporal a sí mismo sin intención suicida, como cortarse, quemarse o tirarse del pelo, y se ha relacionado con problemas de salud mental subyacentes, como la depresión y la ansiedad (Greaves, 2018b). Entre las diferentes acciones que las personas afectadas llevan a cabo dentro del concepto general de autolesión, existen grandes diferencias tanto en los motivos para llevarlas a cabo, como en el género (Rodham, Hawton, y Evans, 2004). La naturaleza a menudo impulsiva de estos actos (especialmente el auto-corte) significa que la prevención debe centrarse en fomentar métodos alternativos de gestión de la ansiedad, la resolución de problemas y la búsqueda de ayuda antes de que se desarrollen pensamientos de autolesión. Dada la gravedad de los síntomas y los riesgos, es importante dedicar esfuerzos a detectar mejor los problemas de salud mental en la sociedad para que puedan recibir la ayuda que necesitan. También se han encontrado diferencias en la forma de comunicarse y en el lenguaje empleado por las personas que sufren problemas de salud mental (Pennebaker, Mehl, y Niederhoffer, 2003). A pesar de que los informes médicos están generalmente escritos por médicos, en ocasiones tratan de plasmar las ideas subyacentes del paciente e incluso escriben literalmente frases o expresiones utilizadas y que puedan denotar de forma clara estados de ánimo o pensamientos. De esta forma, el análisis del lenguaje empleado en estos informes médicos mediante técnicas de Procesamiento del Lenguaje Natural (PLN) pueden ayudar a la detección temprana de pacientes con otros trastornos previos. Sin embargo, cabe señalar que la mayoría de los estudios sobre detección temprana de peligros para la seguridad y la salud se han centrado en el texto en inglés. Por otra parte, hay que señalar que apenas existen conjuntos de datos (datasets o corpora) para entrenar modelos de identificación en las tareas mencionadas, y los existentes se limitan al inglés y son de tamaño reducido, lo cual es un claro indicador del camino que aún queda por recorrer para que los profesionales sanitarios puedan disponer de herramientas maduras de análisis de textos. En este trabajo se presenta un sistema de detección de autolesiones en informes médicos procedentes del servicio de psiquiatría del Hospital Clínico San Carlos de Madrid. Este sistema de detección de autolesiones entrenado y evaluado sobre un corpus anotado de informes médicos, tiene el objetivo de aplicarse a cualquier informe del servicio de psiquiatría para permitir la detección temprana de este trastorno en pacientes que hayan sido tratados por dicho servicio. De esta forma, pacientes con otro tipo de trastornos de carácter más leve, podrían ser diagnosticados y recibir tratamiento antes de adquirir estos hábitos tan perjudiciales. Y es que la detección temprana es clave en el tratamiento de los problemas de salud mental, ya que una intervención rápida mejora las probabilidades de un buen pronóstico.

El resto del artículo se organiza de la siguiente forma: en la Sección 2 se analiza el estado del arte y trabajos relacionados. en la Sección 3 se describe el corpus y las técnicas utilizada para anotarlo. En la Sección 4 se detallan las características del sistema de detección de autolesiones. La Sección 5 se centra en la experimentación y el análisis de resultados. Finalmente, en la Sección 6 se extraen las principales conclusiones y se exponen las líneas de trabajo futuro.

2 Estado del Arte

El estudio de las autolesiones y más concretamente de su investigación a través del análisis de textos no es demasiado extenso. Existen trabajos (Baetens et al., 2011) en los que se investigaron la prevalencia de las autolesiones no suicidas (NSSI) y las autolesiones suicidas (SSI) en una muestra de adolescentes de entre 12 y 18 años, así como las diferencias psicosociales entre los adolescentes que practican NSSI y los que practican SSI. También hay trabajos (Nicolai, Wielgus, y Mezulis, 2016) que apoyan la teoría de la cascada emocio-

nal, en la que la rumiación distingue entre las personas que se autolesionan y las que no lo hacen, y hacen especial hincapié en la relación entre el afecto negativo y las NSSI.

Hay trabajos (Burke, Ammerman, y Jacobucci, 2019) que se han centrado en abordar las limitaciones de los sistemas de detección de riesgo y el tiempo de cómputo utilizando herramientas analíticas avanzadas, como el procesamiento del lenguaje natural (PLN) y el aprendizaje automático. Existen estudios centrados en la ideación de suicidio que usan enfoques de PLN y que han utilizado en gran medida modelos basados en la historia clínica electrónica (HCE) (Haerian, Salmasian, y Friedman, 2012; Kessler et al., 2017) y modelos de predicción basados en PLN y rasgos lingüísticos (Fernandes et al., 2018; McCoy et al., 2016; Poulin et al., 2014).

En 2017, un trabajo (Walsh, Ribeiro, y Franklin, 2017) utilizó aprendizaje automático para predecir el riesgo de suicidio en pacientes de autolesiones a lo largo del tiempo analizando informes médicos de una gran base de datos médica. También se han usado tests adaptativos informatizados (CAT) para entrenar un árbol de decisión con el objetivo de predecir el riesgo de suicidio (Delgado-Gomez et al., 2016). Otros trabajos (Metzger et al., 2017) han empleado algoritmos de clasificación como random forest and naïve Bayes sobre informes médicos para predecir suicidios, demostrando que los métodos de aprendizaje automático pueden mejorar la calidad de los indicadores epidemiológicos en comparación con la actual vigilancia nacional de los intentos de suicidio de un país como Francia. Los árboles de decisión también se han usado en otros trabajos (Mann et al., 2008) para estudiar la correlación entre pacientes de psiquiatría analizando su conducta suicida pasada frente a la ideación de suicidio que mostraban en un momento dado.

