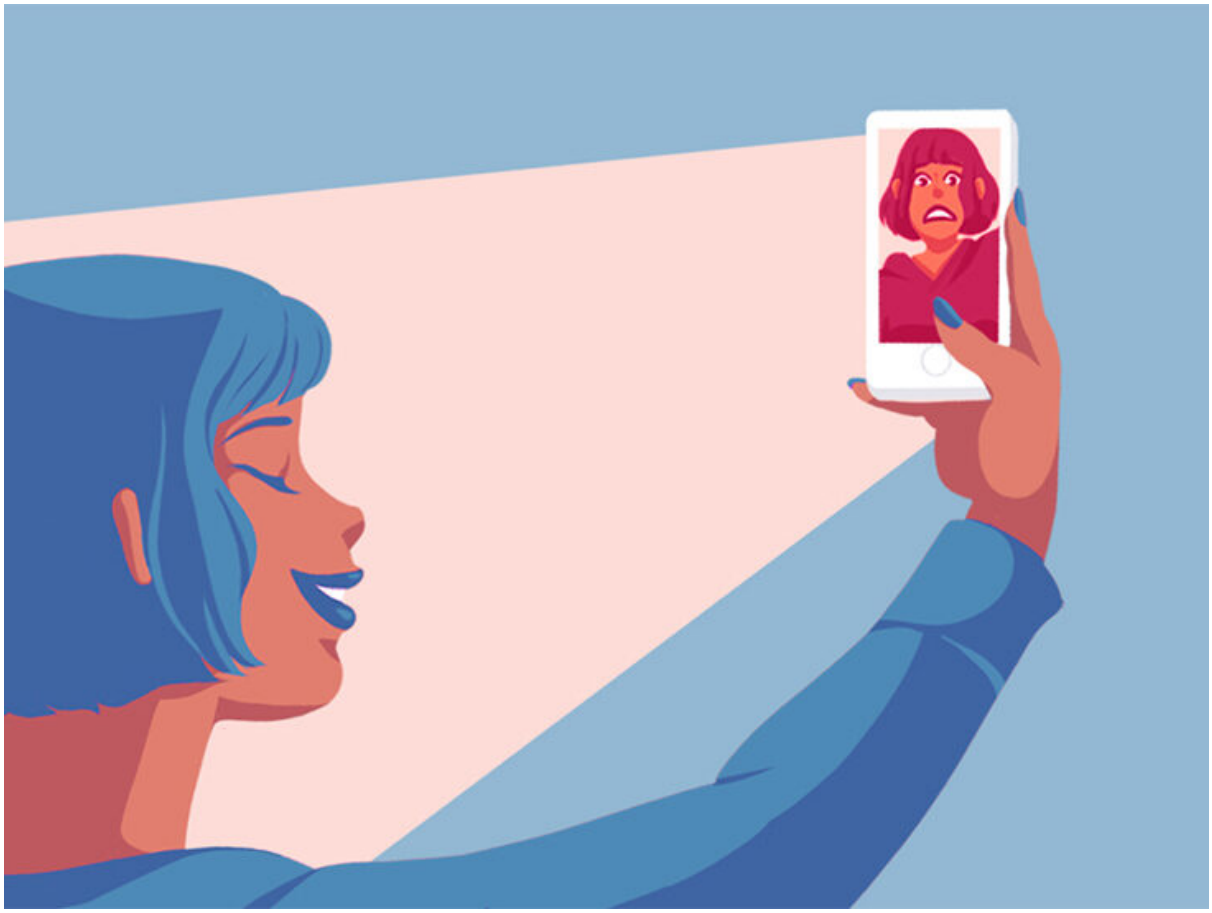
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA
ESCOLA TÉCNICA SUPERIOR DE ENXEÑARÍA



Social Big Data y enfermedad mental: aplicación automática de nuevas técnicas discursivas a colecciones de redes sociales en contextos de enfermedades mentales.

Autor: Raúl Alberto Barrantes Pampillo
Tutores: Patricia Marín Rodilla, David Losada Carril

Máster universitario en Tecnologías de Análisis de Datos Masivos: Big Data
Trabajo de Fin de Master
Febrero 2021

Resumen:

Los textos recopilados de redes sociales han demostrado en investigaciones recientes ser una fuente complementaria para el estudio y detección de riesgos tempranos de padecer enfermedad mental. [1]
Estas diferencias en el lenguaje natural usado en redes sociales han mostrado resultados satisfactorios a nivel léxico y/o gramatical, estudiándose mucho menos el nivel discursivo.
Centrándonos en detección de depresión y a nivel discursivo, puede haber diferencias entre textos escritos por personas con diagnóstico no depresivo y los escritos por personas con diagnóstico depresivo. Aplicando paradigmas de formalización del discurso, como puede ser RST (Rhetorical Structure Theory) [19], se pueden evaluar los patrones en coherencia en común de esos textos.
Por medio de este paradigma, se representan los textos por medio de estructuras jerárquicas llamadas Árboles de Discurso. Las hojas de un árbol de discurso corresponden a unidades de texto atómicas (EDUs). Las EDUs son unidades que funcionan como bloques de construcción. Los EDUs adyacentes son conectadas por relaciones de coherencia (Elaboración, Contraste…) formando unidades de discurso más grandes (representados por nodos internos).
Estos patrones o pistas lingüísticas pueden servir de indicador de una patología mental que podrían facilitar o advertir el diagnóstico de tales enfermedades.
Siguiendo esta hipótesis de investigación, este TFM realiza un análisis Big Data de colecciones de datos textuales desde redes sociales con el objetivo de analizar de forma automática las estructuras discursivas surgidas de los mismos e inferir posibles patrones discursivos de relevancia en enfermedad mental.
Tomando diferentes colecciones de texto obtenidas de redes sociales [1], se realiza tanto un análisis general como un estudio de coherencia por medio del framework CODRA [7], el cual es un framework que hace un análisis discursivo por medio de la generación de árboles RST de cada uno de los textos que se encuentran en el corpus. En base a esos árboles, se analiza su estructura y los patrones en común entre otras mismas estructuras dentro de los corpus.
Nuestro estudio permitió el análisis de 45026 posts de la colección de datos eRisk que comprende 1076582 entradas correspondientes a 887 usuarios [1]. En base a estos análisis se pudieron detectar varios patrones, entre ellos la presencia de textos más extensos entre el corpus relacionado con personas con diagnóstico de depresión.

