UNIVERSIDAD AUTÓNOMA METROPOLITANA
Casa abierta al tiempo
Unidad Cuajimalpa

Universidad Autónoma Metropolitana

División de Ciencias de la Comunicación y Diseño

Textos de usuarios con depresión: atributos que los representan y posible detección

Proyecto Terminal

Licenciatura en Tecnologías y Sistemas de Información

Presenta:

TONANTZIN ANGÉLICA SIUROB PALOMERO

Director: Dr. Esaú Villatoro Tello

Co-Directora: Mtra. Adriana Gabriela Ramírez de la Rosa

Cuajimalpa, Ciudad de México, México, Julio 2019.

Universidad Autónoma Metropolitana
Licenciatura en Tecnologías y Sistemas de Información

Textos de usuarios con depresión: atributos que los representan y posible detección

Proyecto Terminal realizado por:

TONANTZIN ANGÉLICA SIUROB PALOMERO

Cuajimalpa, Ciudad de México, México, Julio 2019.

Universidad Autónoma Metropolitana
Licenciatura en Tecnologías y Sistemas de Información

Contenido

Lista de Figuras

Lista de Tablas

I

Planteamiento del problema de investigación

Capitulo 1: Planteamiento del problema de investigación

1.1. Introducción

La depresión, según la Organización Mundial de la Salud OMS, 2018, es un trastorno mental frecuente y se calcula que afecta a más de 300 millones de personas en el mundo. Se caracteriza por la presencia de tristeza, pérdida de interés o placer, sentimientos de culpa, falta de autoestima, trastornos del sueño o del apetito, sensación de cansancio y falta de concentración. Puede causar gran sufrimiento y alterar las actividades laborales, escolares y familiares. En el peor de los casos puede llevar al suicidio, siendo éste la segunda causa de muerte en las personas entre 15 y 29 años.

Los trastornos mentales en México constituyen un importante problema de salud pública. En relación con los años de vida perdidos por incapacidad, el trastorno depresivo ocupa el primer lugar en mujeres y el quinto en hombres OMS-IESM, 2011.

Dentro del Informe de la evaluación del Sistema de Salud Mental en México OMS-IESM, 2011 se indica que la Secretaría de Salud , de su presupuesto asignado (121.67 Mil Millones Transparencia-Presupuestaria, 2019) destina únicamente el 2 % del mismo a la salud mental y, de esa parte, el 80 % se asigna a la manutención de hospitales psiquiátricos. Esto resalta una limitada inversión gubernamental para el diagnóstico y la atención temprana de los trastornos mentales.

En materia de salud los términos de mortalidad, duración, severidad, comorbilidad y discapacidad asociados a enfermedades actualmente se agrupan en un solo indicador denominado “carga de enfermedad” Lozano y col., 2014.

La depresión por sí sola equivale 4.3 % de la carga mundial de enfermedad y esto se traduce como el 11 % de todos los años vividos con discapacidad a escala mundial, de acuerdo con información de la Secretaría de Salud SSA, 2015. Adicionalmente se estima que cuatro de cinco personas con trastornos mentales importantes no reciben atención en los países de ingreso bajo y medio.

Así, conforme lo mostrado en el Programa de Acción Específico (PAE) en materia de Salud Mental SSA, 2015: uno de cada cuatro mexicanos entre 18 y 65 años habrá padecido algún trastorno mental en algún punto de su vida, pero sólo uno de cada cinco recibirá tratamiento y tardará de 4 a 20 años en ser atendido en un centro de salud público.

El PAE SSA, 2015, en su objetivo 1 establece: *Desarrollar acciones de protección y promoción de salud mental, así como detección oportuna y prevención de los trastornos mentales*. El objetivo anterior indica que la detección oportuna de los trastornos mentales es una prioridad dentro de las políticas públicas nacionales.

Desde la perspectiva de la psicología, el Instituto Nacional de Salud Mental de los Estados Unidos NIMH, 2009 informa que para el tratamiento de la depresión, la terapia cognitivo-conductual es de gran ayuda. Permite al individuo concebir y cultivar nuevas maneras de pensar y de actuar; para identificar, comprender y cambiar aquellas ideas, actitudes y conductas que le dañan y avivan los episodios depresivos.

De este modo, el especialista en salud mental acompaña al individuo a descubrir sus

patrones y hábitos a través de un exhaustivo análisis de las motivaciones, emociones y actos presentes en su vida cotidiana. Es de esta forma que la expresión de las ideas a través del lenguaje entra en juego, como instrumento fundamental para el reconocimiento de tendencias de un individuo.

Considerando que el lenguaje es la forma más común y confiable con la que alguien cuenta para traducir sus pensamientos y emociones de manera que los demás puedan entenderlos Tausczik y Pennebaker, 2010.

En su libro **La vida secreta de los pronombres** J. Pennebaker, 2011 dice que "Las emociones cambian la forma en que la gente ve el mundo y lo que piensa del mismo [...] Las emociones guían nuestro pensamiento y afectan la forma en que hablamos con otros y nos llevamos bien con ellos."

Si las emociones cambian la forma en la que se percibe el mundo y el lenguaje permite explicar a otros los estados mentales, es posible encontrar emociones predominantes y por ejemplo, averiguar si la tristeza está presente en lo expresado en el uso del lenguaje.

Pero, ¿cómo se pueden detectar las emociones de la gente a través de las palabras? J. Pennebaker, 2011 dice "[...] parece una tarea tan simple como asumir que alguien usa palabras alegres cuando está alegre o palabras tristes si está triste [...] Ésta puede ser una aproximación inicial pero, perdemos de vista un punto central: **las emociones afectan la manera de pensar de las personas.**"

Quizá vale el esfuerzo tener claridad sobre los tipos de palabras empleadas al escribir. Contemplando primero dos grandes grupos: por un lado están las palabras que tienen significado o carga semántica y por otro, aparecen las palabras funcionales (que reflejan un estilo de escritura como las preposiciones, artículos, ...). Considerando que, al investigar sobre la presencia de dichas palabras en los textos, aparece la posibilidad de reconocer patrones ahí contenidos.

J. Pennebaker, 2011 afirma que, se pueden usar las palabras funcionales para seguir la pauta de las maneras de pensar y, por lo tanto, también los estados emocionales. La tristeza vuelve más introspectivo al individuo, los pronombres toman importancia y se emplean mucho las referencias al yo, la persona se enfoca en el pasado o en el futuro.

Medir características de un texto al contar las palabras empleadas, el tipo de palabras que aparecen, conocer el tamaño del léxico de un conjunto de textos son tareas relacionadas con el Análisis de Textos. Si se cuenta con una cantidad de textos generados por algún individuo, se pueden tomar como ejemplos y llevarlos a representaciones adecuadas que posibiliten un análisis automático mediante modelos computacionales y permitan determinar si los textos de un autor en específico presentan señales de depresión.

Las redes sociales son cada vez más utilizadas, según el estudio sobre los hábitos de los Usuarios de Internet en México AIMx, 2019 en México el 82 % de los usuarios de internet tiene como principal actividad acceder a redes sociales, el 99 % de ellos tiene un perfil en Facebook y para 31 % de su tiempo conectado a internet en alguna red social.

Mediante el uso de redes sociales los individuos comparten sus opiniones y pensamientos con sus contactos. Hacer publicaciones en redes sociales forma parte de las actividades cotidianas, lo que provee un recurso que captura atributos de comportamiento relevantes a características del individuo tales como: su manera de pensar, estado de

ánimo, comunicación, actividades y socialización De Choudhury, Gamon, Counts y Horvitz, 2013.

Las emociones y el lenguaje empleados en las publicaciones en redes sociales indican sentimientos de inutilidad, culpa, impotencia o autorechazo que son característicos de un episodio de depresión. De forma similar Moreno y col., 2011 ha demostrado que las actualizaciones en los estados de Facebook pueden revelar síntomas correspondientes a episodios de depresión mayor.

De esta forma el material extraído de redes sociales, que contiene los síntomas de estados depresivos presentes en la enorme cantidad de publicaciones realizadas en redes sociales, sirve de insumo a tareas de análisis de textos; que en este trabajo corresponde a la determinación de atributos y representaciones que permitan la detección de depresión en textos.

Todos aquellos usuarios de redes sociales que, aunque no acudan con los especialistas en salud mental por diversos factores personales y sociales, se pueden ver favorecidos con la detección de episodios de depresión por medio de textos. Por ejemplo, según Moreno y col., 2011,los estudiantes universitarios en riesgo dan pistas a investigadores y especialistas en salud mental cuando crean un perfil en redes sociales porque muestran aspectos de su comportamiento que no siempre son aparentes en su vida fuera de las redes sociales.

Finalmente, la motivación del presente trabajo es colaborar en los objetivos planteados en el Programa Nacional de Salud Mental al proveer una herramienta de apoyo a los expertos en salud mental dentro de sus labores de diagnóstico, explorar la posibilidad de crear sistemas automáticos que coadyuven a la detección de indicios textuales correspondientes a los síntomas del trastorno mental depresivo. Lo anterior, a través de evaluar hasta qué punto son pertinentes las técnicas y estrategias empleadas por el Procesamiento de Lenguaje Natural y la Clasificación de Textos en la identificación de elementos que permitan hacer visibles algunos rasgos depresivos en textos.

1.2. Objetivos

1.2.1. Objetivo general

Determinar la pertinencia de la información psicolingüística en la identificación de usuarios con depresión a través de técnicas de lingüística computacional.

1.2.2. Objetivos específicos

- Examinar los textos de usuarios con depresión y sin depresión a través del estudio estadístico de un corpus específico para seleccionar aspectos relevantes que los distingan.

- Replicar el uso de las representaciones de los textos que han mostrado ser más eficientes en la detección temprana de depresión, con técnicas de aprendizaje automático, y diseñar representaciones derivadas de los aspectos distintivos seleccionados.

- Experimentar, por medio de la clasificación de textos y algoritmos de clasificación específicos, con las representaciones diseñadas para determinar las ventajas y desventajas de cada representación.
- Analizar cuantitativa y cualitativamente el desempeño de los atributos considerados dentro de las representaciones en los resultados experimentales para elegir atributos distintivos de la depresión.

Este trabajo está dividido de la siguiente forma: en la sección 2 que corresponde al Marco Teórico se encontrarán todos aquellos conceptos necesarios para una mejor comprensión del presente documento: tales como Aprendizaje Automático, Clasificación de Textos, Formas de representación o Métricas y formas de evaluación . La sección 3 corresponde al Estado del Arte y presenta los puntos principales de los trabajos relacionados con la detección de depresión en textos. En la sección 4 se describen las estadísticas generadas sobre el corpus y el análisis correspondiente a la selección de atributos6ein los experimentos y las conclusiones a las que llega este trabajo.

II

Marco teórico

2.1. Aprendizaje Automático (Machine Learning)

El Aprendizaje Automático es una rama de la inteligencia artificial que tiene por objetivo permitirle a las máquinas recuperar patrones encontrados en las instancias de entrada para integrarlos a su toma de decisiones y a realizar ciertas tareas, basándose en métodos de aprendizaje estadístico Mohammed, Khan y Bashier, 2016. Machine Learning (ML) también puede ser definido como aquellos métodos computacionales que emplean la *experiencia* para mejorar su desempeño o para hacer predicciones exactas. Así, en este contexto, experiencia quiere decir: datos electrónicos disponibles para hacer el análisis y que provienen de los conjuntos de entrenamiento ya etiquetados (corpus). El Aprendizaje Automático se centra en el diseño de ***algoritmos de predicción*** eficientes y precisos Mohri, Rostamizadeh y Talwalkar, 2018. Se pueden encontrar diferentes técnicas empleadas en el Aprendizaje Automático: Aprendizaje Supervisado, No Supervisado, Semi-supervisado y por refuerzo. En el presente trabajo es de particular interés el aprendizaje supervisado.

2.2. Aprendizaje Supervisado

En la técnica de *aprendizaje supervisado* o *aprendizaje con maestro*, los datos disponibles son proporcionados en parejas entrada-salida. Particularmente, cada muestra de datos consiste de un vector de entrada específico y su valor de salida relacionado.

El propósito principal de este paradigma de aprendizaje es encontrar una función que genere el valor de salida correcto para un patrón de entrada dado. El término aprendizaje supervisado proviene del hecho de que los objetos empleados en el proceso de entrenamiento ya están asociados con un valor objetivo (clase o etiqueta) el cual puede tomar un valor entero representativo de la clase a la que pertenecen los valores reales Sotiropoulos y Tsihrintzis, 2016.

2.3. Clasificación de Textos

El aprendizaje automático se encarga de resolver diferentes problemas y uno de los más importantes es la *clasificación de textos* o *categorización de textos*, lo cual involucra organizar documentos en diversas categorías basadas en propiedades inherentes de los textos. La clasificación de documentos es un problema genérico que también se aplica a otro tipo de objetos como música, vídeo u otros medios Sarkar, 2016.

La forma más común es la *clasificación binaria*, es decir, asignar una de dos categorías a todos los documentos en un corpus. Entonces, básicamente para la clasificación de textos se extraen un conjunto de características (features) que describan a los documentos del corpus y después, se aplica a los mismos un algoritmo diseñado para procesar y usar dichas características o atributos y así poder seleccionar la categoría apropiada de un documento específico Miner y col., 2012.

2.4. Procesamiento de Lenguaje Natural

El término **Procesamiento de Lenguaje Natural (PLN)** comúnmente es usado para describir la función de componentes de software o hardware dentro de un sistema computacional el cual analiza o sintetiza lenguaje hablado o escrito. El epíteto 'natural' tiene la intención de distinguir el habla y la escritura humanas de lenguajes más formales como las matemáticas o lenguajes de programación Jackson y Moulinier, 2002.

PLN se relaciona con modelos o formalismos lingüísticos que son desarrollados para la implementación de proyectos de Tecnología de la Información como: software que nos permite procesar el lenguaje o técnicas de procesamiento de señales para el reconocimiento de unidades lingüísticas (por ejemplo fonemas, sílabas o palabras). De esta forma, podemos afirmar que una parte del campo de PLN se encarga del desarrollo de herramientas para el entendimiento del lenguaje natural. Por ejemplo: etiquetadores de partes de la oración (POS Part-Of-Speech), analizadores morfológicos, analizadores sintácticos (parsers) de diferentes tipos, entre otros Kurdi, 2016.