La clasificación de textos clínicos mediante redes neuronales ha resultado una herramienta de gran utilidad en problemas como la identificación de fenotipos en informes médicos para pacientes con un conjunto determinado de signos y síntomas clínicos (Obeid et al., 2019). En los últimos años se han producido avances significativos en los enfoques de aprendizaje profundo, como las redes neuronales convolucionales (CNN), y su aplicación ha sido un éxito en problemas como el procesamiento y la clasificación de textos o el reconocimiento del habla (LeCun, Bengio, y Hinton, 2015).

Recientemente ha surgido un estudio (Obeid et al., 2020) que aprovecha la información de las notas clínicas utilizando redes neuronales profundas (DNNs) para identificar los pacientes tratados por autolesión intencional y predecir futuros eventos de autolesión. Los autores utilizaron dos modelos basados en una CNN y en una LSTM con resultados prometedores. También se han usado técnicas de clasificación como el Gradient Boosting para la detección de autolesiones e ideación suicida en las notas de triaje de los servicios de urgencias (Rozova et al., 2022).

En los últimos años han aparecido bastantes trabajos en el ámbito de las redes sociales y la salud mental. Aunque el formato del texto de las redes sociales y en lenguaje escrito en primera persona hacen abordar este problema desde un enfoque diferente, queremos destacar algunos trabajos por su relevancia y su cercanía al problema de las autolesiones.

Desde 2017 y de forma anual se celebra la tarea competitiva eRisk (Losada, Crestani, y Parapar, 2019; Losada, Crestani, y Parapar, 2020; Parapar et al., 2021), dentro del congreso CLEF (Cross Language Evaluation Forum). eRisk trata de avanzar en la predicción temprana en redes sociales de problemas relacionados con la salud mental. Depresión, anorexia, ludopatía y autolesiones desde el año 2019 han sido los trastornos elegidos por los organizadores. Dentro de esta competición, un sistema con resultados prometedores fue el equipo iLab (Martınez-Castano et al., 2020) en el que los investigadores propusieron un sistema de clasificación basado en BERT y transformers. En contraposición al uso pesado de las redes neuronales y los transformers, también resultan interesantes las participaciones del grupo NLP-UNED (Ageitos, Martínez-Romo, y Araujo, 2020; Campillo-Ageitos et al., 2021), que emplearon técnicas de PLN y análisis de sentimientos para alimentar un clasificador rápido y eficiente. Finalmente, un trabajo que obtuvo buenos resultados con un sistema innovador fue el del grupo UNSL(Loyola et al., 2021), que empleó políticas de alerta, un sistema basado en reglas y un modelo de aprendizaje por refuerzo.

3 Corpus

El corpus de evaluación procede de un conjunto de informes médicos anonimizados pro-

cedentes del servicio de psiquiatría del Hospital Clínico San Carlos de Madrid en España.

La preparación de los informes para su análisis ha sido desarrollada por la Unidad de Innovación del Hospital Clínico San Carlos, a partir de la descarga autorizada de informes informatizados del Servicio de Psiquiatría correspondientes a un periodo de cuatro años. Dicha preparación ha consistido en tres fases: limpieza de los informes, compleción y anonimización. Previamente, esta cesión de datos fue evaluada y aprobada por el Comité de Ética de la Investigación (20/586-E).

A partir de este conjunto de informes anonimizados, se llevó a cabo un proceso de anotación por parte de expertos dando lugar a un corpus de 1252 informes anotados. Los diagnósticos de estos informes son diversos, pero entre ellos no se incluye sufrir autolesiones. Por ello ha sido necesaria una anotación manual supervisada por los médicos expertos en base al contenido textual de los informes. Tras la anotación manual en busca de indicios de autolesiones, 1138 han sido anotados como negativos y 114 como positivos. Durante el proceso de anotación, se buscaban indicios claros de que el profesional sanitario que hubiera atendido al paciente indicara que las autolesiones se habían producido. La mera ideación o pensamiento de esta situación fue tratada como un caso negativo. En el caso de situaciones en las que las autolesiones tenían un fin autolítico también fueron etiquetadas como casos negativos al buscar otro fin diferente al ansiolítico. Estos últimos casos deberían tratarse en un estudio diferente como parte de los pacientes con riesgo de suicidio. Los informes tienen una media de 1310 palabras y 7566 caracteres por cada informe, teniendo el informe de mayor tamaño 1639 palabras y 32767 caracteres y el de menor tamaño 84 palabras y 510 caracteres. El corpus se ha dividido en dos conjuntos de entrenamiento (80 %) y test (20 %), resultando dos conjuntos de 1001 y 251 informes respectivamente. La división se ha llevado a cabo de forma estratificada para respetar la proporción de clases en los conjuntos de entrenamiento y test.

4 Sistema de Detección de Autolesiones

Para la tarea de detección de autolesiones se han desarrollado dos sistemas supervisados, uno de ellos basado en la extracción de rasgos y la aplicación de algoritmos clásicos de clasificación y el otro basado en redes neuronales con la aplicación de un modelo BERT para el tokenizado. Los dos sistemas desarrollados solo analizan el texto anonimizado del informe sin tener en cuenta el diagnóstico que aparece en otro campo y que sólo ha sido tenido en cuenta en el proceso de anotación manual para ayudar a los expertos en caso de duda.

4.1 Sistema de Aprendizaje Automático

El sistema supervisado está compuesto de tres módulos diferentes que se encargan de las tareas de pre-procesamiento, extracción de rasgos y aplicación de algoritmos de clasificación.