Tabla de Contenidos

1. Problema a resolver *4*

2. Métricas *4*

2.1 Rhetorical Structure Theory **4**

3. Herramientas/Estado del Arte *5*

3.1 Colecciones de Texto **5**

3.2 CODRA **5**

3.3 Gensim **5**

4. Proceso de análisis *6*

5. Solución: Estudio Big Data *6*

5.1 Análisis descriptivo de los datos **6**

5.1.1 Conteo de las palabras en el texto (rango de 1 a 100, sin contar los datos que están vacíos). 7

5.1.2 Conteo de caracteres en el texto (rango de 1 a 1000, sin contar los datos que están vacíos). 7

5.1.3 Promedio de caracteres por palabra en el texto (mayores que 0). 8

5.1.4 Número de stopwords en el texto (cuando el texto no es vacío) 8

5.1.5 Porcentaje de stopwords en el texto (cuando el texto no es vacío) 9

5.2 Estadísticas de análisis de coherencia **9**

5.2.1 Número de frases 9

5.2.2 Conteo de frases de elaboration. 10

5.2.3 Conteo de frases de enablement. 10

5.2.4 Conteo de frases de contrast. 11

5.2.5 Conteo de frases de núcleo que son leaf. 11

5.2.6 Conteo de frases de satélite que son leaf. 12

5.2.7 Conteo de frases de núcleo que son span 12

5.2.8 Conteo de frases de satélite que son span. 13

5.3 Palabras más frecuentes y centrales **13**

5.4 Análisis de patrones en los textos procesados **14**

5.4.1 Características de textos no analizados 14

5.4.2 Patrones en los textos 15

5.4.3 Tipos de patrones en los textos 15

5.4.5 Cálculo de la profundidad de los textos 16

6. Conclusiones *16*

7. Referencias *16*

1. Problema a resolver

En la década del 2010 en la cual han surgido redes sociales como Facebook, Twitter, Reddit e Instagram, también se ha notado un incremento notable de los diagnósticos de depresión. La falta de un tratamiento puede llevar a ansiedad, episodios psicóticos, depresión y en el peor de los casos, autolesiones o al suicidio.
Muchas personas usan las redes sociales para compartir sus pensamientos y sentimientos, como extensión de su vida social, personal y familiar. Debido a esto, sus contribuciones en redes sociales en forma de posts o comentarios en lenguaje natural, pueden reflejar una condición mental. Estos textos han sido analizados recientemente por investigadores en busca de indicadores que permitan encontrar patrones y factores diferenciales para la detección temprana de enfermedad mental, especialmente y debido a la gran cantidad de casos diagnosticados en el mundo, en el caso de la depresión, y así poder compararlos con textos de personas que no tienen tal diagnóstico.
Estos análisis aplican métricas y patrones discursivos como factor diferencial e indicativo de textos producidos por personas diagnosticadas con depresión, comparándolos con población no diagnosticada.
En base a esto, el objetivo es analizar las estructuras discursivas surgidas de los textos a gran escala y de esta forma, inferir posibles patrones discursivos de relevancia en enfermedad mental.
Primeramente, dentro de los textos en general se puede medir la coherencia por medio de la frecuencia de palabras dentro de las oraciones, mediante lo cual se asigna un índice de coherencia discursiva [3]. Por otro lado, existen palabras o tópicos que son más comunes dentro de los encontrados en los foros de redes sociales relacionados con salud mental [2, 11, 12, 13]. Utilizando elementos procedentes de este tipo de análisis se plantea evaluar factores en común entre los textos escritos por los diagnosticados por depresión.
Tomando como base del trabajo diferentes de textos de prueba [1] y varios estudios que se han hecho al respecto [7, 14, 19], se pretende implementar y comparar una serie de métricas de coherencia discursiva con el fin de determinar qué elementos puede ser diferenciadores de las personas que sufren trastornos psicológicos como la depresión [9].

2. Métricas

2.1 Rhetorical Structure Theory

La teoría de discurso utilizada es la Rhetorical Structure Theory (RST), en la cual se representan los textos por medio de estructuras jerárquicas llamadas Árboles de Discurso. Este es el paradigma más formal y computacional de discurso que hay en la actualidad. Las hojas de un árbol de discurso corresponden a **unidades de texto atómicas (EDUs).** Las EDUs son unidades que funcionan como bloques de construcción. Los EDUs adyacentes, son conectadas por relaciones de coherencia (*Elaboration*, *Contrast*,) formando unidades de discurso más grandes (representados por nodos internos). Unidades de discurso enlazadas por una relación retórica son distinguidas por su importancia en el texto: el **núcleo** consiste en la parte principal de la relación, así como los **satélites** son partes periféricas de la oración. Asimismo, tanto estructuras de núcleo como de satélite pueden ser **contenedores** de otras estructuras -llamadas **span**-, o bien hojas que abarcan estructuras enteras -llamadas **leaf**-. Asimismo, las relaciones entre las unidades de discurso pueden entrar en cada una de estas categorías: *Elaboration, Circumstance, Solutionhood, Volitional Cause, Conditional Result, Non-Volitional Cause, Non-Volitional Result, Purpose, Condition, Otherwise, Interpretation, Evaluation,*