2.5. Corpus (Texto)

El corpus de texto o simplemente corpus, son los datos de los que dependen todas las tareas en PLN. Es un gran conjunto de datos en texto que puede estar en uno de los idiomas como inglés, francés, etcétera. Puede estar formado de un solo documento o varios de ellos (Arumugam y Shanmugamani, 2018) Arumugam y Shanmugamani, 2018.

Otra definición es que los corpora (plural de corpus) son una grande y estructurada colección de textos o datos textuales, usualmente conformada por cuerpos de texto escrito o hablado, frecuentemente almacenado de manera electrónica. Su propósito principal es ser aprovechados al máximo para análisis tanto lingüístico como estadístico y posteriormente emplearlos en la construcción de herramientas de PLN Sarkar, 2016.

Las fuentes de un corpus pueden ser los sitios de redes sociales como Twitter, sitios de blogs, foros de discusión abierta como Stack Overflow, libros, entre otros. Dependiendo de la tarea a tratar, el corpus podría estar integrado por archivos de distintos formatos como sonido, imágenes o video.

De manera general dentro de las tareas de PLN, el corpus es dividido en partes más pequeñas, conocidas como *chunk* en inglés, para poder analizarlo. Estas partes pueden ser a nivel de párrafo, enunciado, palabra o letra.

2.6. Atributos (Features)

Para poder llevar a cabo el aprendizaje supervisado y generar nuestro algoritmo de predicción, es muy importante poner especial atención tanto a la cantidad y el tipo de características a incluir como a la forma en la que las tomamos en cuenta para conformar el vector que representa cada documento, Lo anterior debido a que los algoritmos de clasificación de textos determinarán la predicción de la clase con base en la presencia o ausencia de características particulares del texto.

Inevitablemente, las palabras individuales (términos o tokens) son las características primarias a considerar. La creación de características adicionales requiere de aplicar acciones de *pre-procesamiento* al texto.

Adicionalmente se pueden usar características que no sean texto, por ejemplo la longitud del documento, cantidad de palabras por sección, cantidad de palabras en el título, cantidad de emojis, etcétera.

2.7. Formas de representación

Las características que conformen representación vectorial de los documentos pueden tomar una de las siguientes tres formas: representación binaria, entera (frecuencia de aparición del término) o flotante (peso calculado).

Otro aspecto a considerar al decidir cómo estará integrado el vector de cada documento es por medio de las formas de representación como la Bolsa de Palabras (BoW por sus siglas en inglés) o la matriz de n-gramas.

En la BoW cada token es una característica y no guarda el orden de aparición de las palabras en el texto. En la matriz de n-gramas, cada característica del vector se considerará como una subsecuencia de n términos de una secuencia dada (los documentos del corpus).

Los valores que toman las características en las matrices son asignados si la palabra (o subsecuencia de términos) aparecen en el documento o se asigna 0 cuando no aparecen.

2.8. Algoritmos de aprendizaje

Para los experimentos propuestos en este trabajo, se emplearon tres algoritmos de aprendizaje: Bayes Ingenuo Multinomial (Multinomial Naive Bayes), Árboles de decisión (Decision Tree Classifier) y Máquinas de Vectores de Soporte (Support Vector Classifier).

2.8.1. Bayes Ingenuo Multinomial (Multinomial Naive Bayes)

Los métodos Bayes ingenuos son un conjunto de algoritmos basados en la aplicación del teorema de Bayes con la suposición "ingenua" de independencia condicional entre cada par de características dado el valor de la clase variable. El teorema de Bayes establece la siguiente relación, dada la clase variable y y el vector de características dependiente x_i hasta x_n:

$$P(y|x_1, \ldots, x_n) = \frac{P(y)P(x_1, \ldots, x_n|y)}{P(x_1, \ldots, x_n)}$$

Usando la siguiente suposición ingenua de independencia condicional

$$P(x_i|y, x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n) = P(x_i|y),$$

Para toda i, esta relación es simplificada a

$$P(y|x_1,\ldots,x_n) = \frac{P(y)\prod_{i=1}^{n} P(x_i|y)}{P(x_1,\ldots,x_n)}$$

Como $P(x_1,\ldots,x_n)$ es constante dada la entrada, podemos usar la siguiente regla de clasificación:

$$P(y \mid x_1,\ldots,x_n) \propto P(y)\prod_{i=1}^{n} P(x_i \mid y)$$

$$\Downarrow \qquad (2.1)$$

$$\hat{y} = \arg\max_{y} P(y)\prod_{i=1}^{n} P(x_i \mid y),$$

y podemos usar la estimación Maximum A Posteriori (MAP) para estimar $P(y)$ y $P(x_i|y)$; el primero es entonces la frecuencia relativa de la clase y en el conjunto de entrenamiento.

MultinomialNB implementa el algoritmo Bayes ingenuo para datos multinomialmente, y es una de las dos variantes ingenuas clásicas usadas en la clasificación de textos (donde los datos están típicamente representados como vectores de conteos de palabras, aunque se sabe que los vectores tf-idf funcionan bien en la práctica). La distribución tiene como parámetros los vectores $\partial_y = (\partial_{y1},\ldots,\partial_{yn})$ para cada clase y, donde n es el número de características (en clasificación de textos, el tamaño del vocabulario) y ∂_{yi} es la probabilidad $P(x_i|y)$ de que la característica i aparezca en una muestra perteneciente a la clase y.

Los parámetros ∂_y son estimados a través de una versión suavizada de máxima verosimilitud, por ejemplo para el conteo relativo de frecuencias:

$$\hat{\partial}_{yi} = \frac{N_{yi} + a}{N_y + an}$$

donde $N_{yi} = \sum_{x\in T} x_i$ es el número de veces que la característica i aparece en una muestra y en el conjunto de entrenamiento T, y $N_y = \sum_{i=1}^{n} N_{yi}$ es la cuenta total de todas las características para la clase y.

El suavizado anterior $a \geq 0$ cuenta para las características no presentes en las muestras de aprendizaje y previene probabilidades con valor de cero en los siguientes cálculos. Cuando estableces $a = 1$ es llamado suavizado Laplace, mientras que establecer $a < 1$ es llamado suavizado Lidstone.

2.8.2. Árboles de decisión (Decision Tree Classifier)

Un árbol de decisión es uno de los más conocidos y usados en aprendizaje automático supervisado para llevar a cabo tareas tanto de regresión como de clasificación. Para cada atributo del conjunto de datos, el algoritmo de árbol de decisión forma un nodo, donde el atributo más importante es colocado en el nodo raíz. El proceso comienza en el nodo raíz

y se hace el recorrido a través de árbol siguiendo el nodo correspondiente que cumpla con la condición o "decisión". Este proceso continúa hasta que un nodo hoja es alcanzado, el cual contenga el resultado del árbol de decisión.

Por ejemplo, en un escenario donde una persona pide prestado un carro por un día, el dueño tiene que decidir si lo presta o no. Existen varios factores que ayudarán a tomar la decisión:

1. ¿Esta persona es un amigo cercano o sólo una conocida del dueño? Si la persona es sólo una conocida, entonces la solicitud será rechazada; si la persona es un amigo, entonces se continúa con el siguiente paso.
2. ¿La persona está pidiendo por primera vez el carro? En caso que si, se acepta a prestarle el carro, de otro modo se continúa con el siguiente paso.
3. ¿El carro estaba dañado la última vez que la persona lo regresó? En caso afirmativo, se niega el préstamo; En caso negativo, se le presta el carro.

Para este ejemplo el árbol de decisión se ve así:

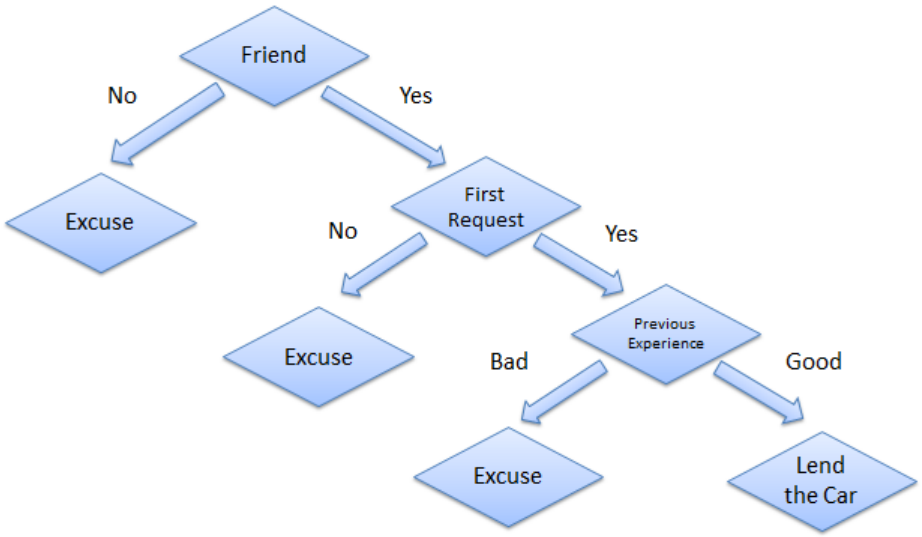

Figura 2.1: *Ejemplo de estructura de un árbol de decisión*

2.8.3. Máquinas de Vectores de Soporte (Support Vector Classifier)

Dado un conjunto etiquetado de datos (en aprendizaje supervisado), el algoritmo proporciona como salida un hiperplano óptimo el cual categoriza nuevos ejemplos. En un espacio de dos dimensiones este hiperplano es una línea que divide un plano en dos partes, separando en cada lado cada clase.

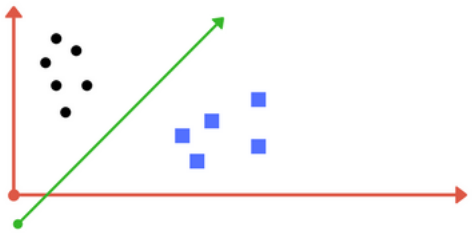

Figura 2.2: *Ejemplo de hiperplano que separa elementos de dos clases*

Cuando tenemos datos que no es posible separar por medio de una recta, se aplica una transformación para agregar una dimensión más y se busca obtener aquella recta que sea capaz de separar los datos.

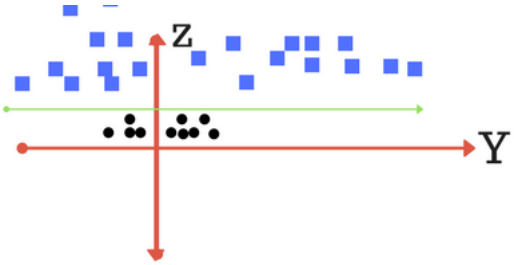

Figura 2.3: *Ejemplo de hiperplano generado de una tranformación a tres dimensiones*

Después, se transforma de regreso esa línea al plano original y esas transformaciones son llamadas **kernels**. Cuenta también con un parámetro de regularización **C** que permite qué tanto se quiere evitar clasificar erróneamente cada ejemplo de entrenamiento. Cuando C es mayor se tiene un margen más pequeño y cuando C en menor se cuenta con un margen más grande. Otro parámetro es **gamma**, que define qué distancia de separación entre un punto a clasificar y el hiperplano debe existir para etiquetarlo con una clase determinada. Finalmente se encuentra el **margen** que es la separación que existe entre la línea del hiperplano y los puntos a clasificar más cercanos. Donde un buen margen es equidistante tanto a los puntos de una clase, como de la otra.

2.9. Métricas y formas de evaluación

Para hacer la evaluación del desempeño del clasificador hay dos opciones:

- Contar con dos partes del corpus: una para entrenamiento y la otra para prueba (propiamente la validación).
- Validación cruzada de n pliegues: El corpus completo es dividido en n subconjuntos de igual tamaño y el proceso de entrenamiento – prueba se ejecuta n veces, usando una parte diferente de las divisiones hechas como conjunto de prueba por cada vuelta. Para que al final los resultados de cada vuelta sean promediados.

Para los problemas de clasificación binaria se recurre a las métricas: Accuracy (Exactitud), Recall (recuperación), Precision (precisión) y F1.

Considerando el desempeño del clasificador de la siguiente forma:

	Etiqueta Positivo	**Etiqueta Negativo**
Clasificador Positivo	a	b
Clasificador Negativo	c	d

Tabla 2.1: *Representación de resultados para métricas de evaluación*

$$Accuracy = \frac{a+b}{a+b+c+d}$$

$$Recall(R) = \frac{a}{a+c}$$

$$Precision(P) = \frac{a}{a+b}$$

$$F1 = \frac{2PR}{P+R}$$

III

Estado del Arte

Teniendo en cuenta la ya mencionada prevalencia de los trastornos depresivos, su baja detección y atención temprana, ha incrementado el interés de especialistas en PLN así como el número de planteamientos e investigaciones realizadas sobre este tema.

3.1. Características basadas en Información sintáctica

Hay trabajos que han empleado múltiples métodos para resolver la tarea de clasificación de textos de usuarios con depresión y han tomado múltiples consideraciones para las representaciones de los textos.

A. En el trabajo de Villegas, Funez, Ucelay, Cagnina y Errecalde, 2017, se emplean representaciones con BoW con pesos booleano, tf y tf-idf y también se contemplan n-gramas de 3 caracteres para la construcción de su matriz término-documento (document-term matrix -dtm-). Su baseline en F1 measure con representación de BOW es de 0.24.

B. En la investigación de Almeida, Briand y Meurs, 2017, los autores decidieron usar representación en BOW junto con bigramas. Se consideran también elementos POS (categorías gramaticales) como: adjective (JJ), noun (NN), predeterminer (PDT),particle (RP) y verb(VB) .