4.1.1 Pre-procesamiento de los Informes Médicos

En cuanto al pre-procesamiento se han aplicado las técnicas habituales, como son la conversión a minúsculas del texto, la eliminación de caracteres especiales, la normalización de determinados conectores, el tokenizado del texto y el borrado de palabras vacías. También se ha llevado a cabo un proceso de stemming mediante el segundo algoritmo de Porter para extraer la raíz de las palabras.

4.1.2 Extracción de Rasgos y Algoritmos de Clasificación

La extracción de rasgos se puede agrupar en cinco conjuntos de características:

- **CountVectorizer**: En primer lugar se ha utilizado la herramienta CountVectorizer de la biblioteca scikit-learn en Python para obtener vectores de palabras a partir de los informes médicos. Esta herramienta se utiliza para transformar un texto dado en un vector sobre la base de la frecuencia de cada palabra que aparece en todo el texto. La función crea una matriz en la que cada palabra única está representada por una columna de la matriz, y cada muestra de texto del documento es una fila en dicha matriz. El valor de cada celda no es más que la frecuencia de la palabra en esa muestra de texto en particular.
- **Vocabulario de Autolesiones**: Se ha compilado un conjunto de 53 palabras relacionadas con el contexto de las autolesiones. En este conjunto hay palabras

como morder, cortar, pellizcar, etc. Este conjunto de palabras, se emplea como entrada del CountVectorizer para no usar todo el vocabulario completo sino solo estas 53 palabras, algo que proporciona mayor rapidez y precisión.

- **Diccionario NSSI**: Greaves (Greaves, 2018a) desarrolló un trabajo en el que llevó a cabo la clasificación de un conjunto de conceptos relacionados con las autolesiones. El resultado de este trabajo es un diccionario de palabras relacionadas con la autolesión llamado Diccionario Non-Suicidal Self-Injury (NSSI), donde las palabras se dividen en cinco categorías: 1) Métodos de NSSI; 2) Términos de NSSI; 3) Instrumentos utilizados; 4) Razones de NSSI; y (5) Términos específicos de cortes. De esta forma, se han creado cuatro rasgos de NSSI, uno para cada categoría. Estas características cuentan la frecuencia de las palabras de su categoría en el texto.
- **Distancia de Términos de Autolesiones**: Existen numerosos trabajos que han probado la relevancia de las primeras palabras de un documento en relación al texto completo. En este grupo de rasgos se ha tratado de medir por un lado la distancia entre el inicio del documento y la primera palabra del vocabulario de autolesiones presente en el texto y por otro lado la distancia media entre palabras del vocabulario de autolesiones. Estas medidas se han realizado en función del número de palabras y del número de caracteres, dando lugar a cuatro rasgos.
- **Negación**: Se ha llevado a cabo un proceso de detección de la negación mediante una arquitectura (Fabregat, Araujo Serna, y Martínez Romo, 2019; Fabregat et al., 2019) basada en aprendizaje profundo. La detección de la negación se ha aplicado a los grupos de rasgos definidos anteriormente para eliminar la presencia de los términos de autolesiones que han sido negados y de esta forma restar su incidencia. Es decir, si en el texto aparece una afirmación como "No se aprecian cortes", la detección de negación evita que el término "cortes" se contabilice en ninguno de los rasgos calculados en este trabajo.

Una vez extraídos los rasgos descritos anteriormente y con la ayuda del corpus anotado, se ha llevado a cabo la aplicación de los algoritmos más efectivos según el estado del arte en este tipo tareas.

4.2 Sistema basado en Aprendizaje Profundo

El segundo sistema usa redes neuronales con una arquitectura en la que se disponen tres capas de redes neuronales convolucionales y para la que se ha adaptado la tecnología de BERT (Devlin et al., 2018) para el proceso de tokenizado, que está basado en la representación de codificadores binarios a partir de Transformers. En este caso, hemos adaptado esta tecnología para la clasificación de textos.

Aparte de la preparación del texto, para el tokenizado de los textos médicos, hemos usado dos modelos pre-entrenados: Un modelo base de BERT[1] que está disponible en seis idiomas, incluido el español, y fue creado para tareas de clasificación de textos. Y el modelo RoBERTa-base-bne[2], que es un modelo de lenguaje enmascarado basado en transformers para el español. Está basado en el modelo base de RoBERTa y ha sido pre-entrenado utilizando el mayor corpus en español conocido hasta la fecha, con un total de 570GB de texto limpio y procesado expresamente para este trabajo. El texto procede de una compilación de páginas web realizada por la Biblioteca Nacional Española desde 2009 hasta 2019.

De esta forma, se ha adoptado una arquitectura que consta de tres capas de redes neuronales convolucionales concatenadas. La arquitectura del sistema de aprendizaje profundo usada para este trabajo puede apreciarse en la Figura 1, en la que se muestra a nivel general la arquitectura de la red neuronal, con tres capas de redes neuronales convolucionales (CNN) y dos capas de redes neuronales densamente conectadas, la última empleada como capa de clasificación.

