UNIVERSIDAD AUTÓNOMA DE CHIHAUHUA

FACULTAD DE INGENIERÍA

SECRETARÍA DE INVESTIGACIÓN Y POSGRADO

ESTRATEGIAS BASADAS EN APRENDIZAJE COMPUTACIONAL PARA DETECTAR DEPRESIÓN EN USUARIOS DE REDES SOCIALES

POR:

ING. ALÁN ALÉXIS FARÍAS ANZALDÚA

TESIS PRESENTADA COMO REQUISITO PARA OBTENER EL GRADO DE MAESTRO EN INGENIERÍA EN COMPUTACIÓN

CHIHUAHUA, CHIH., MÉXICO **MAYO DE 2018**

Estrategias Basadas en Aprendizaje Computacional para Detectar Depresión en Usuarios de Redes Sociales. Tesis presentada por Ing. Alán Alexis Farías Anzaldúa, como requisito parcial para obtener el grado de Maestro En Ingeniería En Computación, ha sido aprobada y aceptada por:

M.I. Javier González Cantú
Director de la Facultad de Ingeniería

Dr. Fernando Rafael Astorga Bustillos
Secretario de Investigación y Posgrado

M.S.I. Karina Rocío Requena Yáñez
Coordinador(a) Académico

Dr. Luis Carlos González Gurrola
Director(a) de Tesis

Mayo de 2018

Comité:
Dr. Luis Carlos González Gurrola
Dr. Manuel Montes y Gómez
Dra. Graciela María de Jesús Ramírez Alonso
M.I. Jesús Roberto López Santillán

UNIVERSIDAD AUTÓNOMA DE
CHIHUAHUA

22 de mayo de 2018

ING. ALÁN ALEXIS FARÍAS ANZALDÚA

Presente

En atención a su solicitud relativa al trabajo de tesis para obtener el grado de Maestro en Ingeniería en Computación, nos es grato transcribirle el tema aprobado por esta Dirección, propuesto y dirigido por el director **Dr. Luis Carlos González Gurrola** para que lo desarrolle como tesis, con el título: **"ESTRATEGIAS BASADAS EN APREDIZAJE COMPUTACIONAL PARA DETECTAR DEPRESIÓN EN USUARIOS DE REDES SOCIALES".**

ÍNDICE

FACULTAD DE INGENIERÍA
Circuito No.1, Campus Universitario 2
Chihuahua, Chih., México. C.P. 31125
Tel. (614) 442-95-00
www.fing.uach.mx

UNIVERSIDAD AUTÓNOMA DE
CHIHUAHUA

4. Experimentación y Resultados

I. Métricas de desempeño

II. Resultado del experimento base (bolsa de palabras)

III. Resultados del uso de ganancia de información en bolsa de palabras

IV. Resultados del uso de esquemas de pesado

V. Resultados de clasificación en experimentos temporales, textuales y mixtos

VI. Resultados de palabras más relevantes según el procesamiento de texto

VII. Resultados del concurso y evaluación ERDE

VIII. Hallazgos sobre el dataset

5. Conclusiones

I. Trabajo futuro

Solicitamos a Usted tomar nota de que el título del trabajo se imprima en lugar visible de los ejemplares de las tesis.

A T E N T A M E N T E
"Naturam subiecit aliis"

EL DIRECTOR

M.I. JAVIER GONZÁLEZ CANTÚ

FACULTAD DE INGENIERIA U.A.CH
CHIHUAHUA
DIRECCION

EL SECRETARIO DE INVESTIGACIÓN Y POSGRADO

DR. FERNANDO RAFAEL ASTORGA BUSTILLOS

FACULTAD DE INGENIERÍA
Circuito No.1, Campus Universitario 2
Chihuahua, Chih., México. C.P. 31125
Tel. (614) 442-95-00
www.fing.uach.mx

Dedicatoria

A mis padres, por todo su apoyo y amor. Sin ellos no hubiera sido posible completar este trabajo.

Al doctor Luis Carlos Gonzalez Gurrola, por creer en mi, tenerme paciencia, y por todos sus consejos.

