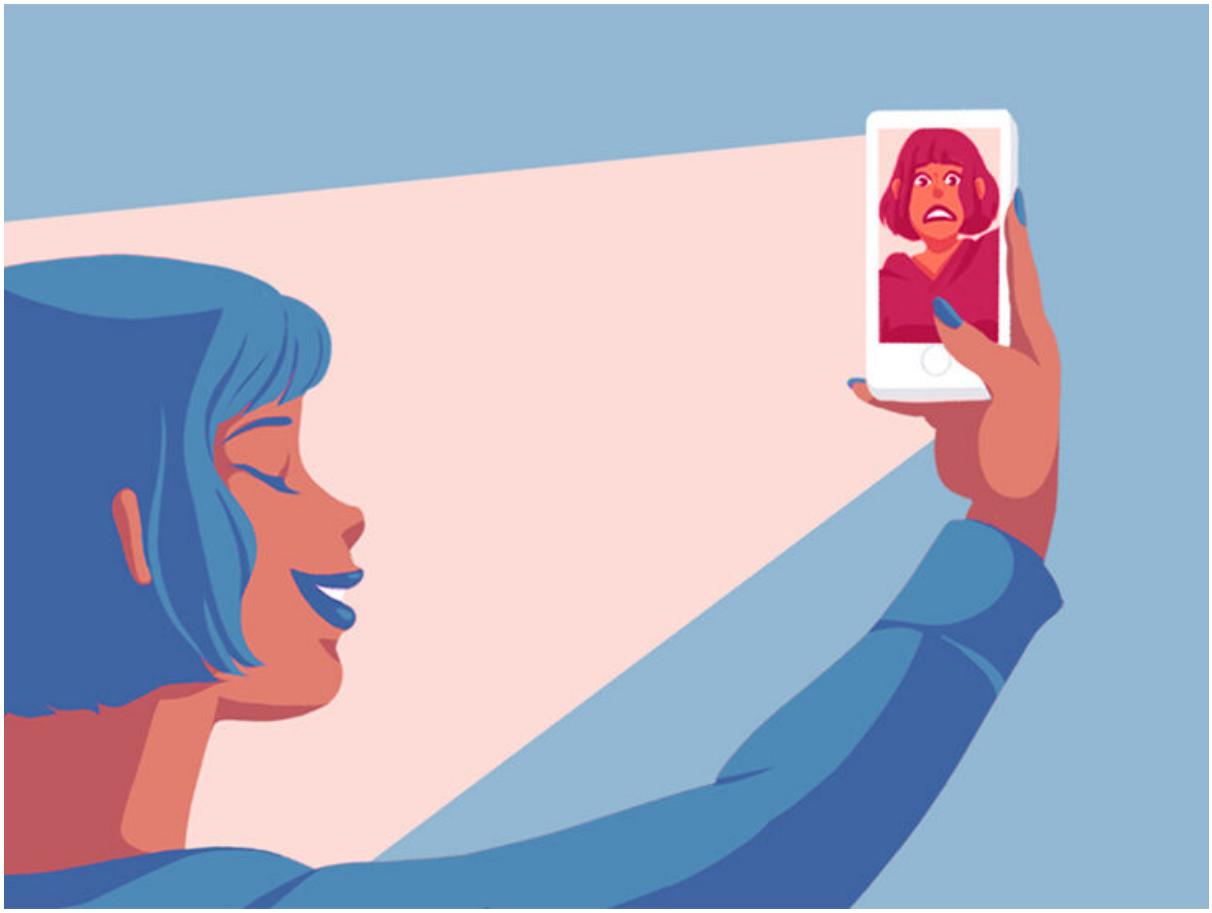
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA
ESCOLA TÉCNICA SUPERIOR DE ENXEÑARÍA



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

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

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.
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 hechos por personas con diagnostico no depresivo y los hechos por personas con diagnostico depresivo. Aplicando paradigmas de formalización del discurso, como puede ser RST (Rethorical Structure Theory) [19], se pueden evaluar los patrones en coherencia en común de esos textos.
Por medio de este paradigma, se representan los textos mediando por medio de estructuras jerárquicas llamadas Arboles 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 tomadas de redes sociales [1], se analizan tanto generalmente y así también se analiza su 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 esas arboles, se analiza sus estructura y los patrones en común entre otras mismas estructuras dentro de los corpus.
Nuestro estudio permitió el análisis de 45026 posts de las colección de datos eRisk que comprenden 1076582 entradas corresponden a 887 usuarios [1]. En base a estos análisis se pudieron detectar varios patrones, entre ellos la presencia de textos mas extensos entre el corpus relacionado a personas con diagnostico de depresión.