C. En la aproximación de Trotzek, Koitka y Friedrich, 2017 se determina qué tokens proporcionan más ganancia de información (Information Gain -IG-) dentro de los uni-, bi- y trigramas encontrados en al menos dos documentos. Se generó un ensamble con representación BOW y diferentes pesos del tipo tf-idf, todos normalizados .

F. Farias-Anzaldua, Montes-y-Gómez, López-Monroy y González-Gurrola, 2017, presentan en su trabajo una perspectiva de clasificación en dos pasos. Primero se considera el análisis de las publicaciones individualmente para discernir los mensajes producidos por cada clase de usuario. El segundo paso pretende modelar el comportamiento de los usuarios con base en su categoría. Emplean uni-, bi- y trigramas.Conservan los 10,000 items más frecuentes y se seleccionan aquellos que proporcionen una IG mayor a cero.

3.2. Características basadas en Información semántica

A. Al usar información semántica, Villegas y col., 2017 emplean características basadas en listados de palabras como LIWC J. W. Pennebaker, Francis y Booth, 2019 considerando aquellas palabras que indican emociones positivas, negativas y relacionadas con la muerte (familias Emopos, Emoneg y Muerte; respectivamente). Cabe señalar que emplean un método denominado Temporal Variation of Terms, el cual través de vectores de conceptos permite capturar las variaciones en los textos de los

individuos. Logrando un resultado en F1 measure de 0.59 con el uso de métodos combinados.

B. Por parte de Almeida y col., 2017, se toman en cuenta varios diccionarios conformados por listados de palabras relacionadas con depresión: sentimientos, nombres de medicamentos o enfermedades mentales. Se mezclaron el uso de atributos *sintácticos, semánticos y de comportamiento*.

C. Adicionalmente, Trotzek y col., 2017 calcularon características sobre nombres de medicamentos, menciones explícitas del diagnóstico que incluían la palabra ***depresión***. Su modelo final toma en cuenta el desequilibrio de las clases penalizando a los falsos negativos. El resultado de esta configuración de F1 measure es: 0.64.

D. En el trabajo de Sadeque, Xu y Bethard, 2017 se consideran dos tipos de características: el primero basado en vocabulario relacionado con depresión y el segundo en conceptos extraídos con Metamap. Para el primer tipo se utilizaron las 110 palabras más frecuentes extraídas de las palabras más asociadas a la raíz *deprimir* en el foro de salud mental de Yahoo! Answers. Recuperaron 404 Identificadores Únicos de Conceptos (CUI) como características para cada post. De esta configuración se observa un resultado de F1 measure de: 0.40.

E. En la investigación de Malam y col., 2017 se establecen características de tipo semántico y estadísticas. Las características de tipo semántico contemplan entre otros aspectos: Uso de referencias a sí mismo, términos de generalización, análisis de sentimiento, análisis de emociones, síntomas de depresión o palabras en pasado. Dentro de los atributos calculados de tipo estadístico se encuentran: las variaciones en el número de publicaciones, medias de publicaciones, palabras, comentarios o palabras por comentario. Se obtiene un resultado de F1 measure de: 0.47.

3.3. Características basadas en comportamiento

B. Almeida y col., 2017, adicionalmente calculan atributos con base en las frecuencias de las publicaciones. Todas las configuraciones ya descritas (sintácticas, semánticas y de comportamiento) y empleadas, les devuelve un resultado de F1 measure de 0.42.

F. Se observa en Farias-Anzaldua y col., 2017 que juntan las características sintácticas mencionadas en la sección 3.1 con dos atributos de tiempo (de comportamiento): uno la hora del post y otro binario indicando si éste se produjo o no en fin de semana. Estas características de comportamiento capturan las proporciones de publicaciones depresivas, no depresivas, los horarios (mañana, tarde y noche) en los que se generaron los posts y los posibles cambios u oscilaciones entre clases. El resultado de F1 measure obtenido es: 0.48.

3.4. Conclusiones sobre los trabajos consultados

En general, las aproximaciones que están compuestas por demasiados pasos tienden a complicar el proceso de clasificación y en consecuencia es difícil determinar qué tan bien funciona cada atributo seleccionado dentro de las representaciones propuestas.

Prácticamente todos los trabajos consultados emplean recursos basados en listados de palabras, la mitad emplea algún tipo de rastreo del comportamiento y sólo uno hacer clasificaciones por publicación (considerando el tamaño de la instancia como cada publicación del sujeto).Los trabajos referidos en esta sección, no toman en cuenta aspectos como las faltas de ortografía o slang/calo que emplean los individuos al escribir sus textos.

Tampoco se contempla en estos trabajos la búsqueda de respuestas (parciales o totales) a cuestionarios de diagnóstico clásico de depresión,los cuales incluyen preguntas como: ¿Tiene problemas para dormir? ¿Duerme demasiado? ¿Poco apetito? ¿Tiene problemas de concentración? (Kroenke, Spitzer y Williams, 2001).

Por último dentro de estas aproximaciones se excluye consideración alguna de la sintomatología de la depresión en hombres, cuyas demostraciones suelen ser por ejemplo la búsqueda de situaciones extremas o de riesgo (Pérez, Castro y Rodrıguez, 2017), por lo que deberían incluirse atributos que representen el género del sujeto.

Para la presente investigación, se consideran de interés las características basadas en información del tipo ***sintáctico, semántico y de comportamiento***. Dentro de aquellas consideradas como características de información sintáctica, se incluyen las categorías gramaticales del POS Treetagger descritos en la sección 4.2.4. En el caso de las contempladas como características de información semántica están las basadas en las familias de LIWC, explicadas en la sección 4.2.3. Y por último, las características de información de comportamiento se generan con base en la fecha y hora de emisión de las publicaciones y se detallan en la sección 4.2.2. En la tabla 3.1, se muestra un resumen sobre las características incluidas en las diferentes representaciones consideradas como Estado del Arte. La columna denominada ***G***, hace referencia al presente trabajo de investigación.

	A	B	C	D	E	F	G
Características de comportamiento		✓	✓			✓	✓
Características sintácticas		✓	✓		✓		✓
Características semánticas (diccionarios de palabras)	✓	✓	✓	✓	✓		✓
Características con pesos (binario, tf o tf-idf)	✓	✓	✓				
Características estadísticas del texto			✓		✓	✓	
Análisis a nivel post						✓	
Resultado F measure	0.59	0.48	0.64	0.45	0.47	0.48	

Tabla 3.1: *Comparativa de características utilizadas en los trabajos relacionados con el presente Proyecto Terminal.*

La presente propuesta se centró en las características de comportamiento, semánticas

y sintácticas porque al incluirlas como representaciones y poder interpretarlas como patrones conductuales, cobran sentido los resultados encontrando pautas que favorezcan la clasificación de textos de personas con depresión.

Todas las aproximaciones para la generación de las características provienen de intuiciones sobre la conducta que presentan los individuos que atraviesan por un episodio depresivo: entonces las características incluidas en el presente trabajo tienen coherencia desde una perspectiva psicológica y del comportamiento humano.

Con las características denominadas de Comportamiento se pretende examinar si en la noche los sujetos con depresión escriben más, pensando que en ese momento el entorno les incitara a hacerlo. O tal vez tuvieran espacios de mayor soledad durante el invierno y fuera más propicio externar sus estados emocionales a través de las publicaciones en redes sociales.

En los trabajos referidos en esta sección, sólo la mitad incorporan características de comportamiento en sus representaciones. Es así que la presente propuesta se distingue al incorporar características calculadas a partir de los tiempos en los que las publicaciones son realizadas. Añadiendo varios niveles de detalle en estos períodos de tiempo que van desde las estaciones del año en que se emiten las publicaciones hasta el rango del día en que las mismas se generan. Además, se procedió de una forma diferente al calcular las características de comportamiento estableciendo las representaciones en términos de los porcentajes de publicaciones realizadas en los distintos niveles contemplados.

Respecto de las características de origen semántico, en la presente propuesta se intenta indagar si es que los individuos que tienen episodios depresivos son más autocentrados. Buscando qué tantas referencias hacen hacia sí mismos y revisando si es que sus publicaciones se ven impregnadas con palabras sobre temas depresivos.

Las características semánticas brindan consistencia con los trabajos del estado del arte. El presente trabajo se distingue porque en el cálculo de las características semánticas se obtuvieron porcentajes de uso de palabras por familia de LIWC y aunque no es posible determinar exactamente la manera en que se han generado las características semánticas en los trabajos consultados, si se observa que la mayoría de los mismos emplea alguna clase de listado de palabras temáticas.

Este trabajo se caracteriza al incluir las características de tipo sintáctico en que analiza cómo es que los individuos con depresión escriben sus publicaciones, es decir, averiguar qué tipo de palabras son las que usan más. Saber qué tipo de palabras hacen diferentes los mensajes entre las personas con depresión y sin depresión.

Se distingue de lo que se ha revisado hasta ahora en que, en el presente trabajo las características sintácticas contempladas para crear las representaciones de los textos salen de la tradicional bolsa de palabras para incorporar una representación de los textos como unigramas de categorías gramaticales POS y juntándolo también con los porcentajes de uso por categoría gramatical.

IV

Método

Capitulo 4: Método

El presente trabajo tiene como principal finalidad colaborar en la detección de rasgos indicadores de depresión en textos de usuarios diagnosticados con depresión. Se busca explorar diversas representaciones de textos que sirvan como entrada para algoritmos de clasificación.

Dichas representaciones serán generadas con base en informaciones semánticas, sintácticas y de comportamiento presentes en los textos, considerando la hipótesis general de que la información psicolingüística es útil para capturar rasgos depresivos, cuando dichos rasgos existen en los textos de un individuo.

La hipótesis antes mencionada, se ha explorado en múltiples investigaciones de acuerdo con J. W. Pennebaker, Mehl y Niederhoffer, 2003, en un estudio de habla espontánea de cinco individuos deprimidos de edad madura Bucci y Freedman, 1981 encontraron que la depresión estaba relacionada con un uso elevado de pronombres de primera persona singular y una ausencia de pronombres de segunda y tercera persona.

De forma similar Weintraub, 1981 encontró que cuando a la gente deprimida se le pide hablar acerca de cualquier tema personal por 10 minutos, usan la palabra "yo" a una tasa mayor que individuos sanos. Lo que es confirmado por Rude, Gortner y Pennebaker, 2004, que en su estudio, estudiantes que estaban deprimidos usaron significativamente más pronombres de primera persona del singular en sus ensayos personales que aquellos estudiantes que nunca se habían deprimido.

De este modo, en esta sección se llevará a cabo la descripción del método empleado al definir el modelo de predicción y las representaciones de los documentos, así como la evaluación de dichas representaciones al emplearse en la tarea de identificar a través de los textos las clases: *con depresión* y *sin depresión*.

4.1. Descripción del corpus

El corpus empleado en el presente trabajo pertenece a la competencia eRisk Losada, Crestani y Parapar, 2017, es una colección de escritos (publicaciones o comentarios) de usuarios de Reddit. Los usuarios etiquetados como *con depresión* son aquellos que han mencionado explícitamente su diagnóstico de depresión. En la tabla 4.1 se presenta el corpus dividido en dos subconjuntos: *entrenamiento* y *prueba*. Así como la cantidad de instancias con las que se cuenta por cada subconjunto señalado.

	Total	**con depresión**	**sin depresión**
Entrenamiento (train)	486	83	403
Prueba (test)	401	52	349
Total	887	135	752

Tabla 4.1: *Distribución de sujetos por clase en el corpus*

4.1.1. Estadísticas de Comportamiento

Las estadísticas denominadas de comportamiento, se calculan con base en los momentos en el tiempo en los que se llevan a cabo las publicaciones de los usuarios de la clases con y sin depresión. En adelante referidos como *segmentos de tiempo*, abarcan los siguientes periodos específicos: estaciones del año (season), meses del año (month), días del mes (day) y horario del día (rangeDay). Los *segmentos de tiempo* considerados para las características de comportamiento son:

- **Season.-** Se refiere a la estación del año en la que se hacen las publicaciones. Tiene cuatro posibles valores (*spring* -mar,apr y may-, *summer* -jun,jul y aug-, *autumn* -sep, oct y nov- e *winter* -dec, jan y feb-). Este *segmento de tiempo* se ha incluido con el objetivo de indagar la posible presencia de una periodicidad específica en la aparición de episodios depresivos, que esté directamente vinculada con las estaciones del año. Evaluando si en invierno se manifiestan con mayor frecuencia, o que en verano los sujetos de la clase *con depresión* no deseen expresar tanta emoción por la temporada de sol y vacaciones como lo hacen los sujetos *sin depresión*.

- **Month.-** Toma como valores el número del mes en el que se hacen las publicaciones (1-january, 2-february, ..., 12-december). Con este *segmento de tiempo* se pretende comprobar si hay ciertos meses "detonadores" que, al involucrar ciertas celebraciones, promuevan la actividad o inactividad de los sujetos *con depresión*.

- **Day.-** Hace referencia al número de día del mes en el que se genera la publicación y toma como valores los números del 1 al 31, según corresponda. Al segmentar por día del mes, la búsqueda gira en torno a si existe alguna parte del mes en la que quizá por aumento de estrés generalizado se manifieste un aumento o disminución entre ambas clases respecto de sus publicaciones.

- **RangeDay.-** En este caso se establecieron rangos del día: morning (6:00 – 11:59), afternoon (12:00 – 17:59), night (18:00 – 23:59) y dawning (0:00 – 5:59). Para este *segmento de tiempo*, la idea es buscar si hay algun indicador de mayor presencia de publicaciones por la noche o madrugada por parte de los sujeto *con depresión*. Lo anterior considerando que parte de las preguntas del diagnóstico clásico de depresión es saber si se tienen trastornos de sueño como el insomnio, alentando a los sujetos a escribir y publicar en ese caso.

Para cada partición del corpus (Train y Test), se calcularon las proporciones de las publicaciones (expresados como porcentajes) emitidas en:

- Los horarios del día (RangeDay)
- Los días del mes (Day)
- Los meses del año (Month)
- Las estaciones del año (Season)

Analizando los porcentajes promedio de las publicaciones emitidas por ***rango del día***, mostrados en la Figura 4.1. Se pueden rescatar con diferencias más notorias, dentro de la tabla 4.2, dos rangos del día en particular: la mañana (morning) y la tarde (afternoon).