Al doctor Manuel Montes y Gómez, por creer en mi, apoyarme y guiarme.

A mis amigos, por haber estado conmigo y escucharme cuando los necesitaba.

Agradecimientos

A CONACYT, por la beca 735623/599448 que hizo posible este estudio de posgrado.
Al proyecto “Caracterización de usuarios en redes sociales: hacia un enfoque multimodal y multidominio” CONACYT - Problemas Nacionales, Proyecto 247870, 2015-2017, por el apoyo para participar en el foro eRisk del CLEF 2017 y la estancia de tres meses en Puebla.

Resumen

El objetivo del presente trabajo es el de analizar publicaciones de redes sociales en búsqueda de características textuales y temporales que permitan discriminar usuarios con y sin depresión. La necesidad de esta tarea viene de la enorme cantidad de personas a nivel mundial que sufren de depresión, lo que les impide llevar una vida plena. Ayudando a comprender cómo difieren las publicaciones entre usuarios con y sin depresión se puede llegar a construir un sistema integrado en redes sociales que permita a sus usuarios recibir ayuda profesional.

Para lograr este objetivo se realizó una exploración experimental de características textuales y lingüísticas, buscando la mejor combinación de parámetros. Se obtuvo como resultado un método que combina ambos tipos de características, procesando a los usuarios a dos pasos. El primer paso trata las publicaciones individualmente y el segundo paso realiza un meta análisis de las publicaciones. Dicho método se utilizó como propuesta para participar en el foro eRisk (*EARLY RISK PREDICTION ON THE INTERNET*) del CLEF (*Conference and Labs of the Evaluation Forum*) 2017, en Dublin, Irlanda. Específicamente, se participó en la *Pilot Task: Early Detection of Depression* (lit. Tarea Piloto: Detección temprana de Depresión), el cual distribuyó *dataset* de [1] con el que se realizaron los experimentos y se desarrolló el método propuesto. Este *dataset* fue construido a partir de publicaciones de usuarios de la red social *reddit* y cuenta con etiquetas para distinguir a los usuarios con depresión de los usuarios sin depresión.

Este documento detalla la metodología propuesta para el CLEF, así como la búsqueda que llevó a crear esa propuesta. Se presenta, además, información aprendida durante el trayecto con la esperanza de que resulte de utilidad para quienes busquen encontrar respuestas del psique humano en el mar de información que es el internet.

Índice de Contenido

Índice de figuras

Índice de tablas

1 Introducción

1.1 Definición del problema

Según el *Center for Disease Control and Prevention (CDC)*, 1 de 10 adultos tienen alguna forma de depresión [2]. Las personas que sufren de depresión presentan sentimientos de desesperanza, pérdida de motivación, trastornos alimenticios y de sueño, y se sienten desconectados de otras personas. Si no se trata adecuadamente, la depresión puede llevar al suicidio. Los criterios presentados en el *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5)*, giran entorno a un análisis a largo plazo (dos años o más) realizado por un profesional. Esto requiere que los individuos sean diagnosticados por un psicólogo, algo que muchas personas no son capaces o no están dispuestas a hacer.

Las redes sociales en línea son cada vez más populares, al grado de ser casi ubicuas en países que tienen acceso a internet. Estas plataformas permiten compartir y expresar los pensamientos y emociones libremente con otras personas. Como es de suponerse, esta información es una fuente valiosa de información para tareas de procesamiento de lenguaje natural, tal como: minado de opiniones [3], análisis de sentimientos [4], o inferir problemas de salud mental [5].

En este trabajo se busca detectar la depresión en individuos mediante un análisis de las publicaciones que comparten en las redes sociales. Para ello se hace uso del *dataset* construido para el CLEF [1] que utiliza publicaciones de la red social *reddit*. Los usuarios fueron etiquetados automáticamente según si se encontraba que el usuario reportara haber sido diagnosticado formalmente por un profesional, en cuyo caso se marcaba al usuario como depresivo, o no depresivo en el caso contrario.