Tabla de Contenidos

Problema a resolver

En la ultima década, con el surgimiento de las redes sociales como Facebook, Twitter, Reddit e Instagram, 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 diagnostico.

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 a 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].

Métricas

Rethorical Structure Theory

La teoría de discurso utilizada es la Rethorical Structure Theory (RST), en la cual se representan los textos mediando por medio de estructuras jerárquicas llamadas Arboles 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 retorica 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 sentencia. Asimismo, tanto estructuras de núcleo como de satélite pueden ser **contenedores** de otras estructuras -llamadas **span**-, o si 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.***

Herramientas/Estado del Arte

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 como de prueba de textos con un diagnostico positivo como de textos con diagnostico negativo de depresión. Los datos de los usuarios de la clase positiva (es decir, con depresión) se recopilaron por medio de autoexpresiones de diagnósticos de depresión (p. Ej., 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]

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, segmentador 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.

Gensim

Gensim hace posible análisis en textos usando un Modelo de Espacio Vectorial, 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 lenguajes naturales se puedan agrupar en clústeres de 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 midiendo los términos por medio de TF-IDF, el cual calcula la relevancia de un término dentro del corpus. [25]

Proceso de análisis

Para el desarrollo del proyecto se usaron como base dos grandes 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 -uno con casos clínicamente diagnosticados positivamente de depresión, y otro con casos diagnosticados negativamente-.[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 a 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 arboles con la estructura semántica de cada entrada. La naturaleza de procesamiento por medio de un contenedor docker es muy alta, y se necesita un mínimo de 6 GB de RAM para su correcto procesamiento, 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 improbable en tales condiciones. Debido a esto, se hizo uso de la infraestructura del Centro de Supercomputación de Galicia (CESGA) [26], por medio del cluster Finisterrae se hace uso de la instalación del contenedor de CODRA, en este caso 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 corpus negativo y 41104 datos del corpus positivo, de los cuales se pudieron extraer arboles de coherencia de 26797 y 18229 datos respectivamente de cada corpus.
En base a los arboles 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 mas comunes dentro de las frases, y asimismo como el nivel de profundidad que pueden resultar de los arboles semánticos resultantes para cada corpus de sentencias.
Asimismo, debido a la alta tasa de sentencias procesadas que no resultaron en un árbol de coherencia, se analizaron de nuevo para poder mirar las causas y razones por las que esas entradas textuales no pudieron ser procesadas.

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, así mismo se analizan los textos que no pudieron ser analizados por el parser. Así también se hacen los análisis generados por Gensim. Las categorías que tienen mayor porcentaje que están en cada corpus, se encuentran en **negrita**.

Análisis descriptivo de los datos
Se hace un análisis descriptivo de los datos dentro del corpus positivo (90222 entradas) y el corpus negativo (986360 entradas). Dentro de los análisis que se hacen en el texto se encuentran; el conteo de caracteres, conteo de palabras, promedio de caracteres, numero de stopwords en el texto, porcentaje de stopwords en el texto.

Conteo de las palabras en el texto (de 1 a 100, sin contar los datos que están vacíos). Se cuentan las palabras que se encuentran dentro del corpus de casos positivos y negativos.

	Conteo Positivos:	**Porcentaje Positivos:**	**Conteo Negativos**	**Porcentaje Negativos:**
10-20	**32013**	**35.50%**	**342605**	**36.80%**
< 10	21367	23.70%	252662	27.20%

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%

Conteo de caracteres en el texto (de 1 a 1000, sin contar los datos que están vacíos). Se cuentan los caracteres que se encuentran dentro del corpus de casos positivos y negativos.