Restatement, Summary, Sequence, Contrast, Motivation, Antithesis, Background, Enablement, Evidence, Justify, Concession. [19].
Las relaciones que vamos a usar para este trabajo son las siguientes:
***Elaboration*:** Es una oración con una agregación de detalle a algo que ya ha sido dicho o escrito.
Ejemplo: *A tall man came by;* ***he was wearing an old navy jacket.***
***Contrast*:** Una relación de contraste es una relación interproposicional que expresa que una diferencia entre una proposición y otra es relevante.
Ejemplo: *Animals heal,* ***but trees compartmentalize.***
***Enablement*:** Una relación de habilitación es una relación interproposicional en la que una proposición (es) apoya la directiva de un hablante al mejorar la capacidad del destinatario para cumplirla o aprovecharla.
Ejemplo: *It feels like an impossible task* ***to find scholarships for a non-minority, non-STEM graduate student.***

3. Herramientas/Estado del Arte

3.1 Colecciones de Texto

Para el desarrollo del proyecto se van a usar como base colecciones de texto para estudios de depresión, se trata de publicaciones (posts o comentarios) realizadas por un conjunto de personas en la red social Reddit. Estos textos fueron recopilados en el marco la iniciativa de investigación eRisk, la cual se ha llevado a cabo en cuatro ediciones desde el 2017 hasta el 2020. [21, 22, 23, 24] Esta es una de las primeras iniciativas en reunir a muchos investigadores para estudiar la interacción entre el lenguaje y los trastornos mentales en las redes sociales con el propósito de abordar la detección temprana de la depresión de forma automática . Se han publicado corpus de textos de entrenamiento donde se dispone de textos procedentes de personas con diagnóstico de depresión como de personas de un grupo de control (sin diagnóstico de depresión). Los datos de los usuarios de la clase positiva (es decir, con depresión) se recopilaron por medio de la búsqueda de autoexpresiones de diagnósticos de depresión (por ejemplo, la oración "Me diagnosticaron depresión") y verificando si realmente contenían una declaración de diagnóstico – según el método propuesto por Coppersmith [18]. Los usuarios no deprimidos se recopilaron mediante un muestreo aleatorio del gran conjunto de usuarios disponibles en la plataforma [1].

3.2 CODRA

El método usado para análisis de texto es CODRA (COmplete probabilistic Discriminative framework for performing Rhetorical Analysis) propuesto por Joty [7], en el cual se hace un análisis sintáctico de los textos, y se elabora una estructura en forma de árbol en el cual se puede mostrar la coherencia entre los textos. CODRA comprende un segmentador de discurso y un analizador de discurso. Primero, se realiza una segmentación de discurso, que se basa en un clasificador binario, el cual identifica las unidades elementales del discurso en un texto dado. Luego, el analizador de discurso construye un árbol de discurso RST aplicando un algoritmo de parseo. Así, para cada texto fuente analizado, tendremos un árbol RST con la estructura formal discursiva del texto.

3.3 Gensim

Gensim hace posible análisis en textos usando un Modelo de Espacio Vectorial y técnicas de análisis de tópicos, con el que se pueden resumir los diferentes elementos (palabras,

oraciones…) que se pueden encontrar en los documentos. Para tal análisis se usa el método de modelado por tópicos, en el que los textos en lenguaje natural se puede asociar a un número de conceptos (o temas) subyacentes. Los métodos de modelado por tópicos implementados han sido el Latent Semantic Analysis (LSA), Latent Dirichlet Allocation (LDA) [17] y también se han analizado los términos por medio de TF-IDF, que pondera la relevancia de un término dentro del corpus [25].

4. Proceso de análisis

Para el desarrollo del proyecto se usaron como base dos grandes colecciones de texto para estudios de depresión [1].
El proceso de análisis consistió en 2 fases. En primer lugar, se extrajeron características como número de palabras, número de caracteres, número de stopwords, promedio de caracteres por palabra, entre otros factores, así mismo como porcentajes relacionados con estas estadísticas.
Una segunda parte consistió en la generación de análisis de coherencia, por medio de CODRA [2], el cual es un framework en el cual se generaron árboles con la estructura semántica de cada entrada. La naturaleza de procesamiento es pesada y se realiza por medio de un contenedor docker es muy alta. Para este proceso se necesita un mínimo de 6 GB de RAM para su correcto procesamiento, y si se usa en estas condiciones mínimas el tratamiento de cada dato dura tres minutos – para lo cual es tratamiento de un corpus de mas de 1000000 de datos es inviable. Debido a esto, se hizo uso de la infraestructura del Centro de Supercomputación de Galicia (CESGA) [26] . En concreto, por medio del cluster Finisterrae se hizo uso de la instalación del contenedor de CODRA. El CESGA soporta únicamente Singularity que es una solución que replica el entorno basado en contenedores que utiliza Docker [27].
Usando este entorno se tuvo la capacidad de procesar 55892 datos del grupo de control y 41104 datos del clase positiva, de los cuales se pudieron extraer árboles de coherencia de 26797 y 18229 datos respectivamente de cada corpus.
En base a los árboles resultantes, se extrajeron varios tipos de estadísticas para cada corpus, las cuales señalan la estructura semántica que extrae el parser de codra entre las cuales están; la presencia de frases de núcleo, de contraste, de satélite, de habilitación, entre otros.
En una segunda parte de este análisis, también se extrajeron cuales son las estructuras semánticas más comunes dentro de las oraciones y, adicionalmente, se estudió el nivel de profundidad de los árboles de discurso resultantes de cada corpus de oraciones.
Debido a la alta tasa de oraciones procesadas que no resultaron en un árbol de coherencia, se analizaron de nuevo estas oraciones fallidas para poder mirar las causas y razones por las que esas entradas textuales no pudieron ser procesadas.