5 Resultados

Para la evaluación de los dos sistemas desarrollados vamos a usar las medidas tradicionales de clasificación precisión, cobertura y

[1] https://huggingface.co/nlptown/bert-base-multilingual-uncased-sentiment

[2] https://huggingface.co/PlanTL-GOB-ES/roberta-base-bne

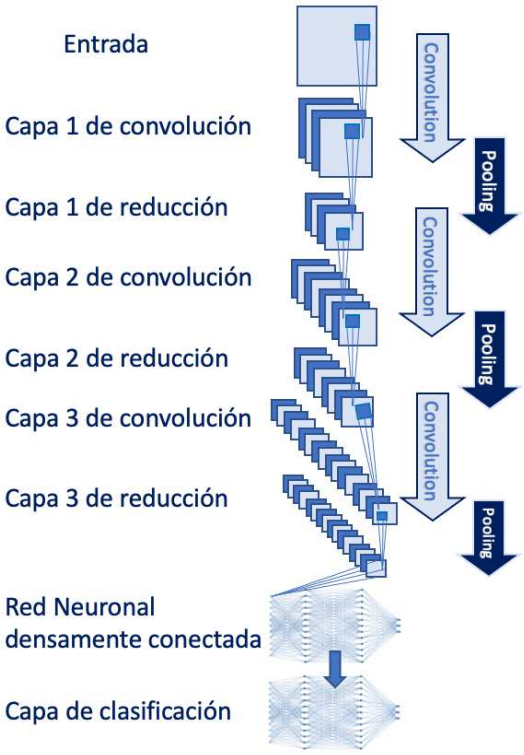

Figura 1: Arquitectura de la red neuronal.

medida-F. Además, como la detección de casos de autolesiones es una tarea cuyas implicaciones requieren una gran precisión, uno de los principales objetivos de este trabajo es la búsqueda de un buen desempeño a la hora de predecir casos positivos. También el hecho de ser una tarea relacionada con la salud, requiere de un grado satisfactorio de explicabilidad de cara a los profesionales que en última instancia deben de tomar las decisiones.

5.1 Baselines

En primer lugar, hemos desarrollado cuatro baselines para medir la calidad de los sistemas supervisados.

- **Most frequent Class (MFC):** El sistema anota todos los casos con la clase más frecuente, que en este caso es la clase negativa.
- **Less frequent Class (LFC):** El sistema anota todos los casos con la clase menos frecuente, que en este caso es la clase positiva.
- **Random prediction:** El sistema asigna una predicción aleatoria a cada instancia.
- **Random Ratio prediction:** El sistema asigna mediante una función de probabilidad una predicción aleatoria a cada instancia, manteniendo el mismo ratio (positivas/negativas) de anotaciones que el conjunto de test.

La Tabla 1 muestra los resultados tras la aplicación de los baselines. Como era de esperar, al tratarse de un corpus desbalanceado, el baseline que mejor rendimiento obtiene es aquel que predice como negativos todos los casos al ser la clase mayoritaria.

5.2 Combinación de Rasgos

En la Tabla 2 se pueden apreciar los resultados obtenidos tras diferentes combinaciones de los rasgos descritos en la sección 4.1.2. Para estos resultados se ha aplicado un algoritmo de regresión logística. De forma evidente en cuanto a la hipótesis de partida, los peores resultados se obtienen con los vectores de palabras formados por el vocabulario completo del corpus (32K palabras) y los mejores se consiguen con la combinación de todos los rasgos computados. En cuanto a la parte más interesante de esta combinación, destaca la diferencia entre la mejora obtenida por la clase negativa y la positiva al introducir los rasgos. La clase negativa solo aumenta tres puntos su medida-F, mientras que la clase positiva aumenta 21 puntos al introducir todos los rasgos. Esta diferencia demuestra la eficiencia de los rasgos introducidos en cuanto al propósito general de mejorar sobre todo la predicción de los casos positivos. En cuanto a los rasgos, analizados de manera individual, destaca la aportación de los vectores de palabras obtenidos a partir del vocabulario compilado manualmente de 53 palabras. La diferencia entre usar el vocabulario completo o solo las 53 palabras, se refleja en un aumento de 14 puntos en la medida-F de la clase positiva, aumentando tanto la precisión como la cobertura. Los rasgos que menos aportación parecen tener en el cómputo global son las distancias entre términos de autolesiones y la negación, quizás debido a que no se producen demasiadas en los informes médicos o su relevancia es menor de la esperada en cuanto al contexto global del informe. En cuanto a la clase positiva, la precisión y cobertura aumentan de forma desigual, teniendo los rasgos computados un impacto mayor en la precisión que en la cobertura. Este hecho era de esperar dado que al menos el rasgo que usa los vectores de palabras con un vocabulario

BASELINES					
Baseline	**F1 Todas Clases**			**Clase Positiva**	
	F1-NO	**F1-SI**	***F1-weighted Avg***	**P-SI**	**R-SI**
MFC	**0.95**	0.00	**0.86**	0.00	0.00
LFC	0.00	**0.17**	0.02	**0.09**	**1.00**
Random Prediction	0.65	0.14	0.60	0.08	0.43
Random Ratio Prediction	0.91	0.05	0.83	0.05	0.04

Tabla 1: F1-SI: F1-measure de los casos positivos, F1-NO: F1-measure de los casos negativos, F1-weighted Avg: F1-measure media de todo el conjunto de test, P-SI: Precisión de los casos positivos, R-SI: Recall de los casos positivos

reducido implica profundizar en esa dirección precisamente.

5.3 Análisis de diferentes algoritmos de Clasificación

En la sección anterior se empleó un algoritmo de regresión logística para la tarea de clasificación. En la Tabla 3 se muestran los resultados al aplicar los diferentes algoritmos de clasificación que mejor rendimiento han obtenido en diferentes trabajos del estado del arte consultados. Para esta comparativa se ha usado la combinación de rasgos que mejor rendimiento obtuvo en la sección anterior y cuyos resultados se pueden observar en la Tabla 2. Como se puede ver, hay tres algoritmos (Logistic Regression, Gradient Boosting y SVM) que obtienen los mejores resultados en cuanto a la medida-F global. Sin embargo, como uno de los objetivos de este trabajo consiste en mejorar la detección de la clase positiva, se observa que el algoritmo "Gradient Boosting" obtiene el mejor rendimiento en la predicción de casos positivos. Esto unido a que era uno de los tres algoritmos que de forma global obtenían mejores resultados lo convierten en la mejor opción para nuestro sistema. Profundizando en los resultados de "Gradient Boosting", aparte de obtener los mejores resultados en las clases positivas, negativas, y de forma global, obtiene mejor cobertura que ningún otro algoritmo. Esta parece ser su mejor aportación, ya que su precisión en la clase positiva es superada por otros algoritmos.

