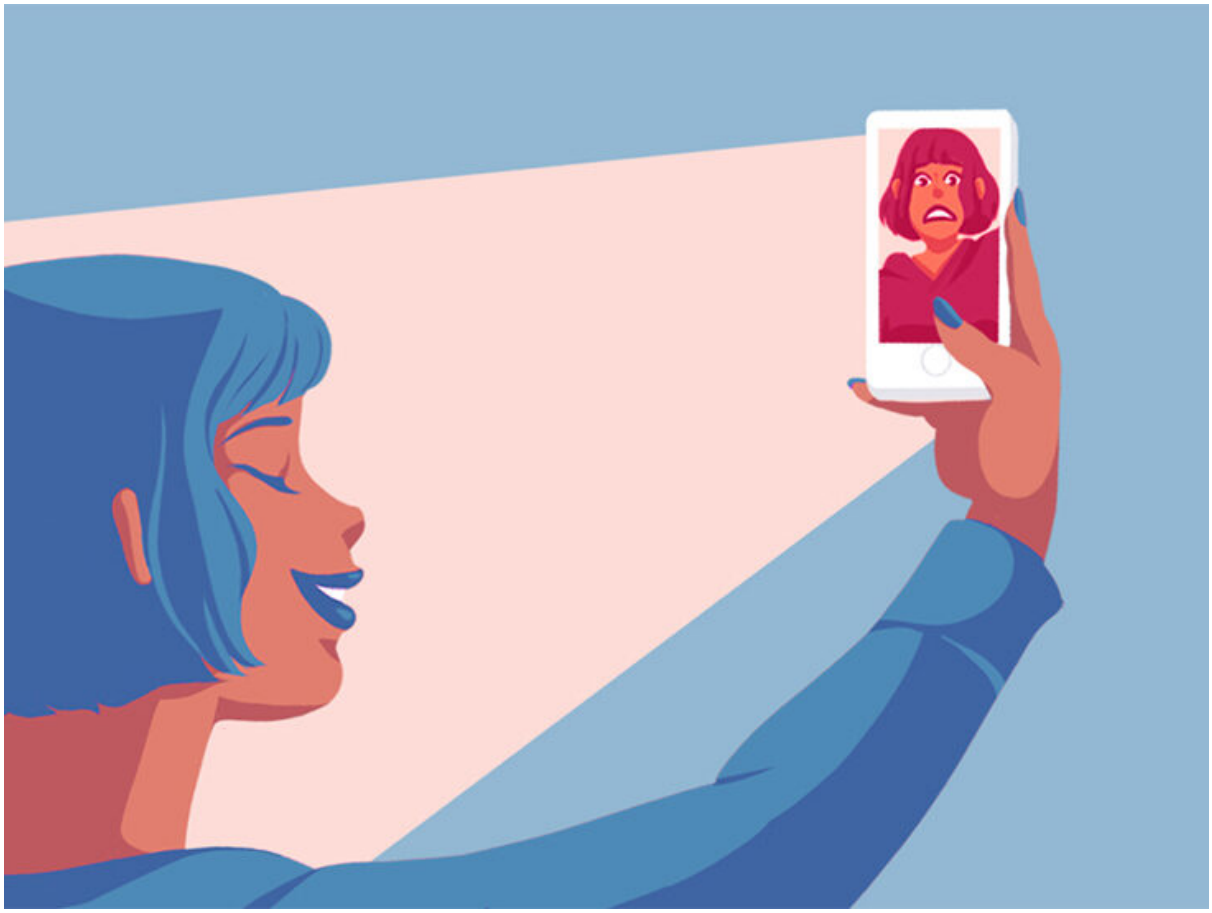
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA
ESCOLA TÉCNICA SUPERIOR DE ENXEÑARÍA



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 Martín-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, en ocasiones se encuentran 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*) [2], se pueden evaluar los patrones en coherencia en común de esos textos.

Por medio de este paradigma, se representan los textos a través 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. Las EDUs adyacentes son conectadas por relaciones de coherencia (Elaboración, Contraste…) formando unidades de discurso más grandes (representadas por nodos internos). Estos patrones o pistas lingüísticas pueden servir de indicadores de patologías mentales 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 [3], el cual 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 los árboles resultantes de diferentes textos del corpus. Nuestro estudio permitió el análisis de 45026 posts de la colección de datos eRisk que comprende 1076582 entradas correspondientes a 887 usuarios en el idioma inglés. Cada entrada se encuentra debidamente etiquetada con información precisa en relación a si es producida por una persona con diagnóstico de depresión o no [1]. En base a estos análisis se pudieron detectar varios patrones, entre ellos la presencia de textos extensos y complejos en el corpus relacionado con personas diagnosticadas de depresión.