Este *dataset* fue construido para explorar la posibilidad de detectar depresión en usuarios de redes sociales mediante sus publicaciones. De ser posible detectar depresión de esta manera, se abriría la posibilidad de acercar a los usuarios con profesionales de la salud mental

en un futuro.

1.2 Estado del Arte

Las redes sociales se han convertido en un lugar atractivo para detectar problemas de salud en comunidades enteras o en individuos. Un estudio presentado en [6] compara comunidades en línea con depresión contra comunidades en línea de control, reportando que usando temas y marcadores de estilo como características es posible diferenciar entre los usuarios.

Dos de las redes sociales comúnmente usadas para minar este tipo de datos son *Facebook* y *Twitter*. Específicamente en *Facebook* se han realizado estudios para identificar distintos contextos de la depresión. En [7], el estudio se centra en predecir la depresión postpartum a partir de datos compartidos. Así mismo, en [8] se presenta un estudio que evalúa síntomas de la depresión a través de una aplicación desarrollada para *Facebook*.

En [9], los autores intentan responder a la pregunta: *¿Qué información de salud pública puede ser inferida de Twitter?* La depresión es uno de los trastornos mentales más comúnmente asociados a un cierto grupo de palabras. Uno de los trabajos que buscan caracterizar usuarios con depresión en *Twitter* es [10], en el que los autores proponen algunos atributos para medir el comportamiento depresivo, tales como interacción social, lenguaje o emoción. Los autores utilizaron un *Support Vector Machine* que logra una presición del 70 % para predecir la aparición de depresión en los usuarios. En [11], Minsu et al. realizaron un estudio cualitativo para investigar las diferencias de percepción que se presentan entre usuarios con y sin depresión de *Twitter*. Encontraron que los usuarios con depresión son más propensos a percibir *Twitter* como una herramienta de consciencia social e interacciones emocionales. Más recientemente, Nadeem et al. [12] compilaron más de 2.5M de *tweets* para identificar el trastorno mayor de depresión (MDD, por sus siglas en inglés “major depressive disorder”), reportando una precisión de más del 81 %. Los autores indican que su método pudiera ayudar a estimar el riesgo de los individuos de tener depresión.

Con un buen número de estudios buscando detectar la depresión, nuevos trabajos sugieren la posibilidad de detectar otro tipo de fenómenos sociales, como la planeación del suicidio [13].

Hasta donde se investigó, no han habido trabajos previos que utilicen la red social *reddit*. El *dataset* de Losada y Crestani [1] fue creado como un primer intento de llenar este vacío. Este *dataset* fue presentado para la *Early Detection of Depression Pilot Task* (eRisk) del CLEF 2017, para investigar la posibilidad de detectar depresión en usuarios de *reddit* mediante sus publicaciones. Este trabajo se basó en dicho *dataset* y se presentó para el CLEF el mejor resultado obtenido.

El eRisk utilizó las siguientes métricas para evaluar los resultados: valor-F, precisión, sensibilidad y ERDE. Las primeras tres métricas tienen un uso bastante difundido en la comunidad y se explican en el capítulo 4. La última métricas, ERDE, fue propuesta por Losada y Crestani junto al mismo *dataset* [1]. En el capítulo 3 se habla más sobre el *dataset*, y en el capítulo 4 se explica con más detalle la métrica ERDE. Pero en resumen, este *dataset* de *reddit* separa la información de cada usuario en bloques de tiempo ordenados cronológicamente, y las medidas ERDE propuestas pretenden evaluar el desempeño de clasificación de los usuarios, penalizando el resultado mientras más tiempo se requiera para clasificarlos.

Los otros participantes del eRisk del CLEF 2017 enfocaron sus esfuerzos principalmente en la caracterización de los textos, más que en los clasificadores utilizados. Algunos de estos trabajos utilizan diccionarios de términos psiquiátricos y emocionales [14, 15], y solo un trabajo optó por construir grafos para la representación de los usuarios [16]. La mayoría optó por construir las representaciones de los documentos mediante un análisis del lenguaje en los documentos del *dataset* [17, 18, 19, 20], y es a este grupo al que pertenece el trabajo presente.