5. Solución: Estudio Big Data

Una vez procesado el texto, se hicieron varios tipos de análisis de datos: primero se hace un análisis descriptivo en el cual se analizan aspectos generales de los corpus de texto, así se analizan estructuradas generadas por CODRA, y se analizan los textos que no pudieron ser analizados por el parser. En la misma línea se hacen los análisis generados por Gensim. En el siguiente informe de resultados, las categorías que tienen mayor porcentaje que están en cada corpus, se presentan en **negrita**.

5.1 Análisis descriptivo de los datos

Se hace un análisis descriptivo de los datos dentro del clase positiva (90222 entradas) y el grupo de control (986360 entradas). Dentro de los análisis que se hacen en el texto se encuentran; el

conteo de caracteres, conteo de palabras, promedio de caracteres, número de stopwords en el texto, porcentaje de stopwords en el texto.

5.1.1 Conteo de las palabras en el texto (rango de 1 a 100, sin contar los datos que están vacíos).

Se cuentan las palabras en cada oración, los resultados son semejantes en ambos corpus.

	Conteo Positivos:	**Porcentaje Positivos:**	**Conteo Control:**	**Porcentaje Control:**
< 10	21367	23.70%	252662	27.20%
10-20	**32013**	**35.50%**	**342605**	**36.80%**
20-30	12669	14.00%	127038	13.70%
30-40	7610	8.40%	72711	7.80%
40-50	5167	5.70%	47363	5.10%
50-60	3559	3.90%	29923	3.20%
60-70	2735	3.00%	20518	2.20%
70-80	2098	2.30%	15424	1.70%
80-90	1673	1.90%	12395	1.30%
90-100	1331	1.50%	9317	1.00%

5.1.2 Conteo de caracteres en el texto (rango de 1 a 1000, sin contar los datos que están vacíos).

Se cuentan los caracteres en cada oración, los resultados son semejantes en ambos corpus.

	Conteo Positivos:	**Porcentaje Positivos:**	**Conteo Control:**	**Porcentaje Control:**
< 25	10439	11.60%	104081	10.60%
25-50	15255	16.90%	176554	17.90%
50-75	11929	13.20%	188881	19.20%
75-100	8749	9.70%	118222	12.00%
100-200	**18726**	**20.80%**	**192659**	**19.50%**
200-300	8401	9.30%	80794	8.20%
300-400	4787	5.30%	37400	3.80%
400-500	3022	3.30%	22771	2.30%
500-600	2038	2.30%	14897	1.50%

5.1.3 Promedio de caracteres por palabra en el texto (mayores que 0).

Se cuentan el promedio de caracteres que se encuentran en ambos corpus. Se puede ver que la diferencia se en cuenta que la mayoría de las oraciones en el corpus de textos producidos por diagnosticados positivos de depresión tienen un promedio de 4 caracteres, pero los textos producidos por el grupo de control tiene un promedio de 5 a 10 caracteres.

	Conteo Positivos:	**Porcentaje Positivos:**	**Conteo Control:**	**Porcentaje Control:**
2	379	0.40%	72	0.00%
3	4387	4.90%	857	0.10%
4	**46534**	**51.60%**	214773	26.20%
5	18825	20.90%	220150	26.90%
5-10	18298	20.30%	**358237**	**43.80%**
10-15	817	0.90%	16288	2.00%

5.1.4 Número de stopwords en el texto (cuando el texto no es vacío).

Stopwords es el nombre que reciben las palabras sin significado como artículos, pronombres, preposiciones, etc. que son filtradas antes o después del procesamiento de datos en lenguaje natural (texto). Se cuentan el número de stopwords en el texto que se encuentran dentro de ambos corpus. Se pueden ver algunas diferencias entre los porcentajes de número de stopwords entre ambos corpus, pero tienen indíces semejantes.

	Conteo Positivos:	**Porcentaje Positivos:**	**Conteo Control:**	**Porcentaje Control:**
0	10712	11.9%	**152536**	**15.5%**
1	7194	8.0%	113691	11.5%
2	7115	7.9%	106011	10.7%
3	6299	7.0%	82498	8.4%
4	5507	6.1%	63711	6.5%
5	4713	5.2%	51333	5.2%
6	4021	4.5%	42438	4.3%
7	3489	3.9%	35489	3.6%
8	3138	3.5%	30552	3.1%
9	2756	3.1%	26600	2.7%
10-20	**15831**	**17.5%**	146104	14.8%
20-30	7064	7.8%	54176	5.5%
30-40	3844	4.3%	27436	2.8%
40-50	2361	2.6%	15778	1.6%
50-60	1519	1.7%	10081	1.0%

5.1.5 Porcentaje de stopwords en el texto (cuando el texto no es vacío)

En base al número de stopwords en el texto que se encuentran dentro del corpus de casos positivos y negativos, se calculan los porcentajes, los cuales son semejantes dentro de ambos corpus.