Abstract

Recent research shows text retrieved from social media being a complementary source for study and early risks detection of suffering mental illness [1].

These differences in natural languages used in social media exposes satisfactory results at lexical and/or grammatical level, with fewer evidences at the discursive level.

Focusing on the detection of depression at a discursive level, differences can be found between texts written by people without depressive diagnosis and text written by people with a depressive diagnosis. Applying paradigms of discourse formalization, like the Rhetorical Structure Theory (RST) [2], it is possible to evaluate the common coherence patterns of these texts.

With this paradigm, texts are represented by hierarchical structures called discourse trees. Discourse tree leaves correspond to Elementary Discourse Units (EDUs). EDUs are units which function as construction blocks. Adjacent EDUs are connected by coherence relations (Elaboration, Contrast…) forming larger discourse units (represented by internal nodes). These linguistic patterns, or clues, can serve as an indicator of a mental pathology that may facilitate a diagnosis of these illnesses.

Following this research hypothesis, this work performs a Big Data analysis of textual data collection from social media. The goal is to automatically analyze these discursive structures, and to infer possible relevant discursive patterns in mental illness.

Taking different text collections obtained from social media [1], we performed a general statistics and coherence study of the trees RST generated by the CODRA framework [3], using a discursive analysis of texts found in the corpus. This analysis was based on these trees, their structure, and the common patterns between the trees resulting from different texts of the corpus. We analysed 45026 posts from the eRisk data collection that comprises 1076582 entries corresponding to 887 users written in the English language. Each entry is duly labelled with precise information regarding whether it is produced by a person with a diagnosis of depression or not [1]. Based on these analyses, several patterns could be detected, including the presence of longer and more complex texts in the corpus related to people diagnosed with depression.

Resumo

Os textos recompilados das redes sociais demostraron en investigacións recentes seren unha fonte complementaria para o estudo e detección anticipada dos riscos de padecer enfermidade mental [1].

Estas diferenzas na linguaxe natural usada en redes sociais amosaron resultados satisfactorios a nivel léxico e/ou gramatical, estudándose moito menos o nivel discursivo.

Centrándonos na detección de depresión e no nivel discursivo, en ocasións atópanse diferenzas entre textos escritos por persoas cun diagnóstico non depresivo e os escritos por persoas con diagnóstico depresivo. Aplicando paradigmas de formalización do discurso, como pode ser o RST (*Rhetorical Structure Theory*) [2], pódense avaliar os patróns de coherencia compartidos por tales textos.

Por medio deste paradigma, represéntanse os textos a través de estruturas xerárquicas chamadas Árbores de Discurso. As follas dunha árbore de discurso corresponden a unidades de texto atómicas (EDUs). As EDUs son unidades que funcionan como bloques de construción. As EDUs adxacentes están conectadas por relacións de coherencia (Elaboración, Contraste…) formando unidades de discurso máis grandes (representadas por nodos internos). Estes patróns ou pistas lingüísticas poden servir de indicadores de patoloxías mentais, de xeito que poderían facilitar ou advertir o diagnóstico de tales enfermidades.

Seguindo esta hipótese de investigación, este TFM realiza unha análise *Big Data* de coleccións de datos textuais dende redes sociais co obxectivo de analizar de forma automática as estruturas discursivas xurdidas dos mesmos e inferir posibles patróns discursivos de relevancia en enfermidade mental.

Tomando diferentes coleccións de texto obtidas de redes sociais [1], realízanse tanto unha análise xeral como un estudo de coherencia por medio do *framework* CODRA [3], o cal fai un análise discursiva por medio da xeración de árbores RST de cada un dos textos que se encontran no corpus. En base a esas árbores, analizase a súa estrutura e os patróns en común entre as árbores resultantes de diferentes textos do corpus. O noso estudo permitiu a análise de 45026 publicacións da colección de datos eRisk que comprende 1076582 entradas correspondentes a 887 usuarios na lingua inglesa. Cada entrada encóntrase debidamente etiquetada con información precisa en relación a se é producida por unha persoa con diagnóstico de depresión ou non [1]. En base a estas análises puidéronse detectar varios patróns, entre eles a presenza de textos extensos e complexos no corpus relacionado con persoas diagnosticadas de depresión.

Tabla de Contenidos

1. Problema a resolver

Estudios han demostrado que el uso de redes sociales está significadamente asociado con la depresión [4]. La falta de un tratamiento puede llevar a ansiedad, episodios psicóticos, 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 lingüísticas surgidas de los textos a gran escala y de esta forma, inferir posibles patrones discursivos de relevancia en enfermedad mental.