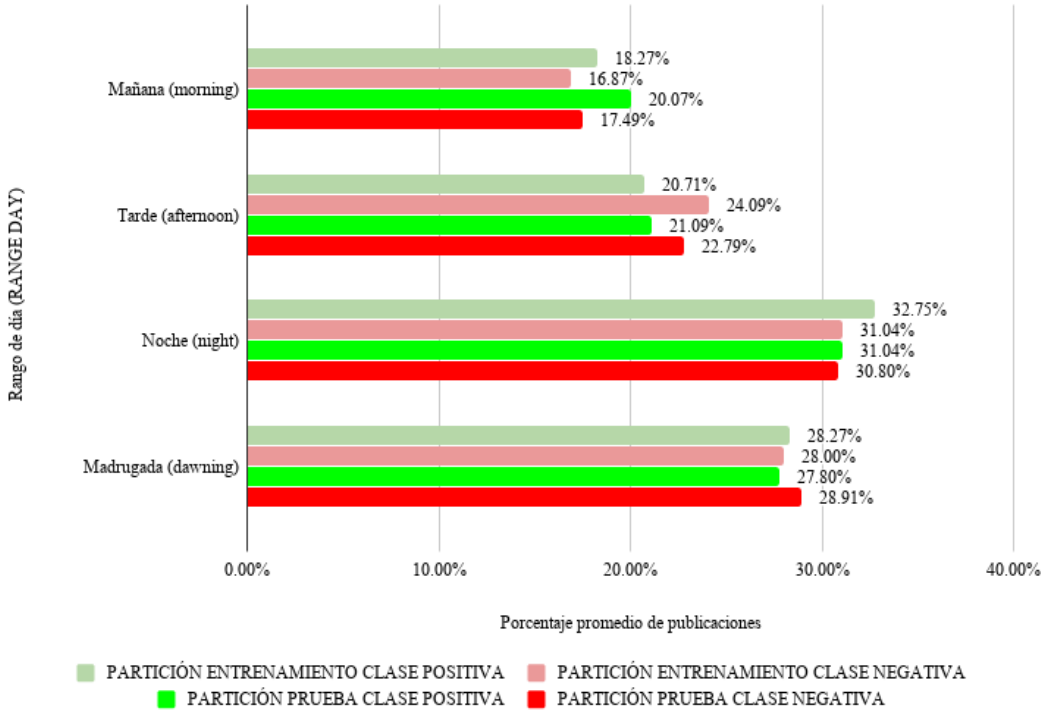

Figura 4.1: *Porcentajes promedio de publicaciones por rango del día.*

	Mañana (morning)		Tarde (afternoon)	
Clase	% Partición Entrenamiento	% Partición Prueba	% Partición Entrenamiento	% Partición Prueba
Positiva	20.71	21.09	18.27	20.07
Negativa	24.09	22.79	16.87	17.49

Tabla 4.2: *Comparativa de los porcentajes promedio de publicaciones para rango del día mañana y tarde.*

Los sujetos de la clase positiva escriben más en el rango **morning**. Entre porcentajes encontramos una diferencia ligeramente mayor a **mayor a 1.4 puntos porcentuales**. El que Los individuos de la clase positiva escriban más por la mañana que los de la clase negativa podría significar que los individuos de la clase positiva tardan más en integrarse a las actividades cotidianas que los de la clase negativa. Es decir, que podría ser que los sujetos de la clase negativa hagan otras cosas en lugar de hacer publicaciones como: desayunar en familia o hacer ejercicio. Podría indicar que los sujetos de la clase positiva se mantienen más "ensimismados" por la mañana, rehusándose a interactuar con el mundo externo y esto entonces convertirse en una mayor cantidad de publicaciones que los sujetos de la clase positiva.

Luego, los sujetos pertenecientes a la clase positiva escriben menos en el rango **after-**

noon comparados con la clase negativa. Se puede ver una diferencia entre porcentajes mayor a **1.7 puntos porcentuales**. Posiblemente la tarde sea el momento "más social del día"donde una persona etiquetada como çon depresión"no se puede aislar tan fácilmente para entregarse a alguna clase de ciclo de lamentaciones. Quizá al no poder aislarse, no les es posible hacer la misma cantidad de publicaciones que los sujetos de la clase negativa.

Examinando los porcentajes de las publicaciones emitidas por ***día del mes***, presentados en la imagen de la Figura 4.2 se puede notar lo siguiente:

Se perciben diferencias leves dentro de los porcentajes: los sujetos de la clase negativa realizan más publicaciones a mediados de mes (día 13 al 17), mientras que los individuos de la clase positiva generan más publicaciones a finales- comienzo de mes (día 26 al 29 y del 02 al 05). Dichas diferencias son menores a un punto porcentual y pudieran deberse a que para los sujetos de la clase positiva les sea más complejo el cambio de un mes a otro, hacer cierres de mes y de las actividades que correspondan al mes que termina.

Observando los porcentajes promedio de las publicaciones generadas por ***mes del año***, contenidos en la Figura 4.3. Es posible hacer énfasis aquellas contenidas en la tabla 4.3, correspondientes a los meses de junio, julio y diciembre.

Los individuos de la clase positiva escriben menos en los meses de **junio** y **julio** comparados con la clase negativa. Contando ambas particiones una diferencia mayor a **3.5 puntos porcentuales** y **9.5 puntos porcentuales** respectivamente. Dichas diferencias quizá puedan apuntar a que la generación de publicaciones de los individuos de las diferentes clases estén cargadas de una cierta temporalidad. Es decir, que en estos meses (junio y julio) es cuando se llevan a cabo las vacaciones de verano y están de forma estereotipada vinculados al sol y a la playa: de tal manera que los sujetos pertenecientes a la clase negativa tuvieran más ánimos para escribir publicaciones.

En el caso del mes de **diciembre** ocurre lo contrario: los sujetos de la clase positiva escriben más en comparación con la clase negativa. Presentándose una diferencia **mayor a 3 puntos porcentuales**. Nuevamente podría existir una influencia temporal vinculada con las celebraciones de fin de año. Pudiendo provocar en los individuos de la clase positiva una mayor nostalgia y por ende la expresión de la misma a través de una mayor cantidad de publicaciones.

	Junio		Julio		Diciembre	
Clase	% Part. Entr.	% Part. Prueba	% Part. Entr.	% Part. Prueba	%Part. Entr.	%Part. Prueba
Positiva	13.73	12.40	11.26	6.45	7.22	8.60
Negativa	17.51	16.02	20.77	17.98	4.20	5.18

Tabla 4.3: *Comparativa de los porcentajes promedio de publicaciones para meses junio, julio y diciembre.*

Revisando los porcentajes promedio de las publicaciones hechas por *estación del año*, presentados en la Figura 4.4, son más evidentes las diferencias correspondientes a las estaciones de verano e invierno presentadas en la tabla 4.4.

Durante la estación del año **verano**, los sujetos de la clase positiva escriben **menos**

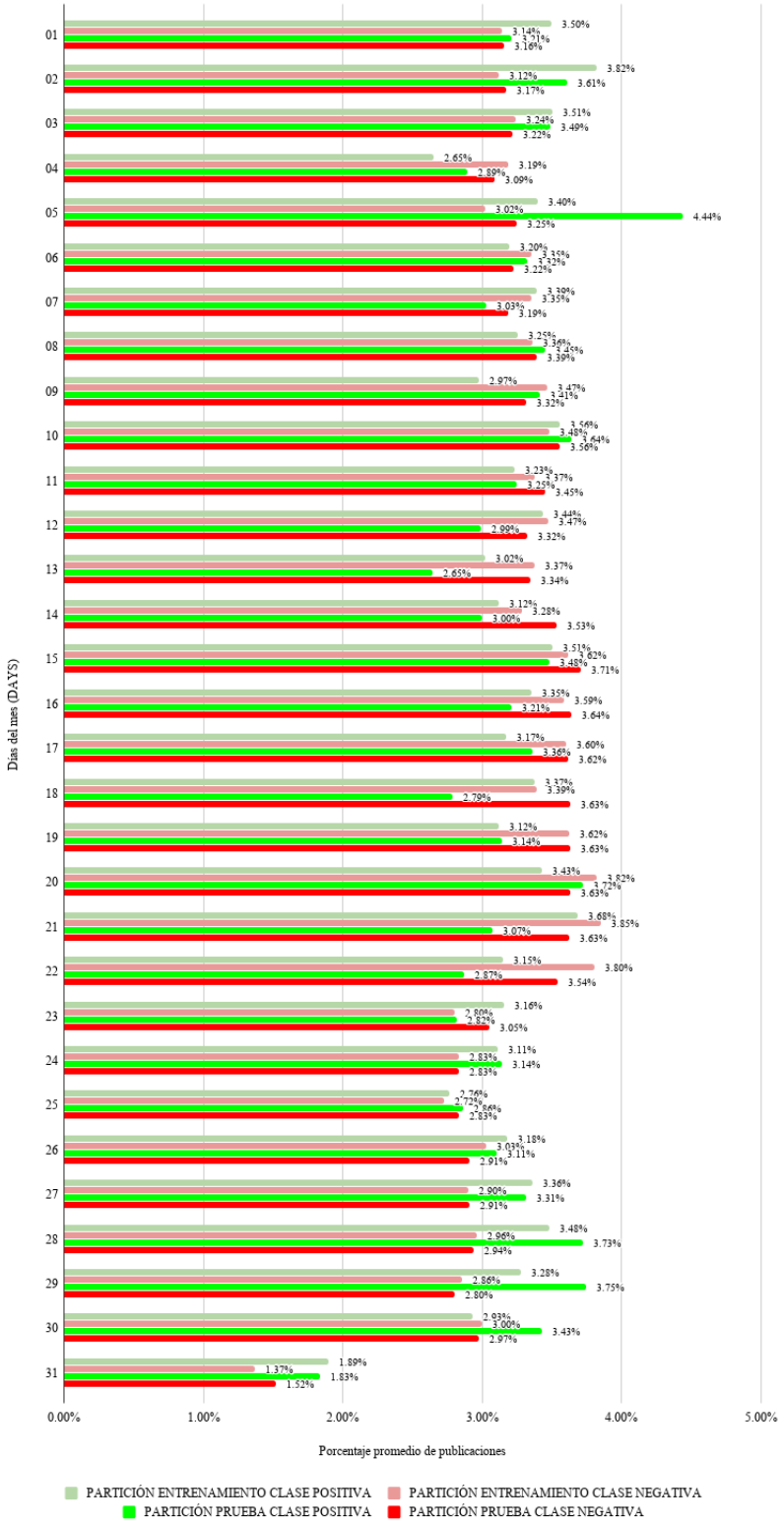

Figura 4.2: *Porcentajes promedio de publicaciones por día del mes.*

que aquellos pertenecientes a la clase negativa, contando con una diferencia **mayor a 14 puntos porcentuales**. Mientras que para la estación del año **invierno** se invierte el

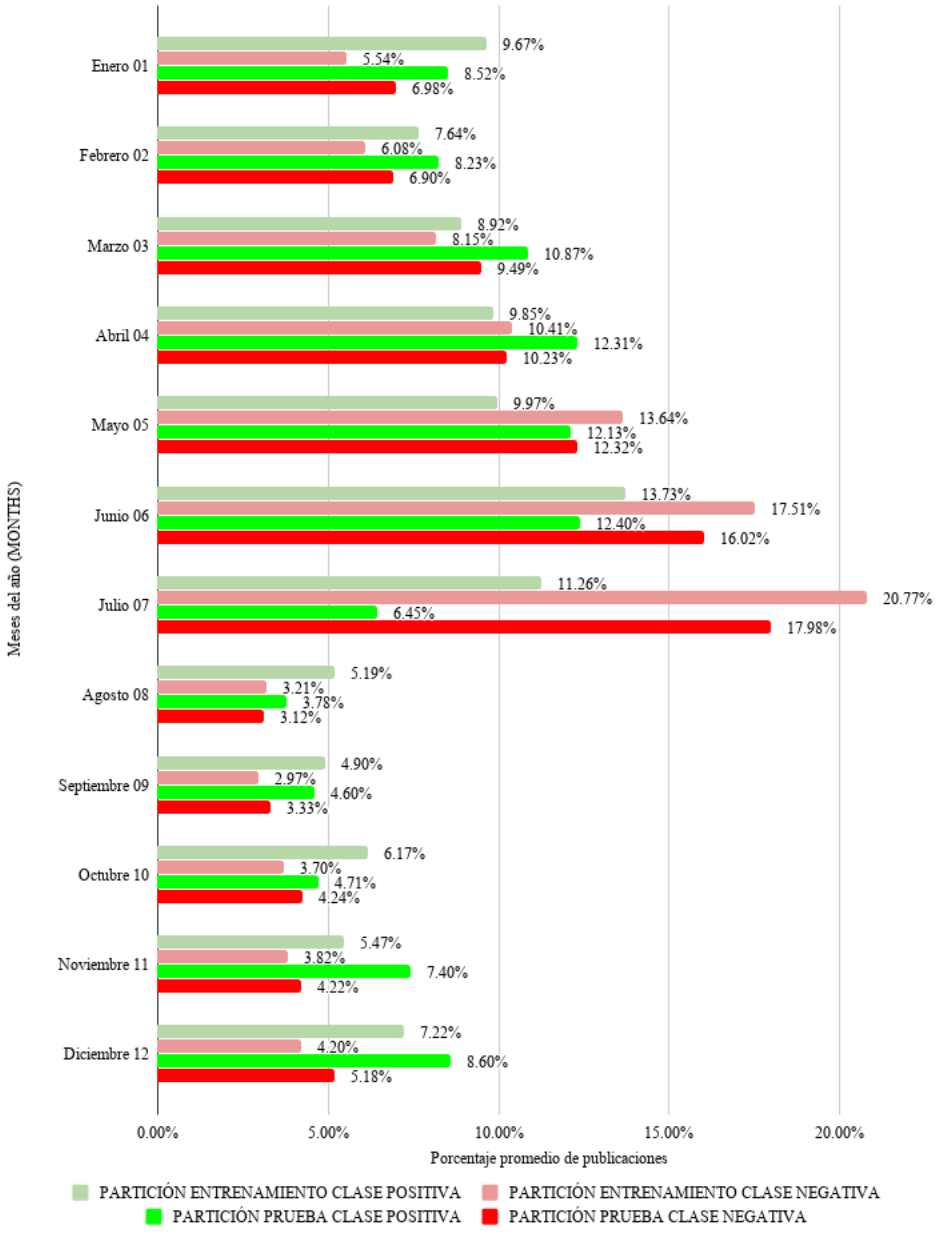

Figura 4.3: *Porcentajes promedio de publicaciones por mes del año.*

comportamiento, contando los individuos de la clase positiva con 6 % más publicaciones.