5.4 Sistema basado en Aprendizaje Profundo

En la Tabla 4 se puede observar el rendimiento de los sistemas basados en redes neuronales. Se ha optado por variar dos hiperparámetros como son el número de épocas y el dropout. En todos los experimentos se han usado embeddings de 200 dimensiones. De los resultados obtenidos en cuanto a las diferentes combinaciones no se pueden obtener demasiadas conclusiones. Quizás se puede observar que un dropout bajo mejora el rendimiento global aunque no es concluyente. De forma general podría decirse que cinco épocas han funcionado mejor que diez, al igual que ocurre para la clase positiva con la que ligeramente se observan mejores resultados. En cuanto a la precisión de los casos positivos, con una combinación se obtienen mucho mejores resultados que con el resto, sin embargo su cobertura y medida-F se ven negativamente afectadas. La única conclusión evidente a nivel de diferentes combinaciones se produce en la cobertura de la clase positiva. En este caso un menor número de épocas y un bajo dropout implican un significativo mejor rendimiento que los casos opuestos con una diferencia de 50 puntos.

5.5 Análisis Global de Resultados

De forma general y tal como se muestra en la Tabla 5, los sistemas desarrollados superan ampliamente a los baselines propuestos al inicio del trabajo. En cuanto a la comparativa entre el mejor sistema supervisado y el mejor sistema basado en redes neuronales, globalmente el sistema de aprendizaje profundo obtiene mejores resultados si atendemos a la medida-F. Sin embargo, la pequeña diferencia a favor de las redes neuronales en relación a la medida-F global y de la clase negativa, se ve ampliamente superada en cuanto a la clase positiva tanto en la medida-F como en la precisión y la cobertura. Destaca notablemente la diferencia en la precisión, de forma muy significativa la medida-F y de forma relevante la cobertura, siendo esta última la medida donde la diferencia de rendimiento es algo

COMBINACIÓN DE RASGOS					
Features	**F1-NO**	**F1-SI**	***F1-W Avg***	**P-SI**	**R-SI**
CV	0.94	0.47	0.90	0.61	0.33
CV + NSSI	0.96	0.49	0.92	0.64	0.39
CV + VAL	0.96	0.61	0.93	0.84	0.50
CV + VAL + NSSI	**0.97**	0.63	0.94	0.86	0.52
CV + VAL + NSSI + DIS + NEG	**0.97**	**0.65**	**0.95**	**0.87**	**0.54**

Tabla 2: CV: CountVectorizer, VAL: Vocabulario de Autolesiones, NSSI: Rasgos de NSSI, DIS: Rasgos de distancia de terminos de autolesion, NEG: Negación

ALGORITMOS DE CLASIFICACIÓN					
Algoritmo	**F1 Todas Clases**			**Clase Positiva**	
	F1-NO	**F1-SI**	***F1-weighted Avg***	**P-SI**	**R-SI**
Logistic Regression	**0.97**	0.65	**0.94**	0.86	0.52
Random Forest	0.95	0.53	0.91	0.50	0.57
Gradient Boosting	**0.97**	**0.68**	**0.94**	0.71	**0.65**
K Neighbours	0.96	0.36	0.91	**1.00**	0.22
SVM	**0.97**	0.59	**0.94**	0.91	0.43
Adaboost	0.96	0.59	0.93	0.62	0.57

Tabla 3: F1-SI: F1-measure de los casos positivos, F1-NO: F1-measure de los casos negativos, F1-weighted Avg: F1-measure media de todo el conjunto de test, P-SI: Precisión de los casos positivos, R-SI: Recall de los casos positivos

menor. De esta forma, y teniendo en cuenta las implicaciones en cuanto a mejor explicabilidad del sistema supervisado basado en el algoritmo "Gradient Boosting", consideramos que la opción más óptima para la tarea concreta en la que se centra este trabajo es dicho sistema.

5.6 Análisis de la Incidencia de las Categorías de Rasgos

Dado que el sistema supervisado ofrece un mejor rendimiento en la detección de casos positivos y además su grado de explicabilidad es mayor, hemos decidido profundizar en los rasgos extraídos y su incidencia en los resultados. En la figura 2 se muestra un gráfico de barras que representa la frecuencia de aparición de los rasgos que componen las diferentes categorías en función de la clase a la que pertenecen. Dicha frecuencia se ha normalizado en función del número de documentos de cada clase y tamaño del vocabulario de cada categoría, dado que el corpus está muy desbalanceado. En esta figura destacan positivamente categorías como el vocabulario de autolesiones, los términos de NSSI, los conceptos de autolesiones por cortes de NSSI y la negación. En cuanto al vocabulario de autolesiones, se intuía esta diferencia debido a los resultados obtenidos. En cuanto a la negación, también se observa una gran disparidad. Finalmente en cuanto a los conceptos de NSSI, los que mejor parecen representar a la clase positiva son los "Términos" y el "Cutting". Sin embargo, los conceptos de "Razones", "Métodos" e "Instrumentos" ofrecen una menor divergencia. Una posible explicación del distinto funcionamiento de estos conceptos de NSSI reside en el hecho de que los informes médicos tratan de reflejar los hechos más relevantes representados seguramente de una forma genérica y sin profundizar en determinados aspectos. De esta forma, los conceptos de "Razones", "Métodos" e "Instrumentos" implican un mayor detalle en la descripción del suceso del que se suele encontrar en un informe.