A nivel discursivo, 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 [5]. Por otro lado, existen estructuras discursivas, conectores y temáticas que son más comunes dentro de los encontrados en los foros de redes sociales relacionados con salud mental [6, 7, 8, 9]. 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 [2, 3, 10], 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 [11].

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 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**-. Estas estructuras en las que se divide el lenguaje natural están relacionadas entre sí mediante relaciones discursivas con significado semántico [2]. Las relaciones de coherencia 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 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.****"*

***Background*:** Una relación de fondo es una relación interproposicional en la que se proporcionan una o más proposiciones como información necesaria para la comprensión adecuada de algunas otras proposiciones.

*Ejemplo: "****Someone left a coffee cup in my office****; would the owner please come and get it?"*

***Evaluation*:** Una relación de evaluación es una relación interproposicional en la que una proposición expresa el juicio de valor del hablante con respecto a la factibilidad o conveniencia de otra proposición [12].

Ejemplo: "Features like our uniquely sealed jacket and protective hub ring make our discs last longer. And a soft inner liner cleans the ultra-smooth disc surface while in use. ***It all adds up to better performance and reliability.****"*

***Attribution*:** Una relación de atribución es una relación de elaboración en la que una proposición describe un atributo de un referente de otra proposición [13].

Ejemplo: "A girl shouted me on the street, ***she was blonde and was wearing a blue dress.****"*

***Summary*:** Una relación de resumen es una relación de contracción en la que una proposición repite, en forma abreviada, la información de un grupo de proposiciones expresadas previamente.

*Ejemplo: "…****The body of a long entertaining text on telegrams had 43 units on the subject, followed by this summary of 5 units: It seems a while since there's been a neatly worded dispatch from the field. (This was followed by 4 units of elaboration.)****"*

***Condition*:** Una relación de condición es una relación de elaboración en la que una proposición depende de la realización de otra proposición. (Presenta una situación hipotética, futura o no realizada).

Ejemplo: "Employees are urged to complete new beneficiary designation forms for retirement or life insurance benefits ***whenever there is a change in marital or family status.****"*

***Joint*:** Una relación de unión es una relación multinuclear, en la cual las conjunciones comparten el mismo tema y están conectadas por una conjunción [14].

*Ejemplo: "****Features like our uniquely sealed jacket and protective hub ring make our discs last longer*** *and a* ***soft inner liner cleans the ultra-smooth disc surface while in use.****"*

***Topic-Comment*:** Una relación de tópico y comentario es una relación de elaboración en la que una proposición llamada tópico se basa en lo que se trata el texto y el comentario son un grupo de proposiciones que añade sobre lo que se dice del tópico [15].

Ejemplo: "Sometimes you try hard to like something and just don't and ***I think that is just fine.****"*

***Textual-Organization*:** Una relación de organización textual es una pseudo-relación multinuclear, donde una proposición añade una expresión o una afirmación a otra proposición.

Ejemplo: "Is there a reason why you HAVE to go ? ***He should understand your anxiety.****"*

***Manner-Means*:** Una relación de forma y medios es una relación de elaboración donde una proposición presenta un método o instrumento que tiende a hacer más probable la realización de otra proposición.

Ejemplo: "...hindering the economy ***by not introducing new gold ;"***

Comparison**:** Una relación de comparación es una relación de elaboración en la que una proposición compara otra proposición.

Ejemplo: "Stock market less ***than when he came into office ,"***

Cause**:** Una relación de cause es una relación de elaboración en la que una proposición brinda una causa para la realización de otra proposición.

Ejemplo: "...and my sense is that the wife doesn't really care about their finances ***because she doesn't…"***

Temporal**:** Una relación de tiempo es una relación de elaboración en la que una proposición muestra el estado de tiempo en que se lleva a cabo otra proposición.

Ejemplo: "Then I went to a house show got my knee kicked in ***while trying to dance ."***

Explanation**:** Una relación de explicación es una relación de elaboración en la que una proposición muestra el causa de la realización de otra proposición.

Ejemplo: "I don't want to stop ***because I don't like who I am when I'm sober . "***

Same-Unit**:** Una relación de "misma unidad" es una pseudo-relación multinuclear, en la cual todas las proposiciones son núcleo, y tienen la misma importancia dentro del texto.

*Ejemplo: "****Once you know how to work two strands at once , graph paper will help you draft your own patterns .****"*

Topic-Change**:** Una relación de "cambio de tema" es una relación de elaboración en la que una proposición muestra un cambio de tema en relación con otra proposición [16].