	Conteo Positivos:	**Porcentaje Positivos:**	**Conteo Control:**	**Porcentaje Control:**
<10	11806	13.10%	187111	19.00%
10-20	8924	9.90%	141601	14.40%
20-30	17904	19.90%	212967	21.60%
30-40	**29045**	**32.20%**	**270228**	**27.40%**
40-50	20009	22.20%	154212	15.60%
50-60	2446	2.70%	19260	2.00%
60-70	56	0.10%	683	0.10%

5.2 Estadísticas de análisis de coherencia

A continuación las estadísticas en base a los árboles RST que son generados por el parser CODRA:

5.2.1 Número de frases.

Se cuentan las frases que se encuentran dentro del corpus de casos positivos y negativos. Se puede ver que el corpus de oraciones de casos positivas tienen más frases que las oraciones presentes dentro del grupo de control.

	Conteo Positivos:	**Porcentaje Positivos:**	**Conteo Control:**	**Porcentaje Control:**
2	3041	16.70%	**7219**	**26.90%**
3	2964	16.30%	4983	18.60%
4	2330	12.80%	3309	12.30%
5	1606	8.80%	2285	8.50%
6	1280	7.00%	1742	6.50%
7	973	5.30%	1246	4.60%
8	756	4.10%	951	3.50%
9	637	3.50%	737	2.80%
10	512	2.80%	596	2.20%
> 10	**4130**	**22.70%**	3729	13.90%

5.2.2 Conteo de frases de elaboration.

Se cuentan las frases de *elaboration* que se encuentran dentro del corpus de casos positivos y negativos, los resultados son semejantes dentro de ambos corpus.

	Conteo Positivos:	**Porcentaje Positivos:**	**Conteo Control:**	**Porcentaje Control:**
0	2833	15.50%	5006	18.70%
1	**4631**	**25.40%**	**8719**	**32.50%**
2	3124	17.10%	4594	17.10%
3	2078	11.40%	2639	9.80%
4	1351	7.40%	1748	6.50%
5	921	5.10%	1101	4.10%
6	728	4.00%	707	2.60%
7	502	2.80%	535	2.00%
8	450	2.50%	404	1.50%
> 10	1083	5.90%	838	3.10%

5.2.3 Conteo de frases de enablement.

Se cuentan las frases de *enablement* que se encuentran dentro del corpus de casos positivos y negativos, los resultados son semejantes dentro de ambos corpus.

	Conteo Positivos:	**Porcentaje Positivos:**	**Conteo Control:**	**Porcentaje Control:**
0	**15898**	**87.20%**	**24106**	**90.00%**
1	1956	10.70%	2383	8.90%
2	300	1.60%	241	0.90%
3	51	0.30%	54	0.20%
4	17	0.10%	6	0.00%
5	4	0.00%	5	0.00%
6	1	0.00%	2	0.00%
7	1	0.00%		
8	1	0.00%		

5.2.4 Conteo de frases de contrast.

Se cuentan las frases de *contrast* que se encuentran dentro del corpus de casos positivos y negativos, los resultados son semejantes dentro de ambos corpus.

	Conteo Positivos:	**Porcentaje Positivos:**	**Conteo Control:**	**Porcentaje Control:**
0	**13048**	**71.60%**	**21682**	**80.90%**
1	2181	12.00%	2236	8.30%
2	2020	11.10%	2143	8.00%
3	487	2.70%	413	1.50%
4	244	1.30%	187	0.70%
5	107	0.60%	61	0.20%
6	66	0.40%	42	0.20%
7	33	0.20%	14	0.10%
8	14	0.10%	6	0.00%
9			7	0.00%
10	19	0.10%		

5.2.5 Conteo de frases de núcleo que son leaf.

Se cuentan las frases de núcleo que son *leaf* dentro del RST (que solo abarcan una frase) que se encuentran dentro de los corpus de casos positivos y negativos, los resultados son semejantes dentro de ambos corpus, excepto que en el corpus de textos producidos por personas con diagnóstico positivo de depresión tienen un porcentaje levemente mayor de oraciones con más de 10 frases de núcleo que son *leaf*.

	Conteo Positivos:	**Porcentaje Positivos:**	**Conteo Control:**	**Porcentaje Control:**
0	**11993**	**65.80%**	**20050**	**74.80%**
1	1485	8.10%	2022	7.50%
2	1463	8.00%	1579	5.90%
3	992	5.40%	962	3.60%
4	546	3.00%	662	2.50%
5	385	2.10%	350	1.30%
6	299	1.60%	293	1.10%
7	182	1.00%	191	0.70%
8	141	0.80%	149	0.60%
> 10	515	2.80%	371	1.40%

5.2.6 Conteo de frases de satélite que son leaf.

Se cuentan las frases de satélite que son *leaf* del RST (que solo abarcan una frase) que se encuentran dentro de los corpus de casos positivos y negativos, los resultados son semejantes dentro de ambos corpus, los resultados son semejantes dentro de ambos corpus, excepto que en el corpus de textos producidos por personas con diagnóstico positivo de depresión tienen un porcentaje levemente mayor de oraciones con más de 10 frases de núcleo que son *leaf*.

	Conteo Positivos:	**Porcentaje Positivos:**	**Conteo Control:**	**Porcentaje Control:**
0	**12417**	**68.10%**	**20414**	**76.20%**
1	2223	12.20%	2791	10.40%
2	1473	8.10%	1491	5.60%
3	746	4.10%	790	2.90%
4	403	2.20%	452	1.70%
5	246	1.30%	285	1.10%
6	175	1.00%	181	0.70%
7	121	0.70%	141	0.50%
8	99	0.50%	59	0.20%
> 10	204	1.10%	104	0.40%

5.2.7 Conteo de frases de núcleo que son span.

Se cuentan las frases de núcleo que son *span* del RST (que abarcan varias frases) que se encuentran dentro de los corpus de casos positivos y negativos, los resultados son semejantes dentro de ambos corpus, excepto que en el corpus de textos producidos por personas con diagnóstico positivo de depresión tienen un porcentaje mayor de oraciones con más de 10 frases de núcleo que son *span* y que en el grupo de control tiene un porcentaje mayor de oraciones que tienen 0 frases de núcleo que son *span* que la clase positiva.