En [14], el equipo de la University of Quebec in Montreal (UQAM) utilizaron dos tipos de diccionarios: de sentimientos y de medicamentos. Construyeron el diccionario de sentimientos a partir del SenticNet [21] y listas de palabras de nombres de antidepresivos, enfermedades afines a depresión y drogas psicoactivas, todas extraídas de *Wikipedia*. Representaron a los usuarios con una bolsa de palabras construida de n-gramas (explicado en la sección 2), aunada a otra bolsa de palabras de los diccionarios, bolsa de palabras de partes de la oración (conteo de verbos, adjetivos, sustantivos, etc. utilizados en los textos) y conteo de número de publicaciones realizadas por el usuario. Para el concurso eRisk presentaron 5 resultados distintos, la mayoría cambiando el algoritmo de clasificación de los usuarios. Sin embargo, uno de los resultados se basaba en la construcción de un motor de búsqueda para comparar

las similitudes entre los textos de los usuarios de prueba con los usuarios de entrenamiento, y clasificarlos en base a esta similitud. Este tipo de enfoque, llamado *Information Retrieval*, fue el único encontrado en este concurso.

El otro trabajo que utilizó diccionarios fue de la University of Arizona [15]. Ellos construyeron un solo diccionario a partir de las palabras más asociadas a la raíz "depress" en el foro de salud mental de *Yahoo! Answers*, seleccionando las primeras 110 según su valor TF-IDF (explicado en la sección 2) de artículos de *Wikipedia*. También utilizaron Metamap, la cual es una herramienta que permite extraer conceptos del diccionario de sinónimos *Unified Medical Language System (UMLS)*. Encontraron experimentalmente que la mejor configuración de la herramienta resultaba en 404 términos médicos. Con estos diccionarios construyeron una bolsa de palabras para representar a los usuarios.

Como se mencionó anteriormente, únicamente el trabajo de la Universidad Autónoma de México [16] utilizó grafos para este concurso. La razón por la que eligieron grafos en lugar de bolsa de palabras u otra representación fue poder conservar el orden de las palabras, además de únicamente la presencia de las palabras. Otra razón por la que usaron un modelo de grafos es que conserva la correlación entre las palabras de manera natural. Para poder clasificar los textos construyeron un grafo prototipo para cada clase con los usuarios de entrenamiento y grafos para cada usuario de prueba, y utilizaron métricas de similitud para aseverar a cuál clase se parece cada usuario.

Finalmente, los trabajos que extrajeron características lingüísticas de las publicaciones para representar a los usuarios resultan tener enfoques bastante diversos. En [19], el *FHDO Biomedical Computer Science Group* construyó bolsas de palabras con unigramas, bigramas y trigramas, y seleccionaron términos por ganancia de información (explicado en la sección 2). Sobre este proceso construyeron características manualmente, como la presencia de ciertos términos y expresiones (por ejemplo, 'my_depression'), conteos de elementos gramaticales (por ejemplo, pronombres posesivos), juntando 16 características. El trabajo de [17] hace un análisis lingüístico similar, pero hay diferencias en las características textuales específicas que terminan usando. Otra diferencia es la inclusión de características estadísticas, como la cantidad de palabras promedio por publicación y la cantidad de publicaciones por usuario. Los últimos dos trabajos también tienen similitudes entre sí. En [18] se utiliza *Concise*

Semantic Analysis (CSA), la cual es una representación tipo bolsa de palabras que disminuye la dimensionalidad utilizando como características grupos semánticos de palabras, en lugar de palabras individuales. Expandiendo sobre esta idea, en [20] utilizan una variación de CSA que genera y utiliza grupos de vocabulario agrupados por tiempo para representar a los usuarios en términos de cómo cambia su vocabulario por el tiempo.

En conclusión, se observó que los otros participantes del eRisk difieren de este trabajo en que todos buscan directamente crear una representación para los usuarios. En el afán de lograr esa meta mezclan la información de todas las publicaciones, únicamente extrayendo estadísticas sobre las publicaciones individuales en algunos casos. En cambio, este trabajo parte de un análisis lingüístico similar pero termina construyendo una representación de usuario en varios pasos que primero clasifica individualmente las publicaciones y utiliza esta información para analizar al usuario.