Ejemplo: "My eyesight isn't even that bad, ***but my optometrist is one of my favorite people."***

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 en el idioma inglés. 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 [17, 18, 19, 20]. 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 [21]. 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 [3], 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 pueden 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) [22] y también se han analizado los términos por medio de TF-IDF, que pondera la relevancia de un término dentro del corpus [23].

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 como porcentajes relacionados con estas estadísticas.

Una segunda parte consistió en la generación de análisis de coherencia, por medio de CODRA [3], el cual es un framework en el cual se generaron árboles con la estructura semántica de cada entrada. 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 el tratamiento de un corpus de mas de 1000000 de datos para el tiempo dedicado al TFM era excesivo. Debido a esto, se hizo uso de la infraestructura del Centro de Supercomputación de Galicia (CESGA) [24]. En concreto, por medio del clúster 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 [25].

Usando este entorno se tuvo la capacidad de procesar 55892 entradas del grupo de control y 41104 entradas del clase positiva, de los cuales se pudieron extraer árboles de coherencia de 26797 y 18229 entradas 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 proposiciones de núcleo, de contraste, de satélite, de habilitación, entre otros. En una segunda parte de este análisis, también se extrajeron cuáles 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 y el porcentaje de stopwords en el texto.

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

Se calcula el número de palabras en cada oración. Al no apreciarse diferencias significativas entre los grupos, no consideramos que en este caso esta variable actúe como posible indicador o marcador diferencial en detección de depresión.

	Conteo Positivos:	**Porcentaje Positivos:**	**Conteo Control:**	**Porcentaje Control:**
< 10	**30313**	**36.6%**	**390807**	**41.6%**
10-20	20258	24.4%	264588	28.2%
20-30	10951	13.2%	105811	11.3%
30-40	6729	8.1%	63078	6.7%
40-50	4442	5.4%	39924	4.3%
50-60	3235	3.9%	25238	2.7%
60-70	2380	2.9%	17875	1.9%
70-80	1944	2.3%	13330	1.4%
80-90	1469	1.8%	10199	1.1%
90-100	1156	1.4%	8150	0.9%

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

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

Se calcula el número de caracteres en el texto. Al no apreciarse diferencias significativas entre los grupos, no consideramos que en este caso esta variable actúe como posible indicador o marcador diferencial en detección de depresión.

	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%

Tabla 2 Número de caracteres en el texto (rango de 1 a 1000, sin contar los datos que están vacíos).

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

Se calcula el promedio de caracteres que se encuentran en ambos corpus. Como se puede apreciar en la tabla 3, en este caso sí que detectamos cierta diferencia significativa entre los textos producidos por los usuarios positivos y negativos. En concreto, 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 tienen un promedio de 5 a 10 caracteres. Esto podría indicar que la longitud de texto es mayor en el grupo de control, y por tanto que los individuos diagnosticados con depresión presentan menor longitud promedio en sus participaciones en la red social. Posteriormente seguiremos observando este fenómeno en otros patrones asociados.

	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%

Tabla 3 Promedio de caracteres por palabra en el texto (mayores que 0).

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

Se calcula el número de stopwords en el texto que se encuentran dentro de ambos corpus. 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). En el grupo de control se encuentra una cantidad menor de stopwords mayoritariamente los textos tienen entre 0, 1 y 2 stopwords. Esto da lugar que en el corpus de control hay un lenguaje más sencillo y atómico, por lo tanto menos elaborado. Se observa en la tabla 4 que los positivos tienen una cantidad mayor de stopwords, lo cual indica un lenguaje más enrevesado y complejo. En otros patrones asociados se va a seguir observando este fenómeno.

	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%

Tabla 4 Número de stopwords en el texto (cuando el texto no es vacío).

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. Si bien existen diferencias en la cantidad de stopwords, en este caso al no apreciarse diferencias significativas entre los grupos, no consideramos que esta variable actúe como posible indicador o marcador diferencial en detección de depresión.

	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%

Tabla 5 Porcentaje de stopwords en el texto (cuando el texto no es vacío)

5.1.6 Prueba de legibilidad de Flesch

En la prueba de legibilidad de Flesch, las puntuaciones más altas indican material que es más fácil de leer; los números más bajos marcan los pasajes que son más difíciles de leer [26]. La formula 1 muestra cómo se calcula la puntuación de la prueba de legibilidad de Flesch.

$$206.835 - 1.015\left(\frac{total\ de\ palabras}{total\ de\ oraciones}\right) - 84.6\left(\frac{total\ de\ sílabas}{total\ del\ palabras}\right)$$

Formula 1 Puntuación de la prueba de legibilidad de Flesch

Esta prueba brinda una puntuación la cual tiene un nivel de equivalencia de acuerdo con el nivel de comprensión del texto. Se puede ver en la tabla 6 que los textos en el corpus positivos tienen un nivel de lectura más fácil y comprensible para lectores novatos. Mientras los textos producidos por el grupo de control tienen un nivel de legibilidad accesible para lectores experimentados.