Dentro de este *segmento de tiempo* se encuentran las diferencias más notorias y mayores. Lo que podría indicar que la temporalidad presente en los datos no se limita a unos meses aislados, sino que se mantiene durante las estaciones completas. Al final con la posibilidad de que, si se integra una representación de los textos en función de las estaciones del año en las que los sujetos realizan sus publicaciones,exista un claro indicador para diferenciar la clase positiva de la negativa.

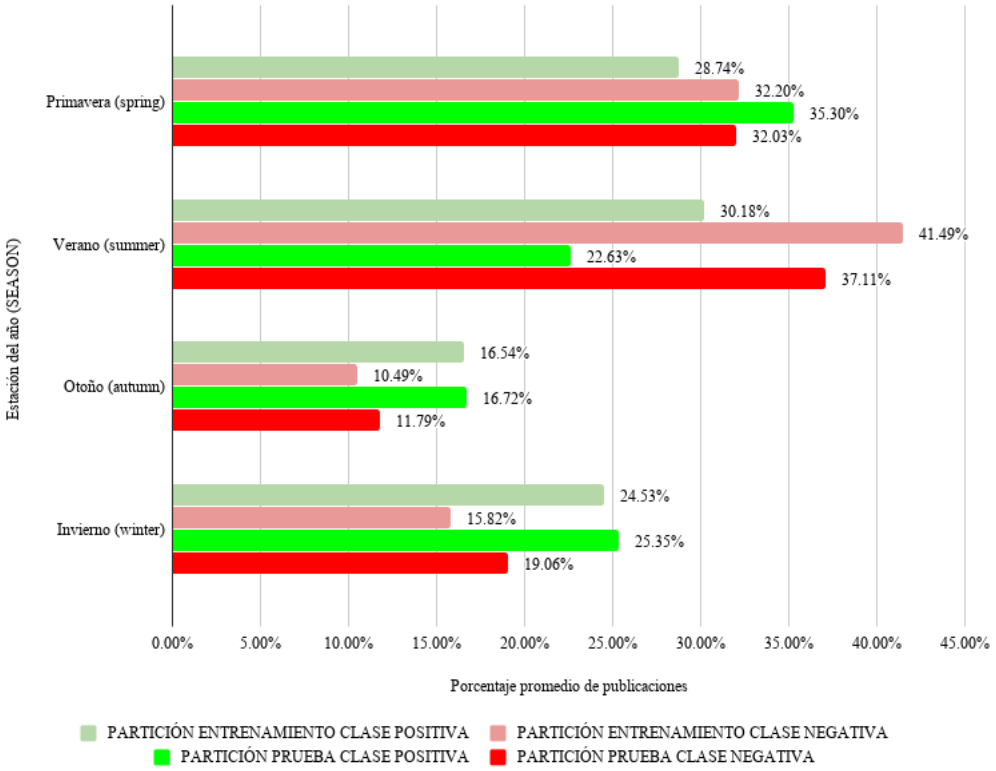

Figura 4.4: *Porcentajes promedio de publicaciones por estación del año.*

	Verano		Invierno	
Clase	% Partición Entrenamiento	% Partición Prueba	% Partición Entrenamiento	% Partición Prueba
Positiva	20.18	22.63	24.53	25.35
Negativa	41.49	37.11	15.82	19.06

Tabla 4.4: *Comparativa de los porcentajes promedio de publicaciones para estaciones del año verano e invierno.*

4.1.2. Estadísticas Semánticas

Las estadísticas denominadas semánticas, se calculan con base en las categorías de palabras definidas por el recurso J. W. Pennebaker, Francis y Booth, 2001. Las categorías de palabras están agrupadas en: dimensiones lingüísticas estandarizadas (pronombres, artículos, ...), aspectos psicológicos (afecto, cognición, ...), dimensiones relacionadas a "relatividad"(tiempo, espacio, ...) y cuestiones personales (trabajo, hogar, ...). Se puede encontrar la lista completa en J. W. Pennebaker y col., 2019.

Para cada partición del corpus (Train y Test), se calcularon los porcentajes de uso por cada familia de LIWC. Es decir, del total de las palabras usadas en las publicaciones de un sujeto; se calculó qué cantidad pertenece a cada una de las categorías de palabras.

Examinando los porcentajes de uso de las *68 familias de LIWC* en la gráfica de la Figura 4.5, se hacen manifiestas diferencias en las familias mostradas en la tabla 4.5.

Los individuos de la clase positiva emplean más palabras de la familias **Total pronom-**

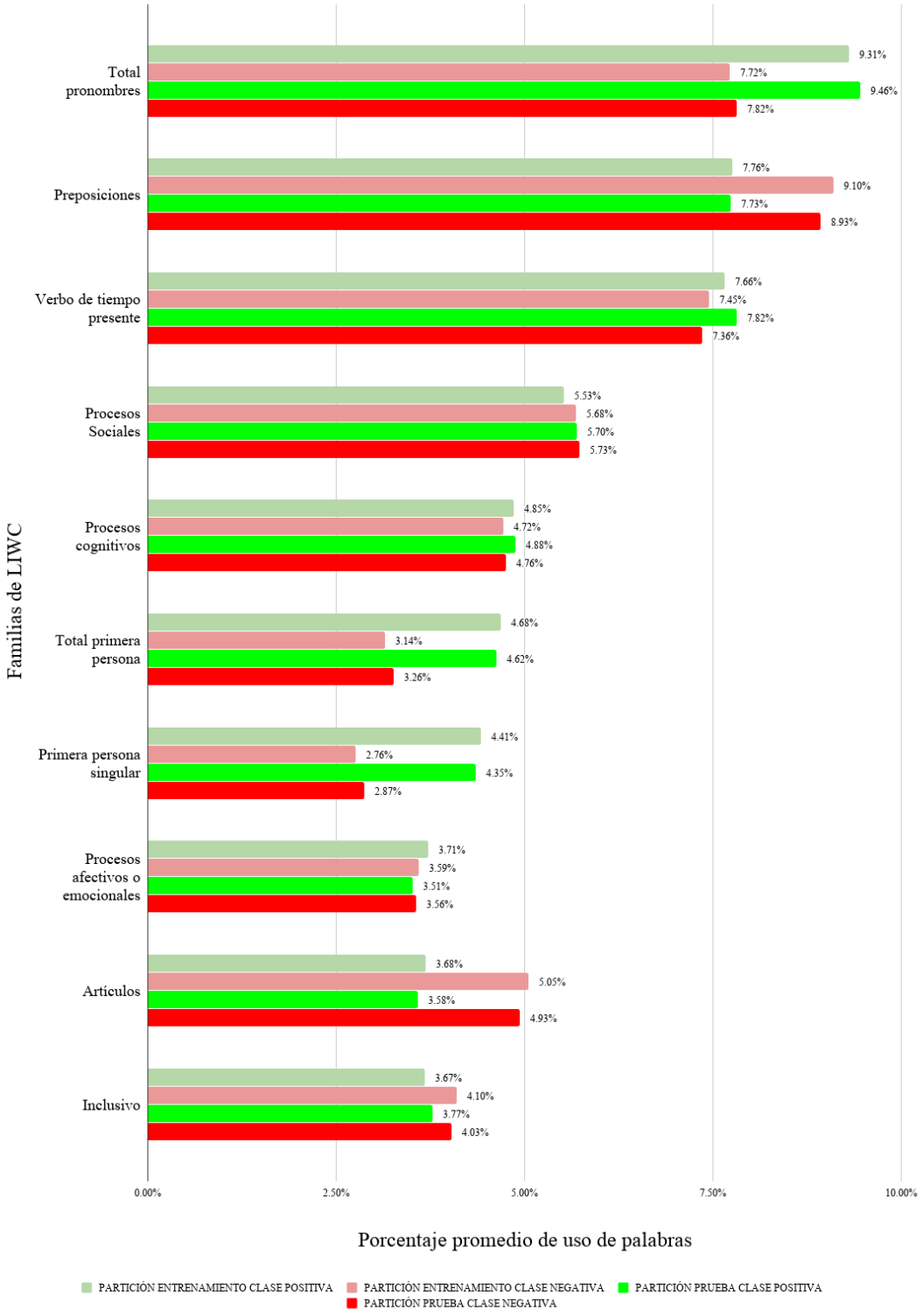

Figura 4.5: *Porcentajes de uso por familia de LIWC*

bres, **Primera persona singular** y **Total primera persona** que los de la clase negativa. Mostrándose diferencias de 1.59, 1.48 y 1.36 puntos porcentuales respectivamente.

En el caso de las familias Total pronombres, Primera persona singular y Total primera

	Total pronombres	
Clase	% Partición Entrenamiento	% Partición Prueba
Positiva	9.31	9.46
Negativa	7.72	7.82
	Primera persona singular	
Positiva	4.41	4.35
Negativa	2.76	2.87
	Total primera persona	
Positiva	4.68	4.62
Negativa	3.14	3.26
	Artículos	
Positiva	3.68	3.58
Negativa	5.05	4.93
	Preposiciones	
Positiva	7.76	7.73
Negativa	9.10	8.93

Tabla 4.5: *Comparativa de los porcentajes promedio de uso de palabras para las familias de LIWC: Total pronombres, Primera persona singular, Total primera persona, Artículos y Preposiciones.*

persona; su mayor uso por parte de los sujetos de la clase positiva podría significar que éstos se encuentran muy atentos a su propia persona, siendo el centro de todas las escenas e historias que alimentan sus publicaciones autocentradas.

Los sujetos de la clase positiva emplean menos palabras de la familia **Artículos** y **Preposiciones** versus los de la clase negativa. Contando con diferencias de 1.35 y 1.2 puntos porcentuales respectivamente.

De tal forma que el mayor uso de palabras contenidas en las familias Artículos y Preposiciones por parte de los sujetos de la clase negativa podría indicar que su forma de escribir es mejor o que su redacción es más rica.

Parece que si se generan representaciones basadas en las familias de LIWC podría ser que los rasgos distintivos aportados por las mismas sean insuficientes para poder discernir entre la clase positiva y la negativa.

4.1.3. Estadísticas Sintácticas (Part of Speech)

Las estadísticas denominadas sintácticas, se calculan con base en las etiquetas POS (part-of-speech) de la herramienta Treetagger Schmid, 2019 desarrollada por Helmut Schmid (36 etiquetas de POS en total).

Para cada partición del corpus (Train y Test), se calcularon los porcentajes de uso por cada Etiqueta POS del Etiquetador Treetagger. Es decir, del total de las palabras usadas en las publicaciones de un sujeto; se calculó qué cantidad pertenece a cada una de las etiquetas POS. La referencia completa de etiquetas POS se puede encontrar en Santorini, 1990.

También se calcularon los porcentajes de uso para un grupo particular de etiquetas

POS, consideradas como relacionadas con síntomas depresivos: VBD (Verb, past tense), VBG (Verb, gerund or present participle), VBN (Verb, past participle), PP (Personal pronoun), PP$ (Possessive pronoun), RB (Adverb), RBR (Adverb, comparative) y RBS (Adverb, superlative). Para esta parte se contemplan entonces, dos grandes grupos: el de la lista antes referida y el resto de las etiquetas POS.

Observando los porcentajes promedio de uso por categoría gramatical de las *36 etiquetas POS* mostrados en las gráfica de la Figuras 4.6, se hacen notorias las diferencias de las categorías contempladas en la tabla 4.6.

	Sustantivo singular NN	
Clase	% Partición Entrenamiento	% Partición Prueba
Positiva	13.50	13.50
Negativa	15.81	15.89
	Pronombre personal PP	
Positiva	10.61	10.60
Negativa	7.56	7.71
	Adverbio RB	
Positiva	8.58	8.45
Negativa	6.80	6.81
	Sustantivo propio NP	
Positiva	2.78	3.26
Negativa	6.40	6.16

Tabla 4.6: *Comparativa de los porcentajes promedio de uso de las categorías gramaticales Treetagger: Sustantivo singular NN, Pronombre personal PP, Adverbio RB y Sustantivo propio NP. Lista completa de categorias en Santorini, 1990*

Calse positiva Pronombres personales (2damás utilizada) Adverbios indicar que intentan describir más ser más descriptivos al contar sus experiencias al reflejarlas en publicaciones

Los sujetos de la clase positiva usan respectivamente un 2.8% y un 1.6% más **PP “Personal pronoun ” - Pronombres Personales** y los **RB “Adverb ” - Adverbios**. Respecto a la frecuencia de uso de estas dos categorías gramaticales podría reflejar que los sujetos de la clase positiva hacen un mayor intento por ser más descriptivos en relación con su experiencia cuando escriben sus publicaciones.

Los individuos de al escrila clase positiva emplean respectivamente un 2.3% y un 3% **menos** los **NN “Noun singular or mass” - Sustantivo singular** y los **NP “Proper noun, singular” - Sustantivos propios** que los de la clase positiva. Y podría sugerir que, al emplear más *Sustantivos propios*, los sujetos de la clase negativa pudieran estar más orientados al exterior, haciendo más referencias a lugares y personas fuera de ellos mismos. En el caso de mayor uso de *Sustantivos singular* por parte de los sujetos de la clase negativa, quizá de un indicio de que éstos se concentran más en qué está pasando que en cómo está pasando al momento de emitir sus publicaciones. Lo que también implicaría que los individuos de la clase positiva generan publicaciones principalmente centradas en su mundo interior y resaltando cómo es que las cosas ocurren.