En la figura 3 se observa un gráfico de barras en el que aparecen las raíces de los términos más frecuentes del vocabulario de autolesiones ordenados por su frecuencia normalizada de aparición y en función de la clase a la que pertenecen. Como se puede observar, raíces como "autolesión", "cort" y "rasg" presentan una gran diferencia a favor de las clases positivas, mientras que otras raíces co-

Red Neuronal					
	F1 Todas Clases			Clase Positiva	
Modelo	**F1-NO**	**F1-SI**	***F1-weighted Avg***	**P-SI**	**R-SI**
Bert Multilingüe					
CVN + BERT (5ep,0.05do)	0.97	0.56	**0.95**	0.50	**0.64**
CVN + BERT (5ep,0.1do)	0.96	0.51	0.94	0.43	**0.64**
CVN + BERT (5ep,0.2do)	0.96	0.47	0.94	0.40	0.57
CVN + BERT (5ep,0.3do)	**0.98**	**0.57**	**0.95**	0.57	0.57
CVN + BERT (5ep,0.4do)	0.96	0.44	0.93	0.36	0.57
CVN + BERT (10ep,0.05do)	0.97	0.55	**0.95**	0.53	0.57
CVN + BERT (10ep,0.1do)	**0.98**	0.33	0.94	**0.75**	0.21
CVN + BERT (10ep,0.2do)	0.97	0.22	0.93	0.50	0.14
CVN + BERT (10ep,0.3do)	0.97	0.24	0.93	0.67	0.14
CVN + BERT (10ep,0.4do)	0.97	0.20	0.93	0.33	0.14
RoBERTa					
CVN + RoBERTa (5ep,0.05do)	0.96	0.39	0.92	0.43	0.35
CVN + RoBERTa (5ep,0.1do)	0.96	0.17	0.91	0.33	0.12
CVN + RoBERTa (5ep,0.2do)	0.96	0.37	0.93	0.50	0.29
CVN + RoBERTa (5ep,0.3do)	0.96	0.48	0.93	0.50	0.29
CVN + RoBERTa (5ep,0.4do)	0.96	0.54	0.94	0.50	0.59
CVN + RoBERTa (10ep,0.05do)	0.96	0.48	0.93	0.50	0.47
CVN + RoBERTa (10ep,0.1do)	0.96	0.41	0.93	0.50	0.35
CVN + RoBERTa (10ep,0.2do)	0.95	0.28	0.90	0.26	0.29
CVN + RoBERTa (10ep,0.3do)	0.97	0.55	**0.95**	0.67	0.47
CVN + RoBERTa (10ep,0.4do)	0.96	0.41	0.93	0.50	0.35

Tabla 4: F1-SI: F1-measure de los casos positivos, F1-NO: F1-measure de los casos negativos, F1-weighted Avg: F1-measure media de todo el conjunto de test, P-SI: Precisión de los casos positivos, R-SI: Recall de los casos positivos. Los sistemas varían en función del número de épocas (5-10 ep) y el dropout (0.05-0.4 do).

COMPARATIVA DE RESULTADOS					
Sistema	**F1 Todas Clases**			**Clase Positiva**	
	F1-NO	**F1-SI**	***F1-weighted Avg***	**P-SI**	**R-SI**
Baseline MFC	0.95	0.00	0.86	0.00	0.00
Baseline LFC	0.00	0.17	0.02	0.09	**1.00**
Gradient Boosting	0.97	**0.68**	0.94	**0.71**	**0.65**
CVN + BERT (5ep,0.3do)	**0.98**	0.57	**0.95**	0.57	0.57

Tabla 5: F1-SI: F1-measure de los casos positivos, F1-NO: F1-measure de los casos negativos, F1-weighted Avg: F1-measure media de todo el conjunto de test, P-SI: Precisión de los casos positivos, R-SI: Recall de los casos positivos

mo "tir", e "inger" muestran una aparición más equilibrada dado los conceptos más neutrales que representan en cuanto a las autolesiones. Destaca la raíz "sangr", teniendo más peso en la clase negativa debido seguramente a que las lesiones relacionadas con la sangre no son las más frecuentes en el problema estudiado.

6 Conclusiones y trabajo futuro

En este trabajo se presenta un sistema de detección de indicios de autolesiones no suicidas, basado en el análisis del lenguaje, sobre un conjunto anotado de informes médicos obtenidos del servicio de psiquiatría de un Hospital público madrileño. Dada la naturaleza tan crítica de la tarea y las implicaciones a

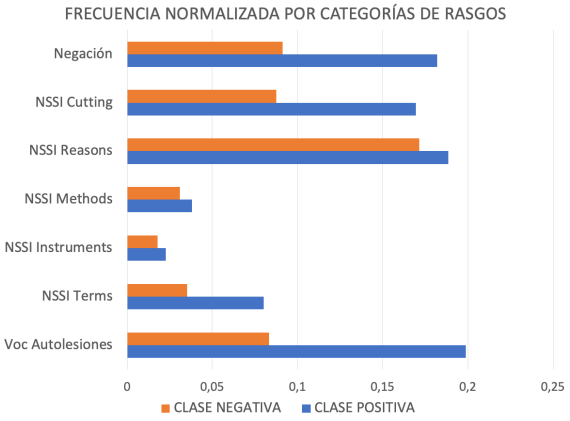

Figura 2: Frecuencia normalizada de las diferentes categorías de rasgos en función de la clase.