1.3 Hipótesis

Es posible encontrar patrones en las publicaciones de usuarios de redes sociales que sean capaces de diferenciar a usuarios con depresión de usuarios sin depresión.

1.4 Objetivo general

Diseñar un método que permita discriminar usuarios de redes sociales con y sin depresión, utilizando características textuales y temporales extraídas de sus publicaciones.

1.5 Objetivos específicos

- Estudiar el estado de arte sobre el tema de detección de depresión.
- Analizar el corpus de publicaciones de *reddit* para identificar características útiles para los vectores de representación.
- Explorar técnicas de procesamiento de lenguaje natural para vectorización.
- Explorar características temporales para mejorar el resultado de clasificación.

- Explorar métodos que combinen características textuales y temporales.
- Obtener y analizar resultados.

1.6 Justificación

Resulta fascinante observar que sea posible identificar en un mensaje escrito el estado anímico de la persona que lo escribió sin que lo haya mencionado explícitamente. La sutileza del uso del idioma permite que exista esta información subtextual, pero identificar estas sutilezas resulta complejo. ¿Qué hay en un texto que revele que esa persona está triste, feliz o molesta? ¿La gramática? ¿Las palabras usadas? ¿Las ideas expresadas? Más aún, esta observación puede llevar a darse cuenta de que además del texto, todo lo que una persona puede crear está cargado de información de quien lo creó, como si fuera firmada por la persona. ¿Pero cómo se sabe esto? ¿En qué yace esta información? Poder hacer esta apreciación resulta de un proceso intuitivo que parece ser como caja negra para el observador, por lo que sigue siendo un misterio. Este trabajo pretende analizar el uso de las palabras, buscando encontrar qué pudiera delatar la depresión de un usuario en los textos que escribe, mediante el uso de un sistema impersonal y objetivo como es una computadora.

La importancia de la detección de la depresión en particular va más allá de la comprensión del uso del lenguaje y de las interacciones sociales en línea. La personas presas de la depresión tienen dificultades en su vida diaria, de manera laboral, interpersonal y emocional, y esto les impide llevar una vida plena. Además, en el peor de los casos la depresión es que puede llevar al suicidio. Lograr detectar la depresión de forma automatizada en las redes sociales permite brindar ayuda más rápidamente a las personas, evitando así estos problemas.

Entender el habla de las personas es solo uno de los pasos que llevan al misterio de la consciencia. El proceso de toma de decisiones, el por qué se tienen distintas opiniones sobre un mismo tema, el funcionamiento de las emociones mismas, etc., es basto el universo de preguntas entorno a la consciencia. El uso de las computadoras puede ayudar a resolver estos enigmas, a entendernos a nosotros mismos. Este trabajo es solo uno de muchos pasos que debemos tomar de forma colaborativa para llegar a resolver todas estas preguntas. Quizá algún día se llegue a comprender qué hace que «Yo» se pregunte «¿Qué soy yo?».

1.7 Aportación metodológica

La exploración de características lingüísticas y temporales y el análisis de su resultado sobre la clasificación de los usuarios llevó a la construcción de un método de clasificación de usuarios a dos pasos, mediante características mixtas. No se se había reportado en la literatura, ni en el concurso eRisk, una metodología similar.

A grandes rasgos, la clasificación de usuarios a dos pasos consiste de un primer paso en el que se procesan las publicaciones de los usuarios individualmente, y en un segundo paso se toma la información extraída de las publicaciones individuales para realizar un metaanálisis con el cual caracterizar y representar a los usuarios.

1.8 Aportación práctica

Fruto de este trabajo de investigación resulta un programa capaz de diferenciar, con cierta, precisión, usuarios con depresión de usuarios sin depresión mediante sus publicaciones en redes sociales. Así mismo, sirve de antecedente para el estudio de la detección de depresión y el estudio en el área de procesamiento de lenguaje natural para la Universidad Autónoma de Chihuahua.