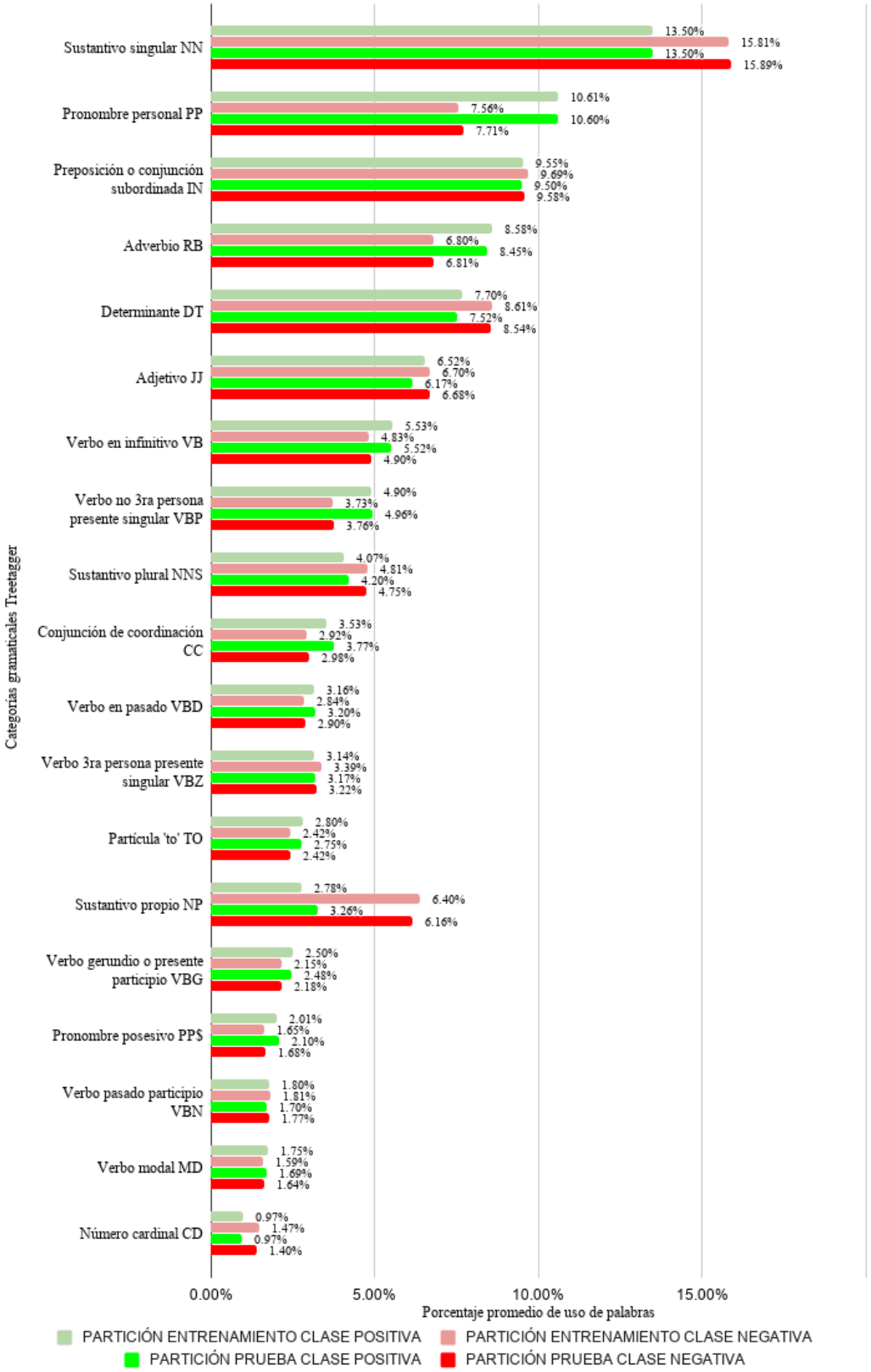

Figura 4.6: *Porcentajes promedio de uso por categoría gramatical de Treetagger.*

Analizando los porcentajes de uso por grupo de *etiquetas vinculadas con depresión*, en la imagen de la Figura 4.7 se puede notar lo siguiente:

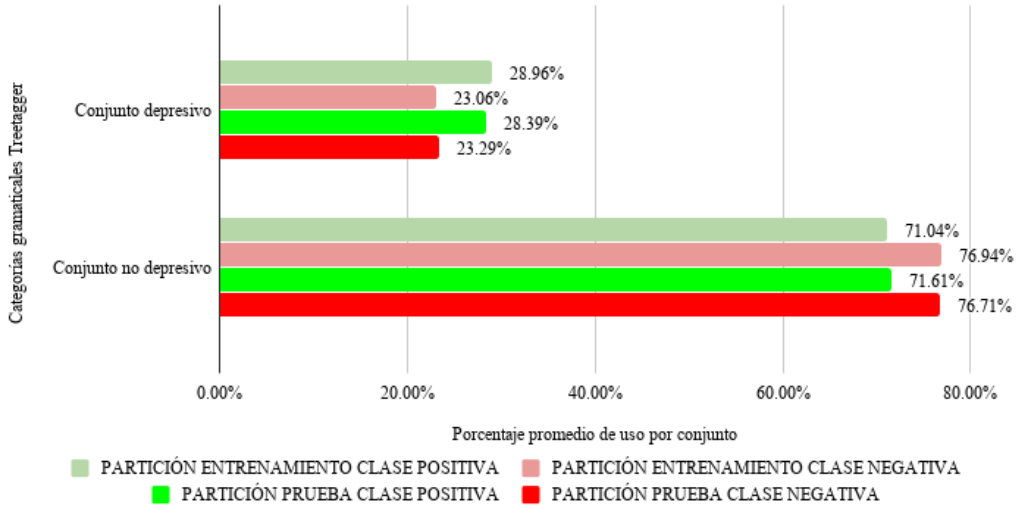

Figura 4.7: *Porcentajes promedio de uso por conjunto de categorías gramaticales de Treetagger.*

Los sujetos de la clase positiva emplean más **VBD, VBG, VBN, PP, PP$, RB, RBR y RBS** comparados con la clase negativa. Teniendo ambas particiones una diferencia apenas mayor a 5 puntos porcentuales. Cuando se hace el cálculo del porcentaje de uso para las etiquetas agrupadas como *relacionadas con depresión (VBD, VBG, VBN, PP, PP$, RB, RBR y RBS)* se detecta una diferencia un poquito mayor de 5 puntos porcentuales. Dato último que permite inferir que si se construye una representación con las etiquetas en grupo se podría generar una mejor diferenciación entre la clase positiva y negativa.

4.2. Representaciones propuestas

4.2.1. Representaciones base

Para el caso del baseline se han seguido las siguientes representaciones: BoW (con pesos binario, tf y tf-idf).

La bolsa de palabras (BoW) se considera que, dado una colección de documentos D con un vocabulario V y t_j términos. La representación en bolsa de palabras para un documento D_i es:

$D_i = \langle t_j : w_j \rangle \mid t_j \in D_i$

Donde:

w_j es el peso del término

t_j en el documento D_i

Para el peso binario si $V_j \in D_i$ entonces $w_j = 1$, de otra forma $w_j = 0$.

En el peso tf $w_t, d = f_t, d$. El peso w es la frecuencia de aparición del t_j término en el d_j documento.

Con respecto al peso tf-idf w_t,

$d = tf \cdot idf$,

$tf = f_t, d$ y

$idf = \frac{N}{|\{d \in D : t \in d\}|}$

4.2.2. Representaciones con atributos de comportamiento

En estas representaciones basadas en las estadísticas de comportamiento, los vectores de características $x \in X$ se construyeron con base en datos específicos de la fecha en la que se realizó cada una de las publicaciones de los sujetos pertenecientes al Corpus de este trabajo.

Los vectores tienen diferentes dimensiones en dependencia del o de los segmentos de tiempo que los conformen.

- Los horarios del día (RangeDay): $x_{RD} = \langle f_{morning}, f_{afternoon}, f_{night}, f_{dawn} \rangle$. Cada característica corresponde a los rangos del día especificados en la sección 4.1.1.
- Los días del mes (Day): $x_{DY} = \langle f_1, ..., f_n \rangle$, $n = \{1, ..., 31\}$ y cada característica corresponde al número del día del mes.
- Los meses del año (Month): $x_{MT} = \langle f_{jan}, f_{feb}, ..., f_{nov}, f_{dec} \rangle$, teniendo una característica por mes del año.
- Las estaciones del año (Season): $x_{SS} = \langle x_{spring}, x_{summer}, x_{autumn}, x_{winter} \rangle$, contando con una característica por cada estación del año.
- Cada característica se calcula en un proceso de dos partes:

  A. Se obtienen las frecuencias absolutas de cada característica.

  B. Se normaliza el cálculo de las frecuencias dividiéndolo entre el total de las publicaciones por sujeto de tal forma que al sumar todas las características de un vector, éste sume 1.

Las representaciones generadas con base en las fechas de las publicaciones y que en el presente trabajo se han denominado como *de comportamiento*, se exponen como de interés debido a que representan las conductas de los individuos que generaron los textos analizados. Considerando que una persona con depresión puede demostrar su estado anímico a través de su actividad de posteo. Entonces, suponiendo que en épocas de fiestas como las decembrinas sea probable que se desencadene un episodio depresivo y que éste sea reflejado en la cantidad de publicaciones realizadas.

4.2.3. Representaciones con atributos semánticos

Para estas representaciones generadas a partir de las estadísticas semánticas, los vectores de características $x \in X$ se generaron a partir de cada uno de los grupos de familias semánticas del recurso LIWC J. W. Pennebaker y col., 2019.

Así, los atributos que constituyen en este caso los vectores de características son los porcentajes de uso de cada una de las familias de LIWC.

$x_{LIWC} = \langle f_{1Pronoun}, f_{2I}, ..., f_m \rangle$, $m = \{1, ..., 68\}$ y cada característica corresponde a una familia del recurso LIWC.

Cada característica se calcula en dos pasos:

A. Obtener las frecuencias absolutas de cada característica.

B. Normalizar el cálculo de las frecuencias dividiéndolo entre el total de las publicaciones por sujeto de tal forma que al sumar todas las características de un vector, éste sume 1.

Las representaciones derivadas de los porcentajes de uso de las palabras de cada una de las familias de LIWC aquí referidas como *semánticas*, se presuponen importantes para discriminar si los textos analizados pertenecen a la clase positiva o negativa a través de determinar la proporción de palabras presentes en ellos que corresponden a cada grupo de palabras de LIWC. Asumiendo de antemano que las personas que transitan por un episodio de depresión se refieren con mayor frecuencia a sí mismas (familia 'yo'), que expresan con mayor frecuencia en sus publicaciones emociones negativas y de tristeza (familias - 'Emociones Negativas' y 'Depresión') o que escriben más sobre temas fatalistas como la muerte (familia 'Muerte'). Lo anterior con el suficiente contraste en relación a los textos publicados por los usuarios que no están atravesando un episodio depresivo para permitir la clasificación.

4.2.4. Representaciones con atributos sintácticos

En las representaciones constituidas de las estadísticas sintácticas, los vectores de características $x \in X$ se generaron a partir de cada una de las 36 etiquetas presentes en el etiquetador POS Schmid, 2019.

De esta forma, los atributos que conforman los vectores de características son los porcentajes de uso de cada una de las etiquetas POS del etiquetador Treetagger.

$x_{unigramPOS} = \langle BoW_{POS-tf} \rangle$

$x_{freqPOS} = \langle f_1, f_2, ..., f_l \rangle$, $l = \{1, ..., 36\}$, y cada característica corresponde a una categoría gramatical del etiquetador Treetagger (Santorini, 1990).

$x_{freqPOS-Dep}$, corresponde a una sola característica que agrupa las siguientes etiquetas POS: *VBD, VBG, VBN, PP, PP$, RB, RBR y RBS*. Consideradas como relacionadas con depresión al reflejar referencias hacia sí mismo y al pasado.

$x_{freqPOS-NonDep}$, corresponden al resto de etiquetas POS del Treetagger.

Las características dentro de $x_{freqPOS}$, $x_{freqPOS-Dep}$ y $x_{freqPOS-NonDep}$ se calcula en dos pasos:

A. Obtener las frecuencias absolutas de cada característica.

B. Normalizar el cálculo de las frecuencias dividiéndolo entre el total de las publicaciones por sujeto de tal forma que al sumar todas las características de un vector, éste sume 1.

$x_{unigramPOS-total} = \langle x_{unigramPOS}, x_{freqPOS} \rangle$, aquí el vector $x_{unigramPOS-total}$ se forma concatenando los dos vectores señalados.

$x_{unigramPOS-group} = \langle x_{unigramPOS}, x_{freqPOS-Dep}, x_{freqPOS-NonDep} \rangle$, en esta representación se concatenan los tres vectores indicados.

Las representaciones formadas a partir de los porcentajes de uso de las palabras clasificadas como una etiqueta POS específica referidas aquí como *sintácticas*, pueden facilitar la identificación de la clase positiva a través de ahondar en el uso de palabras pertenecientes a etiquetas POS específicas que ya se han planteado en la sección 1 donde de forma reiterada se afirma que las personas con depresión escriben con más referencias hacia sí mismas, emplean una mayor cantidad de verbos en pasado y en futuro y una mayor cantidad de pronombres.

V

Experimentos

5.1. Experimentos

En esta sección se explicarán los diferentes experimentos generados en este trabajo. De manera general se presentan tres grandes experimentos: *con representación de comportamiento*, *con representación semántica* y *con representación sintáctica*.