	Equivalencia	**Conteo Positivos:**	**Porcentaje Positivos:**	**Conteo Control:**	**Porcentaje Control:**
0-10	Lectura extremadamente difícil	9742	10.8%	151498	15.4%
10-30	Lectura muy difícil	2956	3.3%	72551	7.4%
30-50	Lectura difícil	13483	15%	205870	**20.9%**
50-60	Lectura relativamente difícil	12043	13.4%	137859	14%
60-70	Lenguaje llano	16591	18.4%	156995	16.0%
70-80	Lectura relativamente fácil	17895	**19.9%**	135223	13.7%
80-90	Lectura fácil	13054	14.5%	94942	9.7%
90-100	Lectura muy fácil	4216	4.7%	28595	2.9%

Tabla 6 Prueba de legibilidad de Flesch

5.2 Estadísticas de análisis de coherencia

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 discursivamente por CODRA:

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 )_!) ))
```

Los tipos de relaciones entre proposiciones que mejor resultado y mayor análisis nos han permitido son: *Enablement, Contrast, Joint, Attribution y Same-Unit.*

5.2.1 Número de proposiciones

Se calcula el número de proposiciones que se encuentran dentro del corpus de casos positivos y negativos. Como se puede observar en la tabla 7, se puede observar que mayoritariamente los positivos tienen 10 o más proposiciones dentro de los textos, mientras que los textos del grupo de control tienen mayoritariamente dos proposiciones. En este y otros patrones analizados parecen indicar una tendencia entre el corpus de casos positivos de textos más largos y complejos.

	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%

Tabla 7 Número de proposiciones

5.2.2 Tipo de relaciones entre proposiciones

Se calcula el total de tipo de relaciones entre proposiciones en los corpus completos de positivos y de control. Por ejemplo, como se puede apreciar en la Tabla 8 los textos de positivos tienen una mayor cantidad de relaciones de tipo *Topic-Change, Contrast* los cuales indican la presencia de textos más complejos en el corpus de positivos.

	Conteo Positivos:	**Porcentaje Positivos:**	**Conteo Control:**	**Porcentaje Control:**
Elaboration	**60423**	**30.77%**	**67507**	**34.27%**
Joint	31377	15.98%	28585	14.51%
Attribution	29810	15.18%	29616	15.03%
Same-Unit	18554	9.45%	19608	9.95%
Contrast	10232	5.21%	9348	4.74%
Topic-Change	8066	4.10%	7038	3.57%
Background	5917	3.01%	5885	2.98%
Enablement	5636	2.87%	6178	3.13%
Explanation	4720	2.40%	3356	1.70%
Temporal	4615	2.35%	4419	2.24%
Topic-Comment	4033	2.05%	3519	1.78%
Condition	3987	2.03%	3713	1.88%
Cause	3156	1.60%	2970	1.50%
Summary	2420	1.23%	2278	1.15%
Evaluation	1216	0.61%	913	0.46%
Comparison	1210	0.61%	1139	0.57%
Manner-Means	508	0.25%	486	0.24%
Textual-Organization	448	0.2%	412	0.20%

Tabla 8 Tipo de relaciones entre proposiciones

5.2.3 Relaciones de Elaboration.

Se calcula el número de relaciones de *elaboration* que se encuentran dentro del corpus de casos positivos y negativos. Aunque no se observan diferencias significativas para el número de proposiciones, se observa un cambio de tendencia en los porcentajes entre los corpus. En el corpus de positivos hay un porcentaje mayor de textos con 10 o más proposiciones de *elaboration* que en el corpus de control donde en su mayoría tienen entre 0, 1 y 2 proposiciones de este tipo. Este tipo de tendencia sigue dentro de los diferentes tipos de proposiciones en el corpus y apoya la hipótesis de patrones anteriores acerca de mayor farragosidad y complejidad estructural en el discurso del corpus de positivos.

	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%

Tabla 9 Relaciones de Elaboration.

5.2.4 Relaciones de Enablement.

Se calcula el número de relaciones de *enablement* que se encuentran dentro del corpus de casos positivos y negativos. Aunque no se observan diferencias significativas para el número de proposiciones, se observa una pequeña diferencia entre los porcentajes entre los corpus. Se puede observar en la Tabla 10 que entre los textos positivos hay una mayor presencia de proposiciones de *enablement* que en los textos negativos.

	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%		

Tabla 10 Conteo de relaciones de Enablement.

5.2.5 Relaciones de Contrast.

Se calcula el número de relaciones de *contrast* que se encuentran dentro del corpus de casos positivos y negativos. No hay grandes diferencias entre los corpus, pero se puede observar que dentro de los positivos hay una mayor presencia de proposiciones de *contrast* que en el corpus de control. Esta diferencia implica que se podría dar una mayor complejidad en los textos producidos por positivos.

	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%		

Tabla 11 Relaciones de Contrast.

5.2.6 Relaciones de Joint