	Conteo Positivos:	**Porcentaje Positivos:**	**Conteo Control:**	**Porcentaje Control:**
0	**4328**	**23.70%**	**9503**	**35.50%**
1	4018	22.00%	6035	22.50%
2	2479	13.60%	3293	12.30%
3	1595	8.70%	2135	8.00%
4	1075	5.90%	1388	5.20%
5	817	4.50%	958	3.60%
6	634	3.50%	688	2.60%
7	510	2.80%	524	2.00%
8	389	2.10%	365	1.40%
> 10	1805	9.90%	1360	5.10%

5.2.8 Conteo de frases de satélite que son span.

Se cuentan las frases de satélite que son *span* del RST (que abarcan varias frases) que se encuentran dentro de los corpus de casos positivos y negativos, excepto que en el corpus de textos producidos por personas con diagnóstico positivo de depresión tienen un porcentaje mayor de oraciones con más de 10 frases de satélite que son *span* y que en el grupo de control tiene un porcentaje mayor de oraciones que tienen 0 frases de satélite que son *span* que la clase positiva.

	Conteo Positivos:	**Porcentaje Positivos:**	**Conteo Control:**	**Porcentaje Control:**
0	**5977**	**32.80%**	**11762**	**43.90%**
1	4307	23.60%	6418	24.00%
2	2423	13.30%	3224	12.00%
3	1522	8.30%	1747	6.50%
4	933	5.10%	1098	4.10%
5	664	3.60%	660	2.50%
6	508	2.80%	479	1.80%
7	366	2.00%	345	1.30%
8	240	1.30%	252	0.90%
> 10	919	5.00%	547	2.00%

5.3 Palabras más frecuentes y centrales

Usando la herramienta de análisis de textos Gensim, que usa el Modelo de Espacio Vectorial para categorizar las palabras dentro de las oraciones, y verificar y frecuencia e importancia. Por ejemplo una de las oraciones analizadas dice lo siguiente: *My favorite snack (pickled herring + saltines).* Al ser analizada por Gensim produce el siguiente resultado, el cual es un vector que contiene cada palabra como su peso dentro de la oración: [['herring', 0.46], ['pickled', 0.47], ['favorite', 0.25], ['my', 0.11], ['saltines', 0.57], ['snack', 0.4]]. En base a estos vectores generados, se hace un conteo de las palabras más frecuentes dentro de los corpus de casos positivos y negativos.
Se hizo un conteo de las palabras más centrales de los corpus negativos y positivos. Resultó llamativo que dentro de las palabras centrales que se encontraron en el corpus de casos positivos de depresión están youll, xxx, xt, wrenching, wrenching, wraps y wonderfully. Dentro de las palabras centrales del corpus del grupo de control están zack, yummy, yrs, yearly, yanked, xbone, yeh, yolk, yoda, ymmv, wrestlers, wounded y wouldve.

5.4 Análisis de patrones en los textos procesados

Se pudieron procesar 18229 ítems del corpus positivo y 26797 ítems del corpus del grupo de control. Dentro de análisis de los textos, también hubo 22875 ítems del corpus positivo y 29095 ítems del corpus negativo que no se procesaron en CODRA.
Por ejemplo un texto analizado:
My parents grew up here (New Holland). It's like my second home. What a beautiful part of the country! Can't leave without getting Achenbachs. (and smelling the "fresh" air)
Una vez procesado este texto dentro de CODRA, se genera una estructura RST de la siguiente forma:

```
( Root (span 1 6)
        ( Nucleus (span 1 5) (rel2par span)
                ( Nucleus (span 1 4) (rel2par span)
                        ( Nucleus (span 1 3) (rel2par span)
                                ( Nucleus (span 1 2) (rel2par span)
                                        ( Nucleus (leaf 1) (rel2par span)
                                        (text _!My parents grew up here_!) )\r\n
                                        ( Satellite (leaf 2) (rel2par Summary)
                                        (text _!( New Holland ) ._!) )
                                ( Satellite (leaf 3) (rel2par Elaboration)
                                (text _!Its like my second home ._!) ))
                        ( Satellite (leaf 4) (rel2par Elaboration)
                        (text _!What a beautiful part of the country !_!) )
                ( Satellite (leaf 5) (rel2par Elaboration)
                (text _!Cant leave without getting Achenbachs ._!) )
        ( Satellite (leaf 6) (rel2par Elaboration)
        (text _!( and smelling the \\\\\\" fresh \\\\\\" air )_!) ))
```

5.4.1 Características de textos no analizados

Las características en común de los textos que no pudieron ser analizados -que el parser CODRA no fue capaz de analizar- son las siguientes:

- Usos de emojis. (Ejemplo: *It was all me, then. ; D*)
- Uso de caracteres especiales, y puntos suspensivos. (Ejemplo: *I'm not sure...was yours covered in sacrificial blood..?*)
- Publicaciones de una sola oración sin una estructura clara. (Ejemplo: *This is the best shit*)
- Oraciones de forma exclamativa -que terminan con !- (Ejemplo: *Holy pepperoni pizza! That looks like some kind of swamp monster. Creepy!*)
- Textos con URL y direcciones web.

Asimismo también hubo textos cuyo procesamiento en CODRA excedía 5 minutos, así que fueron descartados del análisis.