	Conteo Positivos:	**Porcentaje Positivos:**	**Conteo Negativos**	**Porcentaje Negativos:**
100-200	**18726**	**20.80%**	**192659**	**19.50%**
50-75	11929	13.20%	188881	19.20%
25-50	15255	16.90%	176554	17.90%
75-100	8749	9.70%	118222	12.00%
< 25	10439	11.60%	104081	10.60%
200-300	8401	9.30%	80794	8.20%
300-400	4787	5.30%	37400	3.80%
900-1000	3631	4.00%	26225	2.70%
400-500	3022	3.30%	22771	2.30%
500-600	2038	2.30%	14897	1.50%

Promedio de caracteres por palabra en el texto (mayores que 0). Se cuentan el promedio de caracteres que se encuentran dentro del corpus de casos positivos y negativos.

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

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 del corpus de casos positivos y negativos.

	Conteo Positivos:	**Porcentaje Positivos:**	**Conteo Negativos**	**Porcentaje Negativos:**
0	**152536**	**15.50%**	10712	11.90%

10-20	146104	14.80%	**15831**	**17.50%**
1	113691	11.50%	7194	8.00%
2	106011	10.70%	7115	7.90%
3	82498	8.40%	6299	7.00%
4	63711	6.50%	5507	6.10%
20-30	54176	5.50%	7064	7.80%
5	51333	5.20%	4713	5.20%
6	42438	4.30%	4021	4.50%

Porcentaje de *stopwords* en el texto (cuando el texto no es vacío) En base número de stopwords en el texto que se encuentran dentro del corpus de casos positivos y negativos, se calculan los porcentajes.

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

Estadísticas de análisis de coherencia

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

Numero de frases. Se cuentan las frases que se encuentran dentro del corpus de casos positivos y negativos.

	Conteo Positivos:	**Porcentaje Positivos:**	**Conteo Negativos**	**Porcentaje Negativos:**
> 10	**4130**	**22.70%**	3729	13.90%
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%

Conteo de frases de *elaboration*. Se cuentan las frases de *elaboration* que se encuentran dentro del corpus de casos positivos y negativos.

	Conteo Positivos:	**Porcentaje Positivos:**	**Conteo Negativos**	**Porcentaje Negativos:**
1	**4631**	**25.40%**	**8719**	**32.50%**

2	3124	17.10%	4594	17.10%
0	2833	15.50%	5006	18.70%
3	2078	11.40%	2639	9.80%
4	1351	7.40%	1748	6.50%
> 10	1083	5.90%	838	3.10%
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%

Conteo de frases de *enablement*. Se cuentan las frases de *enablement* que se encuentran dentro del corpus de casos positivos y negativos¶.

	Conteo Positivos:	**Porcentaje Positivos:**	**Conteo Negativos**	**Porcentaje Negativos:**
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%
8	1	0.00%		
7	1	0.00%		
6	1	0.00%	2	0.00%

Conteo de frases de *contrast*. Se cuentan las frases de *contrast* que se encuentran dentro del corpus de casos positivos y negativos¶.

	Conteo Positivos:	**Porcentaje Positivos:**	**Conteo Negativos**	**Porcentaje Negativos:**
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%
10	19	0.10%		
8	14	0.10%	6	0.00%
9			7	0.00%

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¶.

	Conteo Positivos:	**Porcentaje Positivos:**	**Conteo Negativos**	**Porcentaje Negativos:**
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%
> 10	515	2.80%	371	1.40%
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%

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¶.

	Conteo Positivos:	**Porcentaje Positivos:**	**Conteo Negativos**	**Porcentaje Negativos:**
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%
> 10	204	1.10%	104	0.40%
6	175	1.00%	181	0.70%
7	121	0.70%	141	0.50%
8	99	0.50%	59	0.20%

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¶.

	Conteo Positivos:	**Porcentaje Positivos:**	**Conteo Negativos**	**Porcentaje Negativos:**
0	**4328**	**23.70%**	**9503**	**35.50%**
1	4018	22.00%	6035	22.50%
2	2479	13.60%	3293	12.30%
> 10	1805	9.90%	1360	5.10%
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%

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¶.