Se calcula el número de relaciones de *joint* que se encuentran dentro del corpus de casos positivos y negativos. Se puede apreciar que dentro del corpus de positivos hay una mayor cantidad de proposiciones de este tipo de relación multinuclear, lo cual indica que hay mayor cantidad de textos complejos que cuentan una cantidad mayor de proposiciones nucleares.

	Conteo Positivos:	**Porcentaje Positivos:**	**Conteo Control:**	**Porcentaje Control:**
0	**10683**	**58.6%**	**18540**	**69.2%**
1	1	0.0%	1	0.0%
2	4271	23.4%	5408	20.2%
4	1518	8.3%	1542	5.8%
6	686	3.8%	616	2.3%
8	399	2.2%	275	1.0%
> 10	443	2.4%	263	1.0%

Tabla 12 Relaciones de Joint.

5.2.7 Relaciones de Attribution

Se calcula el número de proposiciones de *attribution* que se encuentran dentro del corpus de casos positivos y negativos. En el corpus de positivos hay un porcentaje mayor de textos con algún texto con una relación de *attribution,* lo cual como se está viendo en varios análisis puede afianzar la hipótesis de que existe complejidad mayor en los textos producidos por positivos.

	Conteo Positivos:	**Porcentaje Positivos:**	**Conteo Control:**	**Porcentaje Control:**
0	**9901**	**54.3%**	**16754**	**62.5%**
1	5274	28.9%	7287	27.2%
2	1571	8.6%	1728	6.4%
3	691	3.8%	571	2.1%
4	311	1.7%	233	0.9%
5	189	1.0%	105	0.4%
6	111	0.6%	48	0.2%
7	65	0.4%	36	0.1%
8	52	0.3%	11	0.1%
> 10	27	0.1	16	0.0%

Tabla 13 Relaciones de Attribution

5.2.8 Relaciones de Same-Unit

Se calcula el número de proposiciones con relaciones *Same-Unit* que se encuentran dentro del corpus de casos positivos y negativos. Se puede apreciar un menor porcentaje de proposiciones de este tipo en el grupo de control, lo cual, al ser una relación multinuclear también implica que existe una mayor complejidad, al contener mayor cantidad de proposiciones con muchas ideas principales.

	Conteo Positivos:	**Porcentaje Positivos:**	**Conteo Control:**	**Porcentaje Control:**
0	12575	69.0%	19939	74.4%
2	3697	20.3%	5050	18.8%
4	1105	6.1%	1156	4.3%
6	450	2.5%	385	1.4%
8	198	1.1%	156	0.6%
> 10	113	0.6%	59	0.2%

Tabla 14 Relaciones de Same-Unit

5.2.9 Proposiciones de núcleo que son leaf.

Se calcula el número de proposiciones de núcleo que son *leaf* dentro del RST (que solo abarcan una proposición) que se encuentran dentro de los corpus de casos positivos y negativos. Se puede observar que en el corpus de casos positivos hay una mayor presencia de proposiciones de este tipo que en el corpus de control. Las proposiciones de núcleo dentro de un leaf suelen ser oraciones individuales y/o atómicas. Por este motivo se encuentra un discurso más atómico, menos farragoso y sencillo en el grupo de control, al contener menos de este tipo de estructuras.

	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%

Tabla 15 Proposiciones de núcleo que son leaf.

5.2.10 Proposiciones de satélite que son leaf.

Se calcula el número de proposiciones de satélite que son *leaf* del RST (que solo abarcan una proposición) que se encuentran dentro de los corpus de casos positivos y negativos. Hay un porcentaje mayor de textos dentro del corpus positivo que tienen proposiciones de satélite que son *leaf*, lo cual indica la existencia de estructuras más complejas y menos simples en tal corpus.

	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%

Tabla 16 Proposiciones de satélite que son leaf.

5.2.11 Proposiciones de núcleo que son span.

Se calcula el número de proposiciones de núcleo que son *span* del RST (que abarcan varias proposiciones) que se encuentran dentro de los corpus de casos positivos y negativos, los resultados son semejantes dentro de ambos corpus. Se puede observar que en el corpus de textos producidos por personas con diagnóstico positivo de depresión tienen un porcentaje mayor de oraciones con más de 10 proposiciones de núcleo que son *span* y que en el grupo de control tiene un porcentaje mayor de oraciones que tienen 0 proposiciones de núcleo que son *span* que la clase positiva. Además se observa que hay un porcentaje mayor de textos en el corpus positivos que cuentan con proposiciones de este tipo. Esto implica a nivel discursivo que los textos producidos por positivos podrían ser más complejos.

	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%

Tabla 17 Proposiciones de núcleo que son span.

5.2.12 Proposiciones de satélite que son span.