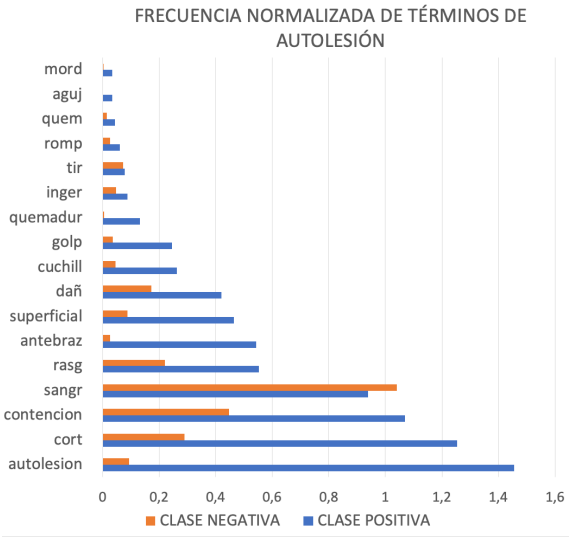

Figura 3: Frecuencia normalizada de términos de autolesiones en función de la clase.

la hora de predecir incorrectamente un caso, que realmente tiene detrás a un ser humano real, es necesario tratar este tipo de trabajos desde un punto de vista diferente al mero resultado obtenido por un conjunto de métricas de evaluación. Al inicio del trabajo se fijaron tres objetivos prioritarios: por un lado el sistema debería obtener un alto rendimiento para impedir en la medida de lo posible las predicciones erróneas, por otro lado la clasificación correcta de los casos positivos debería ser prioritaria, y finalmente se debería buscar la mayor explicabilidad del sistema para que el profesional sanitario pudiera tener la máxima información de cara a tomar una decisión final. Teniendo en cuenta estos requisitos, el trabajo ha cumplido con los objetivos. Por un lado el rendimiento global obtenido alcanza unos valores de medida-F de 0.95, alcanzando un 0.68 de medida-F para los casos positivos, lo cual es una prueba de su buen funcionamiento. Con la extracción de un conjunto de rasgos muy focalizados en alcanzar un mayor rendimiento en cuanto a la detección de los casos positivos, se ha conseguido el segundo objetivo equilibrando y mejorando tanto la precisión como la cobertura de forma significativa. Y finalmente, gracias al esfuerzo realizado para equiparar un sistema supervisado basado en algoritmos tradicionales de clasificación a un sistema basado en redes neuronales, se ha hecho posible el poder elegir el primero de los sistemas ya que con un rendimiento global similar tiene dos ventajas como son el mejor rendimiento en cuanto a la clasificación de casos positivos y una mejor explicabilidad debido a que los rasgos obtenidos forman parte de la decisión tomada finalmente por el sistema. En este último caso, el sistema basado en redes neuronales, a pesar de tener un ligero mejor rendimiento global, obtiene peores resultados en la clase positiva y además sus decisiones a día de hoy son difíciles de explicar de cara a un psicólogo o psiquiatra.

En cuanto al trabajo futuro, consideramos varias líneas de actuación. Por un lado el corpus está desbalanceado y además no dispone de un gran número de casos positivos. De esta forma trabajaremos para obtener un corpus de mayor tamaño con la esperanza de que un mayor número de casos positivos nos ayude a mejorar aún más la detección de este tipo de casos. Por otro lado, debido al gran potencial de tecnologías como las redes neuronales, trabajaremos para optimizar el sistema e incluir transformers con los objetivos de mejorar su rendimiento en cuanto a los casos positivos e iniciar un trabajo de estudio para mejorar la explicabilidad de este tipo de sistemas.

Agradecimientos

This work has been partially supported by the Spanish Ministry of Science and Innovation within the DOTT-HEALTH Project (MCI/AEI/FEDER, UE) under Grant PID2019-106942RB-C32 and the project RAICES (IMIENS 2022).

Bibliografía

Ageitos, E. C., J. Martínez-Romo, y L. Araujo. 2020. Nlp-uned at erisk 2020: Self-harm early risk detection with sentiment

analysis and linguistic features. En *CLEF (Working Notes)*.

Baetens, I., L. Claes, J. Muehlenkamp, H. Grietens, y P. Onghena. 2011. Non-Suicidal and Suicidal Self-Injurious Behavior among Flemish Adolescents: A Web-Survey. *Archives of Suicide Research*, 15(1):56–67.

Burke, T. A., B. A. Ammerman, y R. Jacobucci. 2019. The use of machine learning in the study of suicidal and non-suicidal self-injurious thoughts and behaviors: A systematic review. *Journal of affective disorders*, 245:869–884.

Campillo-Ageitos, E., H. Fabregat, L. Araujo, y J. Martinez-Romo. 2021. Nlp-uned at erisk 2021: self-harm early risk detection with tf-idf and linguistic features. *Working Notes of CLEF*, páginas 21–24.

Delgado-Gomez, D., E. Baca-Garcia, D. Aguado, P. Courtet, y J. Lopez-Castroman. 2016. Computerized adaptive test vs. decision trees: development of a support decision system to identify suicidal behavior. *Journal of affective disorders*, 206:204–209.

Devlin, J., M.-W. Chang, K. Lee, y K. Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Fabregat, H., L. Araujo Serna, y J. Martínez Romo. 2019. Deep learning approach for negation trigger and scope recognition.

Fabregat, H., A. Duque, J. Martínez-Romo, y L. Araujo. 2019. Extending a deep learning approach for negation cues detection in spanish. En *IberLEF@ SEPLN*, páginas 369–377.