5.2. Experimentos base

Con el objetivo de establecer una línea de comparación para el resto de los experimentos, se hizo un experimento con representación de bolsa de palabras (*BoW*) y tres pesos distintos (binario, rf y tf-idf) definido previamente en la sección 4.2.1 para clasificar los textos del corpus en las clases positiva o negativa. Se utiliza la métrica *F1-Score* de la clase positiva para comparar el desempeño entre los algoritmos de aprendizaje empleados (Naive Bayes (NB), Máquinas de Soporte de Vectores (SVM) y Árbol de Decisión (DT)). Se llevaron a cabo dos versiones de este experimento base en relación con el preprocesamiento realizado:

- En la primera versión se han eliminado números, vínculos a páginas en web y símbolos de puntuación.
- En la segunda versión se han remplazado por etiquetas: los números (<number>), vínculos (<link>) y símbolos de puntuación (<punct>)

		BoW(peso binario)	BoW(peso tf)	BoW(peso tf-idf)
Experimento Base 1	NB	0.00	**0.55**	0.00
	SVM	0.54	0.51	0.26
	DT	0.37	0.41	0.40
Experimento Base 2	NB	0.00	**0.54**	0.00
	SVM	0.53	0.52	0.00
	DT	0.35	0.42	0.38

Tabla 5.1: *Resultados F1-Score (clase positiva) de los experimentos base 1 y 2.*

En la tabla 5.1 se muestran los resultados de los experimentos base, reportados con la métrica F1 Score de la clase positiva (con depresión). Es posible observar que para ambas versiones del experimento base, los mejores resultados se han obtenido con el algoritmo de clasificación Naive Bayes (NB) y con una representación tf. Lo anterior podría indicar que las frecuencias con las que aparecen los términos en cada instancia de la matriz término-documento.

En la tabla 5.2 se presentan los mejores resultados, al clasificar la clase positiva, de las matrices de confusión correspondientes a los experimentos base. Se puede ver claramente cómo coinciden los mejores valores de clasificaciones en la clase positiva 42 y 41 con los más altos valores de F1-Score de la clase positiva.

		NotDepressed	Depressed
Experimento Base 1 (tf - NB)	NotDepressed	289	60
	Depressed	10	**42**
		NotDepressed	Depressed
Experimento Base 2 (tf - NB)	NotDepressed	291	58
	Depressed	11	**41**

Tabla 5.2: *Matrices de confusión con los mejores resultados de los experimentos base.*

5.3. Experimentos con representaciones basadas en atributos de comportamiento

- **Hipótesis - 1**: Las representaciones construidas con la proporción de publicaciones por estación del año, diferencia mejor los textos dentro de las clases positiva y negativa.

Configuración experimental.- En este experimento, se formaron múltiples matrices término-documento a partir de los 4 segmentos de tiempo definidos en la sección 4.2.2. De modo tal, que quedaron integradas 15 diferentes representaciones con informaciones de comportamiento. Los algoritmos de aprendizaje empleados son: Naive Bayes (NB), Máquinas de Soporte de Vectores (SVM) y Árbol de Decisión (DT).

En la tabla 5.3, se presentan los resultados del las diferentes variantes del experimento con representaciones basadas en atributos de comportamiento. Las primeras dos columnas describen las 15 representaciones generadas para estos experimentos, mientras que la tercera corresponde a los resultados obtenidos por el algoritmo de Árboles de decisión.

		DT
4 segmentos de tiempo	$\langle x_{RD}, x_{DY}, x_{MT}, x_{SS} \rangle$	0.30
3 segmentos de tiempo	$\langle x_{RD}, x_{DY}, x_{MT} \rangle$	0.30
	$\langle x_{RD}, x_{DY}, x_{SS} \rangle$	0.30
	$\langle x_{RD}, x_{MT}, x_{SS} \rangle$	0.26
	$\langle x_{DY}, x_{MT}, x_{SS} \rangle$	**0.33**
2 segmentos de tiempo	$\langle x_{RD}, x_{DY} \rangle$	0.22
	$\langle x_{RD}, x_{MT} \rangle$	0.29
	$\langle x_{RD}, x_{SS} \rangle$	0.21
	$\langle x_{DY}, x_{MT} \rangle$	0.31
	$\langle x_{DY}, x_{SS} \rangle$	0.26
	$\langle x_{MT}, x_{SS} \rangle$	**0.38**
1 segmento de tiempo	$\langle x_{RD} \rangle$	0.18
	$\langle x_{DY} \rangle$	0.18
	$\langle x_{MT} \rangle$	**0.35**
	$\langle x_{SS} \rangle$	0.17

Tabla 5.3: *Resultados de F1-Score (clase positiva), correspondientes a las variantes del experimento con representaciones basadas en atributos de comportamiento.*

4 segmentos de tiempo	$\langle x_{RD}, x_{DY}, x_{MT}, x_{SS} \rangle$	NotDepressed	NotDepressed 311	Depressed 38
		Depressed	36	**16**
3 segmentos de tiempo	$\langle x_{RD}, x_{DY}, x_{MT} \rangle$	NotDepressed	NotDepressed 295	Depressed 54
		Depressed	33	**19**
2 segmentos de tiempo	$\langle x_{MT}, x_{SS} \rangle$	NotDepressed	NotDepressed 293	Depressed 56
		Depressed	27	**25**
1 segmento de tiempo	$\langle x_{MT} \rangle$	NotDepressed	NotDepressed 294	Depressed 55
		Depressed	29	**23**

Tabla 5.4: *Matrices de confusión con los mejores resultados para cada etapa del experimento con representaciones basadas en atributos de comportamiento.*

Si observamos los resultados obtenidos, es evidente que los mejores resultados se encuentran en aquellas representaciones que incluyen la información tanto del mes como de la estación del año $\langle x_{MT}, x_{SS} \rangle$. Tenemos en el bloque de representaciones con 3 segmentos de tiempo el mejor resultado con la representación $\langle x_{DY}, x_{MT}, x_{SS} \rangle$ y un F1-Score para la clase positiva de 0.33. En el bloque de 2 segmentos de tiempo, el mejor resultado lo obtiene la representación $\langle x_{MT}, x_{SS} \rangle$ con un F1-Score para la clase positiva de **0.38**. Siendo ese resultado es el más alto de todas las variantes realizadas. Lo anterior contrastado en el bloque de 1 segmento de tiempo, donde el mejor resultado se refleja con la representación $\langle x_{MT} \rangle$ con un F1-Score para la clase positiva de 0.35 y que está muy próximo al mejor resultado.

En la tabla 5.4 tenemos los mejores resultados de las matrices de confusión correspondientes a los experimentos realizados con representaciones basadas en atributos de comportamiento, se muestra la mejor matriz de cada bloque de segmentos de tiempo. Las representaciones que mejor clasifican la clase positiva, se corresponden con los mejores resultados de F1-Score para la clase positiva: $\langle x_{MT} \rangle$ y $\langle x_{MT}, x_{SS} \rangle$ habiendo clasificado correctamente 23 y 25 instancias respectivamente.

No obstante, estos resultados antes descritos, siguen muy por debajo de los indicados en la tabla 5.2 sobre experimentos base.

5.4. Experimentos con representaciones basadas en atributos semánticos

- **Hipótesis - 2**: Las representaciones construidas con los promedios de uso de las familias de LIWC, permiten una discriminación mejor entre clases que las representaciones base.

- **Configuración experimental**.- Se generó una matriz término-documento a partir de los porcentajes de uso de palabras por familia de LIWC 4.2.3. Los algoritmos de aprendizaje empleados (Naive Bayes, Máquinas de Soporte de Vectores y Árbol de Decisión).

Para este experimento se obtuvo únicamente resultado de F1-Score (clase positiva) para el algoritmo de aprendizaje de Árbol de decisión (DT), con un valor de **0.29**. Resultado que está por debajo del mejor obtenido en los experimentos base.

En la tabla 5.5 se presenta la matriz de confusión para el experimento con atributos de origen semántico y se observa que la cantidad de instancias asignadas con etiqueta positiva sigue por debajo de lo indicado en los experimentos base.

		0 NotDepressed	1 Depressed
%uso familias LIWC	0 NotDepressed	293	56
	1 Depressed	34	**18**

Tabla 5.5: *Matriz de confusión con mejores resultados para el experimento con atributos de origen semántico.*

5.5. Experimentos con representaciones basadas en atributos sintácticos

- **Hipótesis - 3**: Las representaciones construidas con los promedios de uso de las etiquetas POS del etiquetador gramatical Treetagger, permiten una discriminación mejor entre clases que las representaciones base.

Configuración experimental.- Se generó una matriz término-documento a partir de los porcentajes de uso de palabras por etiqueta POS de Treetagger 4.2.4, una matriz término-documento a partir de los porcentajes de uso de palabras por conjunto de etiquetas POS vinculadas a la depresión como se explica en la sección 4.1.3 y otra matriz término-documento con base en los textos representados como unigramas de etiquetas POS . Los algoritmos de aprendizaje empleados (Naive Bayes (NB), Máquinas de Soporte de Vectores (SVM) y Árbol de Decisión (DT)).

Este experimento se conforma de tres variantes:

- Definición de la matriz término-documento mediante los textos representados como etiquetas POS $x_{unigramPOS} = \langle BoW_{POS-tf} \rangle$.
- Creación de la matriz término-documento con base en la concatenación de la representación del punto anterior, con los porcentajes de uso de palabras por etiqueta POS: $x_{unigramPOS-total} = \langle x_{unigramPOS}, x_{freqPOS} \rangle$.
- Generación de la matriz término-documento al unir la representación de la primera versión $x_{unigramPOS}$ con los porcentajes de uso de palabras por grupo de etiquetas (2 grupos) vinculadas y no vinculadas con estados depresivos $x_{unigramPOS-group} = \langle x_{unigramPOS}, x_{freqPOS-Dep}, x_{freqPOS-NonDep} \rangle$.

	NB	SVM	DT
$x_{unigramPOS}$	0.39	0.40	0.25
$x_{unigramPOS-total}$	0.39	**0.51**	0.34
$x_{unigramPOS-group}$	0.39	0.49	0.35

Tabla 5.6: *Resultados F1-Score (clase positiva), de las variantes del experimento con atributos de origen sintáctico.*

		NotDepressed	Depressed
NB - $x_{unigramPOS}$	NotDepressed	213	136
	Depressed	7	**45**
		NotDepressed	Depressed
NB - $x_{unigramPOS-total}$	NotDepressed	213	136
	Depressed	7	**45**
		NotDepressed	Depressed
NB - $x_{unigramPOS-group}$	NotDepressed	213	136
	Depressed	7	**45**

Tabla 5.7: *Matrices de confusión con los mejores resultados para cada etapa del experimento con atributos de origen sintáctico.*

Es posible observar en la tabla 5.6, los mejores resultados se obtienen con la segunda versión del experimento $x_{unigramPOS-total}$ y el algoritmo de aprendizaje SVM. Se puede ver que para cada representación propuesta en este experimento, dio mejores resultados el algoritmo de clasificación de Máquina de Soporte de Vectores. Además de que el mejor resultado de F1-Score para la clase positiva se acerca más al valor reportado en los experimentos base.

Por otra parte, en la tabla 5.7 se presentan las matrices de confusión que poseen un valor mayor de instancias clasificadas como positivas. Y es posible notar que, para las tres variantes del experimento, los mejores resultados se obtienen del algoritmo de clasificación Bayes Ingenuo. Notanto que, aunque clasifica bien la clase positiva, *confunde* la positiva y debido a eso es que en la métrica F1-Score de la clase positiva no posee el mayor valor.

5.6. Análisis de Resultados

En el problema que atañe al presente proyecto terminal, el de clasificación de textos de usuarios con depresión, lo que realmente importa es no dejar fuera a sujetos que estén atravesando un episodio de depresión al asignarle la etiqueta ***no deprimido*** ya que hacer esto se consideraría como un error muy grave. Sin embargo, si se le asigna a algún sujeto la etiqueta ***deprimido*** no es considerado como un error tan grave considerando las consecuencias que esta asignación pueda traer: ya que como máximo en esta situación el individuo acudiría con un especialista para determinar la certeza de la asignación, algo que siempre es muy positivo al considerarse como un cuidado a la salud mental.

Por lo tanto, para este caso, se considera al *error tipo II* o *falso negativo* como el más importante.

Entonces se puede concluir respecto del experimento de comportamiento, que aunque en apariencia todas las configuraciones contempladas obtuvieron el mismo resultado, la configuración que mejor clasifica los textos es la que junta la información del mes y la estación del año en que se realiza la publicación y empleando el algoritmo de aprendizaje de Árboles de decisión. Aceptando la **Hipótesis - 1** con base en lo observado en las matrices de confusión.

Para el experimento con representación semántica se puede concluir, que aunque en apariencia todos los algoritmos de aprendizaje obtuvieron el mismo resultado, la configuración que mejor clasifica los textos es la que junta la información del porcentaje de uso por familia LIWC y empleando el algoritmo de aprendizaje de Árboles de decisión. Comparando con los resultados de los experimentos base se queda muy por debajo por lo que la **Hipótesis - 2** es rechazada.

Y en el experimento con representación sintáctica se puede concluir que tiene un mejor desempeño al representar los textos como etiquetas POS y agregarle la información correspondiente al porcentaje de uso de palabras por cada etiqueta POS. Sin embargo, el mejor resultado 0.72 % se queda por debajo de los valores de los experimentos base aunque clasifica ligeramente mejor la clase positiva. De acuerdo a lo planteado, se rechaza la **Hipótesis - 3**.

VI

Conclusiones

6.1. Conclusiones

El problema que se abordó en el presente trabajo forma parte de la tarea de clasificación de textos, donde se pretendía identificar si un determinado conjunto atributos para las representaciones de textos de un usuario, facilitaban el reconocimiento de las clases **deprimido** o **no deprimido**. Un problema binario en el que la selección de los atributos para armar las representaciones de los textos puede llevar a la adecuada identificación de la clase **deprimido**.Esto se vuelve crucial porque implicaría un posible desequilibrio en la salud mental de un individuo.

Dado el objetivo general del presente trabajo, lo que se quiso explorar fue qué clase de recursos con información psicolingüística hacían más fácil el proceso de clasificación de textos.