5.4.2 Patrones en los textos

Se extrajeron de los árboles RST, una estructura general de los árboles en un nivel de generalización alto, de forma que se pueden extraer la estructura general de los árboles. Los resultados son semejantes dentro de ambos corpus. Los resultados son semejantes entre ambos corpus, únicamente el porcentaje de *(Nucleus)(Satellite)* es mayor en el corpus de control que en el corpus de casos positivos.

	Conteo Positivos:	**Porcentaje Positivos:**	**Conteo Control:**	**Porcentaje Control:**
(Nucleus)(Satellite)	**1798**	**9.90%**	**5069**	**18.90%**
(Satellite)(Nucleus)	805	4.40%	1440	5.40%
(Nucleus)(Satellite(Nucleus)(Satellite))	519	2.80%	1116	4.20%
(Nucleus)(Nucleus)	437	2.40%	710	2.60%
(Nucleus(Nucleus)(Satellite))(Satellite)	430	2.40%	633	2.40%
(Nucleus)(Satellite(Satellite)(Nucleus))	259	1.40%	424	1.60%
(Satellite)(Nucleus(Nucleus)(Satellite))	257	1.40%	405	1.50%
(Nucleus(Nucleus)(Satellite))(Nucleus)	245	1.30%	370	1.40%
(Nucleus)(Nucleus(Nucleus)(Satellite))	235	1.30%	314	1.20%

5.4.3 Tipos de patrones en los textos

En base a la estructura general, los patrones se categorizan entre diferentes tipos, siendo 1 el patrón más sencillo, así se incrementa hasta patrones más complejos. Por ejemplo siendo el patrón más sencillo de la forma: **(Nucleus)(Satellite)** o **(Satellite)(Nucleus)** es de **tipo 1**, el nivel de complejidad aumenta en las estructuras de los árboles aumenta siendo **(Nucleus)(Satellite(Nucleus)(Satellite))** de tipo 2, o uno de tipo **(Nucleus(Nucleus)(Satellite))(Satellite(Nucleus)(Satellite))** es de tipo 3. El corpus de clase positiva tiene como tipo mayoritario el 1, mientras el corpus de control tiene un porcentaje mayoritario en tipos que son mayores a 10.

	Conteo Positivos:	**Porcentaje Positivos:**	**Conteo Control:**	**Porcentaje Control:**
1	3041	16.70%	**7219**	**26.90%**
2	2964	16.30%	4983	18.60%
3	2330	12.80%	3309	12.30%
4	1606	8.80%	2285	8.50%
5	1280	7.00%	1742	6.50%
6	973	5.30%	1246	4.60%
7	756	4.10%	951	3.50%
8	637	3.50%	737	2.80%
9	512	2.80%	596	2.20%
> 10	**3664**	**20.10%**	3202	11.90%

5.4.5 Cálculo de la profundidad de los textos

Se calcula la profundidad de los árboles, en el cual es categorizado en base a la profundidad mas grande de las ramas de los árboles, por ejemplo un árbol de forma **(Nucleus)(Satellite)** o **(Satellite)(Nucleus)** es de **profundidad 1**, pero árboles de forma **(Nucleus)(Satellite(Nucleus)(Satellite))** o de **(Nucleus(Nucleus)(Satellite))(Satellite(Nucleus)(Satellite))** son de profundidad 2. Los resultados son semejantes entre ambos corpus, únicamente el porcentaje de profundidad 1 es mayor en el corpus de control que en el corpus de casos positivos.

	Conteo Positivos:	**Porcentaje Positivos:**	**Conteo Control:**	**Porcentaje Control:**
1	3041	16.70%	7219	26.90%
2	**4761**	**26.10%**	**7487**	**27.90%**
3	3983	21.80%	5433	20.30%
4	2629	14.40%	3142	11.70%
5	1692	9.30%	1688	6.30%
6	995	5.50%	904	3.40%
7	598	3.30%	470	1.80%
8	283	1.60%	255	1.00%
9	149	0.80%	125	0.50%
10	55	0.30%	44	0.20%

6. Conclusiones

Los resultados producto de estos análisis indican que los textos producidos por personas diagnosticadas con depresión tienden a tener una longitud mayor y por ende tienen una mayor complejidad que los textos emitidos por persona sin diagnóstico de depresión. Al respecto de la estructura general de los textos, los resultados arrojaron resultados similares para ambos corpus. Asimismo con la detección de ciertas palabras dentro de los ítems centrales del corpus de los casos positivos de depresión (como es el caso de *xxx* por ejemplo), se podrían inferir ciertas tendencias dentro de esa población.

Este trabajo da cabida a más oportunidades de investigación, y de esta forma arrojar más patrones con respectos a textos relacionados a este tipo de patologías, por ejemplo en el futuro se puede trabajar en validar estos patrones detectados en otros conjuntos de datos externos.

Asimismo, con un volumen mayor de datos, y. por ende con mayores capacidades de procesamiento, se podrían arrojar más patrones dentro de los corpus de clase positiva o sin diagnostico de depresión

Finalmente, se pueden evaluar estos mismos patrones para conjuntos de otras patologías mentales relacionadas (estados de ansiedad, trastornos alimentarios, esquizofrenia, ansiedad) para ver si son patrones comunes o específicos de depresión.

7. Referencias

[1] Losada David E., Crestani Fabio. (2019). *A Test Collection for Research on Depression and Language Use.* In *International Conference of the Cross-Language Evaluation Forum for European Languages*. 28-39.