Se calcula el número de proposiciones de satélite que son *span* del RST (que abarcan varias proposiciones) 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 proposiciones de satélite que son *span* que en el grupo de control y que en el grupo de control tiene un porcentaje mayor de oraciones que tienen 0 proposiciones de satélite que son *span* que la clase positiva. Como se puede ver en los otros análisis, esto implica que los textos producidos en el corpus positivo tienen un mayor de complejidad dentro de las proposiciones que no son principales en los textos.

	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%

Tabla 18 Proposiciones de satélite que son span

5.2.13 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 entre ambos corpus, únicamente el porcentaje de *(Nucleus)(Satellite)* es mayor en el corpus de control que en el corpus de casos positivos, lo cual confirma lo analizado en puntos anteriores que implica que los textos son más simples en el corpus de control.

	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%

Tabla 19 Patrones en los textos

5.2.14 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. Se puede ver que las estructuras más habituales y coherentes (los tipos con un valor más bajo) los valores mayores se dan en el grupo de control, mientras que en estructuras menos habituales, más farragosas, enrevesadas y con ausencia de coordinadas se dan en el grupo de positivos. Esto viene a afianzar los resultados previos en conteo, longitudes y multinuclearidad, sobre una cierta tendencia del discurso en positivos más enrevesado, largo y desorganizado en el grupo de positivos que en el grupo de control.

	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%

Tabla 20 Tipos de patrones en los textos

5.2.15 Cálculo de la profundidad de los textos

Se calcula la profundidad de los árboles, que es analizada en base a la profundidad más 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. La diferencia que se puede encontrar es que en los niveles de mayor profundidad (a partir de 5) cambia la tendencia: los positivos suelen generar árboles más profundos que indican de nuevo una tendencia a textos con mayor farragosidad y longitud en el corpus de positivos que venimos observando en otros indicadores previos.

	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%

Tabla 21 Cálculo de la profundidad de los textos

5.3 Palabras centrales

Las palabras centrales se calculan en base a la importancia que puede tener una palabra en un corpus determinado. Cada palabra tiene un peso, el cual es simplemente proporcional a la frecuencia del término [27]. Usando la herramienta de análisis de textos Gensim, que usa el Modelo de Espacio Vectorial [23] para categorizar las palabras dentro de las oraciones, y verificar y su frecuencia y peso. 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 calcularon las palabras más centrales de los corpus negativos y positivos. En las tablas 22 y 23 se pueden mirar las palabras centrales en los corpus de positivos y de control. Aunque esta parte del estudio no arrojó patrones léxicos concretos, se documentan aquí los resultados para futuros estudios léxicos que se centren en la construcción de tesauros diferenciadores para el estudio de depresión.

Palabra	Clase
youll	Contracción
xxx	Sustantivo
wrenching	Sustantivo
wraps	Sustantivo
wonderfully	Adverbio
xt	Contracción

Tabla 22 Palabras centrales corpus positivo

Palabra	Clase
zack	Sustantivo
yummy	Sustantivo
yrs	Sustantivo
yearly	Adverbio
wounded	Adjetivo

Tabla 23 Palabras centrales corpus control

5.4 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. En general, creemos que esto es debido a marcas propias del lenguaje coloquial y metalenguaje usado en redes sociales.

Como futuro, se plantea un trabajo mayor de preprocesado y limpieza del corpus de eRisk utilizado que nos permitirá analizar muchos de los textos no analizados aquí. Sin embargo, estos procesos de limpieza son extremadamente costosos en términos de tiempo, y lo que realmente se detecta con este estudio es la necesidad de que las herramientas tecnológicas y algoritmos de análisis de discurso existentes soporten este tipo de características para poder analizar en profundidad la relación entre redes sociales y depresión a nivel discursivo.

6. Conclusiones

En resumen, este trabajo muestra un análisis en términos de coherencia discursiva de los textos escritos por distintos grupos de sujetos. Por tanto, se han conseguido los objetivos marcados acerca de buscar patrones que puedan ser usados para detectar índices de depresión y desarrollando software representando colecciones de textos disponibles en el área de análisis de lenguaje y trastornos depresivos.