	Conteo Positivos:	**Porcentaje Positivos:**	**Conteo Negativos**	**Porcentaje Negativos:**
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%
> 10	919	5.00%	547	2.00%
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%

Palabras más frecuentes y centrales

Usando la herramienta de análisis de textos Gensim, la cual usa Modelo de Espacio Vectorial para categorizar las palabras dentro de las sentencias, 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 sentencia: [['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 mas frecuentes dentro de los corpus de casos positivos y negativos.

Palabras que están dentro de los 20 ítems frecuentes del corpus de casos positivos y negativos	Palabras que solo están en los 20 ítems frecuentes del corpus de casos positivos	Palabras que solo están en los 20 ítems frecuentes del corpus de casos negativos
just	time	make
like	way	did
don	got	new
think	feel	
people		
time		
good		
way		
know		
really		
want		
going		
make		

Asimismo también se hace un conteo de las palabras mas centrales de los corpus negativos y positivos.

Palabras que solo están en los 20 ítems frecuentes del corpus de casos positivos		Palabras que solo están en los 20 ítems frecuentes del corpus de casos negativos	
youll	wrecks	zack	yolk
xxx	wraps	yummy	yoda
xt	wonderfully	yrs	ymmv
wrenching		yearly	wrestlers
xxx		yanked	wounded
		xbone	wouldve
		yeh	

Análisis de patrones en los textos procesados

Se pudieron procesar 18229 ítems del corpus positivo y 26797 ítems del corpus negativo. 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:

I go after what I want. Why should I wait for someone else to make the first move if I have my own expectations!