[2] Choudhury Munmun D., De Sushovan, (2014). *Mental Health Discourse on reddit: Self-Disclosure, Social Support, and Anonymity,* In *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media*. 71-80.
[3] Lapata Mirella, Barzilay Regina. (2014). *Automatic Evaluation of Text Coherence: Models and Representations.* In *IJCAI'05: Proceedings of the 19th international joint conference on Artificial intelligence*. 1085–1090.
[4] McNamara Danielle S., Louwerse Max M., Graesser Arthur C. (2005). *Coh-metrix: Automated cohesion and coherence scores. To predict text readability and facilitate comprehension.* In *Behavior research methods, instruments, & computers: a journal of the Psychonomic Society*. 36. 193-202.
[5] Li Jiwei, Jurafsky Dan. (2017) *Neural Net Models of Open-domain Discourse Coherence.* In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. 198-209.
[6] Deerwester Scott, Dumais Susan T., Furnas George W., Landauer Thomas K. (1990) *Indexing by Latent Semantic Analysis*. In *Journal of the American Society for Information Science*. 41(6):391-407.
[7] Joty Shafiq, Carenini Giuseppe, Ng Raymond T. (2015). *CODRA: A Novel Discriminative Framework for Rhetorical Analysis.* In *Computational Linguistics*. 41(3). 385–435.
[8] Elsner Micha, Austerweil Joseph, and Charniak Eugene. (2007). *A Unified Local and Global Model for Discourse Coherence.* In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*. 436–443.
[9] Barzilay Regina, Lapata Mirella. (2008). *Modeling Local Coherence: An Entity-Based Approach.* In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05).* 141–148.
[10] Petersen Casper, Lioma Christina, Simonsen Jakob Grue, Larsen Birger, (2015). *Entropy and Graph Based Modelling of Document Coherence using Discourse Entities: An Application to IR.* In *Proceedings of the 2015 International Conference on The Theory of Information Retrieval*. 191–200.
[11] Al-Mosaiwi Mohammed, Johnstone Tom. (2018). *In an Absolute State: Elevated Use of Absolutist Words Is a Marker Specific to Anxiety, Depression, and Suicidal Ideation.* In Clinical Psychological Science, 7(3). 636–637.
[12] Ríssola Esteban, Losada David, Crestani Fabio. (2019). *Discovering Latent Depression Patterns in Online Social Media.* In *10th Italian Information Retrieval Workshop.* 13-16.
[13] Morales Michelle Renee, Scherer Stefan, Levitan Rivka. (2017). *A Cross-modal Review of Indicators for Depression Detection Systems.* In *Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology — From Linguistic Signal to Clinical Reality*. 1–12.
[14] Iter Dan, Yoon Jong H., Jurafsky Dan. (2018). *Automatic Detection of Incoherent Speech for Diagnosing Schizophrenia.* In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic.* 136–146.
[15] Duran Nicholas, Bellissens Cedrick, Taylor Roger, McNamara Danielle. (2007). *Quantifying Text Difficulty with Automated Indices of Cohesion and Semantics.* In *Proceedings of the 29th Annual Meeting of the Cognitive Science Society*. 233-238.
[16] Lal Alice, Tetreault Joel. (2018) *Discourse Coherence in the Wild: A Dataset, Evaluation and Methods.* In *Proceedings of the SIGDIAL 2018 Conference*. 214–223.
[17] Rehurek Radim, Sojka Petr. (2010). *Software Framework for Topic Modelling with Large Corpora.* In *Proceedings of LREC 2010 workshop New Challenges for NLP Frameworks*. 46-50.

[18] Coppersmith Glen, Dredze Mark, Harman Craig. (2018). *Quantifying Mental Health Signals in Twitter.* In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*. 51–60.
[19] Mann William, Thompson Sandra. (1988). *Rhetorical structure theory: Toward a functional theory of text organization.* In *Text - Interdisciplinary Journal for the Study of Discourse.* 8(3):243-281.
[20] Martín-Rodilla Patricia. (2020). *Adding Temporal Dimension to Ontology Learning Models for Depression Signs Detection from Social Media Texts*. In *Proceedings of the 15th International Conference on Evaluation of Novel Approaches to Software Engineering.* 323-330.
[21] Losada David E., Crestani Fabio, Parapar Javier. (2020). *eRisk 2020: Self-harm and Depression Challenges.* In *Advances in Information Retrieval. ECIR 2020. Lecture Notes in Computer Science.* 557-563
[22] Losada David E., Crestani Fabio, Parapar Javier. (2019). *Overview of eRisk 2019 Early Risk Prediction on the Internet.* In *Experimental IR Meets Multilinguality, Multimodality, and Interaction. CLEF 2019. Lecture Notes in Computer Science*. 340-357
[23] Losada David E., Crestani Fabio, Parapar Javier. (2018). *Overview of eRisk 2018: Early Risk Prediction on the Internet (extended lab overview).* In *Experimental IR Meets Multilinguality, Multimodality, and Interaction. CLEF 2018. Lecture Notes in Computer Science*. 343-361.
[24] Losada David E., Crestani Fabio, Parapar Javier. (2017) *eRISK 2017: CLEF Lab on Early Risk Prediction on the Internet: Experimental Foundations*. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction. CLEF 2017.* 346-360.
[25] Sparck Jones K. (1972). *A Statistical Interpretation Of Term Specificity And Its Application In Retrieval,* In *Journal Of Documentation*. 28(1): 11-21.
[26] Fundación Pública Galega Centro Tecnolóxico de Supercomputación de Galicia. *Centro de Supercomputación de Galicia.* Centro de Supercomputación de Galicia. https://www.cesga.es/cesga/
[27] Kurtzer Gregory M., Sochat Vanessa, Bauer Michael W. (2017). *Singularity: Scientific containers for mobility of compute*, In *Public Library of Science* 12(5):1-20.