Lo anterior mediante la selección de diferentes atributos para generar las representaciones de los textos.

Desde el principio del trabajo existieron ciertas líneas de aproximación identificadas:

- La posibilidad de que las personas con depresión escribieran más en ciertos momentos del día; particularmente por las noches y madrugadas o que sufrieran depresión durante los meses de invierno al cambiar el clima e impedir de alguna manera la convivencia y aumentando la sensación de soledad.

- La exploración de un recurso en específico llamado LIWC, una herramienta que se ha encontrado particularmente útil y recurrente en múltiples trabajos. Este recurso, con las familias que tiene definidas, permite la identificación de características psicológicas y emocionales en las palabras que las personas usan al comunicarse.

- De la mano del punto anterior, el profundizar sobre el tipo de palabras que se emplean al generar los textos se hizo importante. Esto se consolidó a través del uso del etiquetador de categorías gramaticales Treetagger. Aquí teniendo la corazonada de que el uso de cierto tipo de palabras refleja estados mentales particulares.

Para llevar a cabo el estudio estadístico del corpus se hicieron diversos cálculos siguiendo las líneas antes mencionadas. Antes que otra cosa se agruparon por lapso de tiempo las diversas publicaciones de los sujetos del corpus, se determinaron primero los porcentajes promedio de las publicaciones por cada usuario y segmento de tiempo. Después se sacaron las gráficas correspondientes para poder analizar los resultados. Se hizo lo mismo con las otras dos líneas de aproximación al calcular los porcentajes promedio de uso de los diferentes aspectos y sacando las gráficas correspondientes.

Al final, el análisis antes descrito permitió que se pudieran pre-seleccionar cuáles de estos cálculos podrían lograr mejores resultados al incorporarlos dentro de un modelo de aprendizaje.

Para replicar el uso de las representaciones de los textos, se generaron experimentos base con la representación tradicional de bolsa de palabras, con diferentes pesos (binario, tf y tf-idf) que han sido ampliamente usadas en los trabajos consultados.

De acuerdo con los resultados del análisis del corpus, se diseñaron múltiples representaciones. Para las características de comportamiento se probaron cada una de las posibles combinaciones de los segmentos de tiempo como atributos. En el caso del recurso LIWC se generó una sola versión con un atributo para cada una de las 68 familias del mismo. Y para el etiquetador POS Treetagger, se obtivieron tres vertientes: una como bolsa de palabras de etiquetas POS, otra que juntaba la bolsa de palabras POS con los porcentajes promedio de uso por categoría gramatical y la última juntaba la bolsa de palabras de etiquetas POS con dos atributos más que consideraban a las categorías gramaticales agrupadas como depresivas o no depresivas respectivamente.

Considero que en este punto no fue posible hacer replicas exactas de los trabajos consultados porque en las fuentes no se contaba con el detalle suficiente de cómo integraron los modelos y además porque la aproximación que hicieron los autores de los trabajos consultados estaban enmarcados en la competencia de detección temprana, aspecto que se excluyó del presente trabajo al emplear la totalidad de publicaciones de los sujetos para hacer la clasificación de los textos.

Para determinar las ventajas y desventajas de la incorporación de los atributos seleccionados a los modelos de aprendizaje se decidió emplear tres algoritmos de clasificación ampliamente conocidos: Multinomial Naive Bayes, Support Vector Machine y Decision Tree.

Entonces se hicieron tres ejecuciones por cada una de las representaciones diseñadas para los distintos tipos de características. De aquí se pudieron observar ciertas diferencias entre los resultados tanto evaluando la métrica F1-Score de la clase positiva como observando a detalle las cantidades de sujetos bien clasificados como **deprimido**. Dichas diferencias permitieron saber las ventajas o desventajas de las mismas. Por ejemplo aunque se obtuvieron resultados de F1-Score (clase positiva) de 0.38 con las representaciones de comportamiento, la cantidad de sujetos bien identificados se quedó en 25. Mientras que con las representaciones basadas en características sintácticas se obtiene un F1-Score (clase positiva) de 0.39 pero la cantidad de individuos bien clasificados como deprimido sube a 45.

De acuerdo con el análisis estadístico del corpus y junto con los resultados obtenidos de los experimentos, se puede apreciar cuál ha sido el mejor resultado obtenido. Para al final poder concluir que los atributos que mejor clasifican a los sujetos como deprimidos son aquellos que tienen su origen en las categorías gramaticales (sintáctico).

Y se puede atisbar que para la clasificación de textos de usuarios con depresión, es muy importante el cómo lo dicen más que el qué es lo que dicen. Es decir, que más allá de que alguien use palabras relacionadas con la tristeza o la soledad o la muerte, se hace mucho más importante poner atención a la cantidad de veces que se refiere a sí mismo o si escribe mensajes muy descriptivos porque ahí es donde se puede tener una mejor clasificación para este problema.

También parece ser que no tiene mucha importancia el momento del día en el que se escriben las publicaciones, por supuesto esto visto aisladamente.

El principal aprendizaje que se ha podido obtener al emprender este trabajo es que dependiendo del problema que se esté atendiendo, los resultados más altos no siempre

son los mejores

En este caso y dadas las características del presente problema, más allá de seleccionar una métrica para definir cuál sería el mejor resultado, lo que se convirtió en el aspecto más importante fue la cantidad de individuos con depresión que eran correctamente clasificados.

6.2. Trabajo Futuro

En los resultados de las diferentes versiones del experimento con atributos de comportamiento, aquellas representaciones que incluyeron los meses y las estaciones del año mejoraron la cantidad de individuos bien clasificados como con depresión, podría ser que incorporar este tipo de información a otras representaciones compuestas de diversos tipos de características ayude a mejorar los resultados.

En el caso del trabajo con el recurso LIWC, me parece que se pueden quitar ciertas familias de la representación. Las familias que considero que se deberían quitar son aquellas que son más bien sintácticas y dejar únicamente aquellas familias que representen temas de procesos psicológicos o emocionales. Así se podría saber, de ese subgrupo, cuáles son más adecuadas o contribuyen más a la obtención de un mejor resultado.

Para el caso de las representaciones con características de origen sintáctico, que fueron las que obtuvieron mejores resultados, podría juntarse con los mejores atributos de comportamiento para saber si mejoran los resultados con esa representación.

Capitulo 7: Bibliografía

Bibliografía

AIMx. (yearmonthday). 15 Estudio sobre los Hábitos de los Usuarios de Internet en México 2019 versión pública. Asociación de Internet de México. **retrieved from** https://www.asociaciondeinternet.mx/es/component/remository/Habitos-de-Internet/15-Estudio-sobre-los-Habitos-de-los-Usuarios-de-Internet-en-Mexico-2019-version-publica/lang,es-es/?Itemid=

Almeida, H., Briand, A. y Meurs, M.-J. (yearmonthday). Detecting early risk of depression from social media user-generated content. *Working Notes of CLEF*.

Arumugam, R. y Shanmugamani, R. (yearmonthday). *Hands-On Natural Language Processing with Python [Hands-On Procesamiento de Lenguaje Natural con Pyrthon]*. Reino Unido: Packt.

Bucci, W. y Freedman, N. (yearmonthday). The language of depression. *Bulletin of the Menninger Clinic*, *45*(4), 334.

De Choudhury, M., Gamon, M., Counts, S. y Horvitz, E. (yearmonthday). Predicting depression via social media. En *Seventh international AAAI conference on weblogs and social media*.

Farias-Anzaldua, A. A., Montes-y-Gómez, M., López-Monroy, A. P. y González-Gurrola, L. C. (yearmonthday). UACH-INAOE participation at eRisk2017. En *CLEF*.

Jackson, P. y Moulinier, I. (yearmonthday). *Natural Language Processing for Online Applications. Text retrieval, extraction and categorization [Procesamiento de Lenguaje Natural para Aplicaciones en línea.Recuperación de Textos, extracción y categorización]*. Philadelphia, USA: John Benjamins Publishing Company.

Kroenke, K., Spitzer, R. L. y Williams, J. B. (yearmonthday). The PHQ-9: validity of a brief depression severity measure. *Journal of general internal medicine*, *16*(9), 606-613.

Kurdi, M. Z. (yearmonthday). *Natural language processing and computational linguistics: speech, morphology and syntax [Procesamiento de Lenguaje Natural y Lingüística Computacional.Lenguaje, Morfología y Sintaxis]*. John Wiley & Sons.

Losada, D. E., Crestani, F. y Parapar, J. (yearmonthday). CLEF 2017 eRisk Overview: Early Risk Prediction on the Internet: Experimental Foundations. En *CLEF (Working Notes)*.

Lozano, R., Gómez-Dantés, H., Pelcastre, B., Ruelas, M., Montañez, J., Campuzano, J., ... y González, J. (yearmonthday). Carga de la Enfermedad en México 1990-2010: Nuevos resultados y desafíos. *México DF: Instituto Nacional de Salud Pública*.

Malam, I. A., Arziki, M., Bellazrak, M. N., Benamara, F., Kaidi, A. E., Es-Saghir, B., ... y Ramiandrisoa, F. (yearmonthday). IRIT at e-Risk. En *CLEF*.

Miner, G., Elder IV, J., Fast, A., Hill, T., Nisbet, R. y Delen, D. (yearmonthday). *Practical text mining and statistical analysis for non-structured text data applications [Minería de textos práctica y análisis estadístico para aplicaciones de datos textuales no estructurados]*. Academic Press.

Mohammed, M., Khan, M. B. y Bashier, E. B. M. (yearmonthday). *Machine learning: algorithms and applications [Aprendizaje Automático. Algoritmos y aplicaciones]*. Crc Press.

Mohri, M., Rostamizadeh, A. y Talwalkar, A. (yearmonthday). *Foundations of machine learning [Fundamentos de Aprendizaje Automático. Computación Adaptativa y Aprendizaje Automático]*. MIT press.

Moreno, M. A., Jelenchick, L. A., Egan, K. G., Cox, E., Young, H., Gannon, K. E. y Becker, T. (yearmonthday). Feeling bad on Facebook: Depression disclosures by college students on a social networking site. *Depression and anxiety*, *28*(6), 447-455.

NIMH, N. (yearmonthday). Depresión. *Salud Mental*, 2-3.

OMS. (yearmonthday). Temas de Salud, Depresión. **retrieved from** https://www.who.int/topics/depression/es/

OMS-IESM. (yearmonthday). *Informe de la evaluación del sistema de salud mental en México utilizando el Instrumento de Evaluación para Sistemas de Salud Mental de la Organización Mundial de la Salud (IESM-OMS)*. Secretaría de Salud de México.

Pennebaker, J. (yearmonthday). *The Secret Life of Pronouns. What Our Words Say About Us [La Vida Secreta de los Pronombres. Lo que Nuestras Palabras dicen de Nosotros]*. New York, NY, USA: Bloomsbury Press.

Pennebaker, J. W., Francis, M. E. y Booth, R. J. (yearmonthday). Linguistic inquiry and word count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates*, *71*(2001), 2001.

Pennebaker, J. W., Francis, M. E. y Booth, R. J. (yearmonthday). Tabla 1: LIWC2001 Resultados de la Información de la Variables. **retrieved from** http://www.liwc.net/liwcespanol/descriptiontable1.php

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

Pérez, C. L., Castro, W. P. y Rodrıguez, M. G. (yearmonthday). Sıntomas de depresión en hombres. *Universitas Psychologica*, *16*(4).

Rude, S., Gortner, E.-M. y Pennebaker, J. (yearmonthday). Language use of depressed and depression-vulnerable college students. *Cognition & Emotion*, *18*(8), 1121-1133.

Sadeque, F., Xu, D. y Bethard, S. (yearmonthday). UArizona at the CLEF eRisk 2017 Pilot Task: Linear and Recurrent Models for Early Depression Detection. *CEUR workshop proceedings*, *1866*.

Santorini, B. (yearmonthday). *Part-of-speech tagging guidelines for the Penn Treebank Project*. University of Pennsylvania, School of Engineering y Applied Science ...

Sarkar, D. (yearmonthday). *Text analytics with Python: A practical real-world approach to gaining actionable insights from your data [Analíticas de textos con Python: Una aproximación práctica en el mundo real para obtener percepiones procesables de tus datos]*. Apress.

Schmid, H. (yearmonthday). TreeTagger - a part-of-speech tagger for many languages. **retrieved from** https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

Sotiropoulos, D. N. y Tsihrintzis, G. A. (yearmonthday). *Machine Learning Paradigms: Artificial Immune Systems and Their Applications in Software Personalization [Paradigmas de Aprendizaje Automático. Sistemas inmunes automáticos y sus aplicaciones en personalización de software]*. Springer.

SSA. (yearmonthday). Programa de Acción Especıfico Salud Mental 2013-2018. Obtenido de Programa de Acción Especıfico. **retrieved from** https://www.gob.mx/salud/documentos/programa-de-accion-especifico-salud-mental-2013-2018

Tausczik, Y. R. y Pennebaker, J. W. (yearmonthday). The psychological meaning of words: LIWC and computerized text analysis methods [El significado psicológico de las palabras: LIWC y métodos computarizados de análisis textual]. *Journal of language and social psychology*, *29*(1), 24-54.

Transparencia-Presupuestaria. (yearmonthday). Estándar Internacional de Datos Presupuestarios Abiertos. **retrieved from** https://www.transparenciapresupuestaria.gob.mx/es/PTP/datos_presupuestarios_abiertos

Trotzek, M., Koitka, S. y Friedrich, C. M. (yearmonthday). Linguistic Metadata Augmented Classifiers at the CLEF 2017 Task for Early Detection of Depression. En *CLEF*.

Villegas, M. P., Funez, D. G., Ucelay, M. J. G., Cagnina, L. C. y Errecalde, M. L. (yearmonthday). LIDIC - UNSL's Participation at eRisk 2017: Pilot Task on Early Detection of Depression. En *CLEF*.

Weintraub, W. (yearmonthday). *Verbal behavior: Adaptation and psychopathology*. Springer Publishing Company.