Una vez procesados los textos dentro de CODRA, se generan estructuras 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 )_!) ))
```

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.

Patrones en los textos

Se extrajeron de los arboles RST, una estructura general de los arboles en un nivel de generalización alto, de forma en que se pueden extraer la estructura general de los árboles:

	Conteo Positivos:	**Porcentaje Positivos:**	**Conteo Negativos**	**Porcentaje Negativos:**
(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%

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 arboles aumenta siendo **(Nucleus)(Satellite(Nucleus)(Satellite))** de nivel 2, o uno de tipo **(Nucleus(Nucleus)(Satellite))(Satellite(Nucleus)(Satellite))** es de nivel 3.

	Conteo Positivos:	**Porcentaje Positivos:**	**Conteo Negativos**	**Porcentaje Negativos:**
> 10	**3664**	**20.10%**	3202	11.90%
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%

Cálculo de la profundidad de los textos

Se calcula la profundidad de los arboles, en el cual es categorizado en base a la profundidad mas grande de las ramas de los arboles, por ejemplo un árbol de forma **(Nucleus)(Satellite)** o **(Satellite)(Nucleus)** es de **profundidad 1**, pero arboles de forma **(Nucleus)(Satellite(Nucleus)(Satellite))** o de **(Nucleus(Nucleus)(Satellite))(Satellite(Nucleus)(Satellite))** son de profundidad 2.

	Conteo Positivos:	**Porcentaje Positivos:**	**Conteo Negativos**	**Porcentaje Negativos:**
2	**4761**	**26.10%**	**7487**	**27.90%**
3	3983	21.80%	5433	20.30%
1	3041	16.70%	7219	26.90%
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%

Conclusiones

Los resultados producto de estos análisis emiten 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 diagnostico 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.

En base a este trabajo, los siguientes pasos que hay que tomar son los siguientes:

- Validar estos patrones detectados en otros conjuntos de datos externos.
- Trabajar con un volumen mayor de datos, y. por ende procesarlos con mayores capacidades de procesamiento.
- Validar los 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.

Referencias

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

[2] Munmun De Choudhury, Sushovan De, *Mental Health Discourse on reddit: Self-Disclosure, Social Support, and Anonymity,* In Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media, 2014.

[3] Lapata Mirella, Barzilay Regina, *Automatic Evaluation of Text Coherence: Models and Representations.* In IJCAI'05: Proceedings of the 19th international joint conference on Artificial intelligence, pages 1085–1090, 2014.

[4] McNamara Danielle S., Louwerse Max M., Graesser Arthur C. , *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(2):193-202, 2005.
[5] Li Jiwei and Jurafsky Dan, *Neural Net Models of Open-domain Discourse Coherence.* In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 198-209, 2017.
[6] Deerwester Scott, Dumais Susan T. , Furnas George W., and Landauer Thomas K. *Indexing by Latent Semantic Analysis*, In Journal of the American Society for Information Science. 41(6):391-407, 1990.
[7] Joty Shafiq, Carenini Giuseppe, Ng Raymond T. *CODRA: A Novel Discriminative Framework for Rhetorical Analysis.* In Computational Linguistics, Volume 41, Issue 3, pages 385–435, 2015.
[8] Elsner Micha, Austerweil Joseph, and Charniak Eugene, *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, pages 436–443, 2007.
[9] Barzilay Regina, Lapata Mirella, *Modeling Local Coherence: An Entity-Based Approach.* In Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05), pages 141–148, 2008.
[10] Petersen Casper, Lioma Christina, Simonsen Jakob Grue, Larsen Birger, *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, pages 191–200, 2015.
[11] Al-Mosaiwi, M., & Johnstone, T., Corrigendum: *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, 2018.
[12] Ríssola Esteban, Losada David, Crestani Fabio. *Discovering Latent Depression Patterns in Online Social Media.* In 10th Italian Information Retrieval Workshop, pages 13-16, 2019.
[13] Morales Michelle Renee, Scherer Stefan, Levitan Rivka, *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, pages 1–12, 2017.
[14] Dan Iter, Jong H. Yoon, and Dan Jurafsky, *Automatic Detection of Incoherent Speech for Diagnosing Schizophrenia.* In Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic, pages 136–146, 2018.
[15] Duran Nicholas, Bellissens Cedrick, Taylor Roger, McNamara Danielle. *Quantifying Text Difficulty with Automated Indices of Cohesion and Semantics.* In Proceedings of the 29th Annual Meeting of the Cognitive Science Society, pages 233-238, 2007.
[16] Lal Alice, Tetreault Joel, *Discourse Coherence in the Wild: A Dataset, Evaluation and Methods.* In Proceedings of the SIGDIAL 2018 Conference, pages 214–223, 2018.
[17] Rehurek Radim and Sojka Petr. *Software Framework for Topic Modelling with Large Corpora.* In Proceedings of LREC 2010 workshop New Challenges for NLP Frameworks, pages 46-50, 2010.
[18] Glen Coppersmith, Mark Dredze, and Craig Harman. *Quantifying Mental Health Signals in Twitter.* In Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality. Baltimore, USA, 2018
[19] Mann, William and Sandra Thompson. *Rhetorical structure theory: Toward a functional theory of text organization.* Text, 8(3):243–281, 1988

[20] Martín-Rodilla, Patricia. *Adding Temporal Dimension to Ontology Learning Models for Depression Signs Detection from Social Media Texts*. In ENASE, pp. 323-330. 2020.
[21] Losada D.E., Crestani F., Parapar J. *eRisk 2020: Self-harm and Depression Challenges.* In Advances in Information Retrieval. ECIR 2020. Lecture Notes in Computer Science, 2020
[22] Losada D.E., Crestani F., Parapar J. (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, 2019
[23] Losada D.E., Crestani F., Parapar J. *Overview of eRisk 2018: Early Risk Prediction on the Internet (extended lab overview).* In Conference and Labs of the Evaluation Forum, Avignon (Francia), CEUR Workshop Proceedings, pp. 1-20 , 2018.
[24] Losada D.E., Crestani F., Parapar J. *eRISK 2017: CLEF Lab on Early Risk Prediction on the Internet: Experimental Foundations*. In Experimental IR Meets Multilinguality, Multimodality, and Interaction. CLEF 2017, 2017
[25] Sparck Jones, K., *A Statistical Interpretation Of Term Specificity And Its Application In Retrieval,* In Journal Of Documentation, Vol. 28 No. 1, Pp. 11-21, 1972
[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 GM, Sochat V, Bauer MW. *Singularity: Scientific containers for mobility of compute*, PLoS ONE 12(5), 2017