Fernandes, A. C., R. Dutta, S. Velupillai, J. Sanyal, R. Stewart, y D. Chandran. 2018. Identifying suicide ideation and suicidal attempts in a psychiatric clinical research database using natural language processing. *Scientific reports*, 8(1):1–10.

Greaves, M. M. 2018a. *A Corpus Linguistic Analysis of Public Reddit and Tumblr Blog Posts on Non-Suicidal Self-Injury, An abstract*. Ph.D. tesis, College of Education, Oregon State University.

Greaves, M. M. 2018b. A corpus linguistic analysis of public reddit and tumblr blog posts on non-suicidal self-injury.

Haerian, K., H. Salmasian, y C. Friedman. 2012. Methods for identifying suicide or suicidal ideation in ehrs. En *AMIA annual symposium proceedings*, volumen 2012, página 1244. American Medical Informatics Association.

Kessler, R. C., M. B. Stein, M. V. Petukhova, P. Bliese, R. M. Bossarte, E. J. Bromet, C. S. Fullerton, S. E. Gilman, C. Ivany, L. Lewandowski-Romps, y others. 2017. Predicting suicides after outpatient mental health visits in the army study to assess risk and resilience in servicemembers (army starrs). *Molecular psychiatry*, 22(4):544–551.

LeCun, Y., Y. Bengio, y G. Hinton. 2015. Deep learning. *nature*, 521(7553):436–444.

Losada, D. E., F. Crestani, y J. Parapar. 2019. Overview of erisk 2019 early risk prediction on the internet. En *International Conference of the Cross-Language Evaluation Forum for European Languages*, páginas 340–357. Springer.

Losada, D. E., F. Crestani, y J. Parapar. 2020. erisk 2020: Self-harm and depression challenges. En *European Conference on Information Retrieval*, páginas 557–563. Springer.

Loyola, J. M., S. Burdisso, H. Thompson, L. Cagnina, y M. Errecalde. 2021. Unsl at erisk 2021: A comparison of three early alert policies for early risk detection. En *Working Notes of CLEF 2021-Conference and Labs of the Evaluation Forum, Bucarest, Romania*.

Mann, J. J., S. P. Ellis, C. M. Waternaux, X. Liu, M. A. Oquendo, K. M. Malone, B. S. Brodsky, G. L. Haas, y D. Currier. 2008. Classification trees distinguish suicide attempters in major psychiatric disorders: a model of clinical decision making. *The Journal of clinical psychiatry*, 69(1):2693.

Martınez-Castano, R., A. Htait, L. Azzopardi, y Y. Moshfeghi. 2020. Early risk detection of self-harm and depression severity using bert-based transformers. *Working Notes of CLEF*, página 16.

McCoy, T. H., V. M. Castro, A. M. Roberson, L. A. Snapper, y R. H. Perlis. 2016. Improving prediction of suicide and accidental death after discharge from general hospitals with natural language processing. *JAMA psychiatry*, 73(10):1064–1071.

Metzger, M.-H., N. Tvardik, Q. Gicquel, C. Bouvry, E. Poulet, y V. Potinet-Pagliaroli. 2017. Use of emergency department electronic medical records for automated epidemiological surveillance of suicide attempts: a french pilot study. *International journal of methods in psychiatric research*, 26(2):e1522.

Nicolai, K. A., M. D. Wielgus, y A. Mezulis. 2016. Identifying Risk for Self-Harm: Rumination and Negative Affectivity in the Prospective Prediction of Nonsuicidal Self-Injury. *Suicide and Life-Threatening Behavior*, 46(2):223–233.

Obeid, J. S., J. Dahne, S. Christensen, S. Howard, T. Crawford, L. J. Frey, T. Stecker, y B. E. Bunnell. 2020. Identifying and predicting intentional self-harm in electronic health record clinical notes: deep learning approach. *JMIR medical informatics*, 8(7):e17784.

Obeid, J. S., E. R. Weeda, A. J. Matuskowitz, K. Gagnon, T. Crawford, C. M. Carr, y L. J. Frey. 2019. Automated detection of altered mental status in emergency department clinical notes: a deep learning approach. *BMC medical informatics and decision making*, 19(1):1–9.

Parapar, J., P. Martín-Rodilla, D. E. Losada, y F. Crestani. 2021. Overview of erisk 2021: Early risk prediction on the internet. En *International Conference of the Cross-Language Evaluation Forum for European Languages*, páginas 324–344. Springer.

Pennebaker, J. W., M. R. Mehl, y K. G. Niederhoffer. 2003. Psychological aspects of natural language use: Our words, our selves. *Annual review of psychology*, 54(1):547–577.

Poulin, C., B. Shiner, P. Thompson, L. Vepstas, Y. Young-Xu, B. Goertzel, B. Watts, L. Flashman, y T. McAllister. 2014. Predicting the risk of suicide by analyzing the text of clinical notes. *PloS one*, 9(1):e85733.

Rodham, K., K. Hawton, y E. Evans. 2004. Reasons for deliberate self-harm: Comparison of self-poisoners and self-cutters in a community sample of adolescents. *Journal of the American Academy of Child & Adolescent Psychiatry*, 43(1):80–87.

Rozova, V., K. Witt, J. Robinson, Y. Li, y K. Verspoor. 2022. Detection of self-harm and suicidal ideation in emergency department triage notes. *Journal of the American Medical Informatics Association*, 29(3):472–480.

Walsh, C. G., J. D. Ribeiro, y J. C. Franklin. 2017. Predicting risk of suicide attempts over time through machine learning. *Clinical Psychological Science*, 5(3):457–469.

Young, R., M. Van Beinum, H. Sweeting, y P. West. 2007. Young people who self-harm. *The British Journal of Psychiatry*, 191(1):44–49.