Fruto de este estudio se ha observado como patrón general y diferenciador, apoyado por numerosas métricas, que los textos producidos por el grupo de positivos en depresión en redes sociales tienden a tener una longitud mayor y por ende tienen una mayor complejidad que los textos emitidos por persona sin diagnóstico de depresión. Los textos además muestras estructuras más farragosas y enrevesadas, y la presencia de relaciones de coherencia (*Enablement*, *Contrast*, *Elaboration*…) que pueden soportan esta conclusión. Además estos textos se pueden leer con un nivel de lectura comprensible para lectores sin experiencia. Asimismo con la detección de ciertas palabras dentro de los ítems centrales del corpus de los casos positivos de depresión, se podrían inferir ciertas tendencias dentro de esa población.

En base a esto, es necesario validar los patrones identificados en otros corpus para comprobar el grado de generalización de estos. Estos corpus deben ser, en un primer momento, similares en contenido y usuarios (por ejemplo otros corpus de Reddit para depresión), y en un futuro corpus con fuentes de otras redes sociales.

Este trabajo da cabida a más oportunidades de investigación, pudiendo en un futuro arrojar más patrones con respectos a textos relacionados a este tipo de patologías, por ejemplo 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 realizar más estudios para extraer más patrones dentro de los corpus de clase positiva o sin diagnóstico 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. También, hay que ampliar los conceptos y teoría aplicados a textos escritos en otros idiomas. Por esto hay que estudiar con cuidado si se pueden extrapolar los patrones y métricas a otras lenguas y contextos de redes sociales y de usuarios.

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] 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.
[3] Joty Shafiq, Carenini Giuseppe, Ng Raymond T. (2015). *CODRA: A Novel Discriminative Framework for Rhetorical Analysis.* In *Computational Linguistics*. 41(3). 385–435.
[4] Lin Liu, Sidani, Jaime, Shensa Ariel, Radovic Ana, Miller Elizabeth, Colditz Jason, Hoffman Beth, Giles Leila, Primack, Brian. (2016). *Association between Social Media Use and Depression among U.S. Young Adults.* In *Depression and anxiety*. 33(4):323-331.
[5] 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.
[6] 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.
[7] 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.
[8] Ríssola Esteban, Losada David, Crestani Fabio. (2019). *Discovering Latent Depression Patterns in Online Social Media.* In *10th Italian Information Retrieval Workshop.* 13-16.
[9] 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.
[10] 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.
[11] 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.
[12] Mann William, Taobada Maite. (2018). *The RST Website*. Simon Fraser University. http://www.sfu.ca/rst/index.html
[13] van Dijk, Teun A. (1981). *Studies in the Pragmatics of Discourse.* In *Canadian Modern Language Review*. 40(4): 653a–654
[14] Bateman John A., Jan Rondhuis Klaas. (1997) *Coherence relations: Towards a general specification.* In *Discourse Processes.* 24:1, 3-49.
[15] Büring Daniel (2011). *Topic and Comment.* In *The Cambridge Encyclopedia of the Language Sciences.*
[16] Cassell Justine, Nakano Yukiko, Bickmore Timothy W., Sidner Candace L., Rich Charles. (2001). *Non-verbal cues for discourse structure*. In *Proceedings of the 41st Meeting of the Association for Computational Linguistics.* 17-19
[17] 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

[18] 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
[19] 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.
[20] 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.
[21] 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.
[22] 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.
[23] Sparck Jones K. (1972). *A Statistical Interpretation Of Term Specificity And Its Application In Retrieval,* In *Journal Of Documentation*. 28(1): 11-21.
[24] Fundación Pública Galega Centro Tecnolóxico de Supercomputación de Galicia. *Centro de Supercomputación de Galicia.* Centro de Supercomputación de Galicia. https://www.cesga.es/cesga/
[25] 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.
[26] Kincaid J. Peter, Fishburne Robert P., Rogers Richard L., Chissom Brad .S. (1975). *Derivation of new readability formulas (automated readability index, fog count, and flesch reading ease formula) for Navy enlisted personnel*. In *Research Branch Report*. 8–75.
[27] Luhn, Hans Peter (1957). *A Statistical Approach to Mechanized Encoding and Searching of Literary Information*. In *IBM Journal of Research and Development.* 1 (4): 309–317.