2 Marco Teórico

2.1 Procesamiento de texto

Al ser esta una tarea de procesamiento de texto para clasificación, lo más importante es el tratamiento de los datos en bruto: los documentos. ¿Se debe de conservar todo el texto intacto? Si se ha de modificar algo, ¿qué se debe de modificar? ¿Y para qué? Además, ¿qué es más importante? ¿La presencia misma de las palabras en un texto o la relación entre las palabras?

Este es el tipo de preguntas que surgen en el área de procesamiento de lenguaje natural. Aunque los lingüistas tienen fuertes opiniones al respecto, la experimentación sugiere que abordar el problema de procesamiento de texto desde el enfoque lingüístico tradicional puede entorpecer los resultados. Inclusive, al investigador y líder del grupo de investigación del habla en IBM, Frederick Jelinek, se le atribuye la cita *"cada vez que despido a un lingüista mejora el desempeño del reconocedor [de habla]"* [22]. Se considera que esto se debe a que la lingüística intenta comprender el fenómeno del habla humano desde el enfoque de los humanos, lo que contrasta con el procesamiento de lenguaje natural (NLP). El NLP, siendo parte del área de las ciencias de la computación, debe de comprender el fenómeno del habla humano desde el enfoque de las computadoras.

En esta sección se hablará sobre varias técnicas de procesamiento de texto y los resultados que se obtienen de ello.

2.1.1 Filtrado y sustitución de elementos

Esta sección cubrirá los usos, ventajas y desventajas de filtrar o sustituir elementos en el texto. No hay que olvidar que para los humanos es fácil reconocer cadenas distintas como si fueran la misma. Por ejemplo es posible saber que "Hola" y "hola" son la misma palabra, solo con mayúscula inicial o sin ella, y se puede saber que "don't" y "dont" son lo mismo,

solo que una escrita correctamente y la otra no. Sin embargo, la computadora los ve como cadenas distintas. Es entonces que se tiene que decidir si forzar a que la computadora los vea iguales (por ejemplo, convirtiendo todo a minúsculas o haciendo la ortografía consistente), o si se conservan las diferencias.

Los elementos que se van a cubrir en esta sección son: mayúsculas, símbolos y emoticones, números, URL y corrección ortográfica.

Mayúsculas

Las mayúsculas forman parte del alfabeto latino y de todos los lenguajes que hacen uso de él. Aunque su uso típico está dictado por las reglas de gramática y ortografía de cada idioma, hay otros usos informales frecuentemente encontrados, especialmente en las comunidades en línea. Estos usos incluyen: hacer énfasis en una palabra o en toda una oración, o dar la impresión de que se alza el volumen de la voz. Inclusive se llegan a encontrar textos escritos exclusivamente en mayúsculas (sobre todo mensajes breves), y suelen ser percibidos con inmadurez por parte del autor [23].

Debe de considerarse si detectar o ignorar el uso de mayúsculas en los textos, ya que puede ser una característica útil en el dominio específico que se esté tratando [24].

Símbolos y emoticones

Los símbolos lingüísticos forman parte de las normas gramaticales de los idiomas, pero también pueden considerarse características estilísticas que permitan perfilar a usuarios en distintos grupos [24].

Otro uso para los símbolos lingüísticos comúnmente encontrado en línea son los emoticones. Los emoticones son secuencias de caracteres que se usan para representar una emoción o reacción, por su similitud con un gesto facial o corporal. Ejemplos de emoticones categorizados por el tipo del sentimiento que representan se pueden encontrar en la tabla 2.1. Cabe señalar que el significado exacto de los emoticones suele ser tema de debate, además de que el significado puede cambiar según si se usa con ironía o con seriedad.

Dado que los emoticones suelen ser usados para representar emociones, vale la pena considerar conservarlos en estudios donde se necesite capturar el sentimiento en un texto o el

Tabla 2.1: Ejemplos de distintos tipos de emoticones y significado aproximado

Positivos	Significado	Negativos	Significado
:D	Sonrisa con boca abierta	D:<	Repudio o asco
:)	Sonrisa con boca cerrada	):	Tristeza
XD	Risa	:c	Tristeza
n_n	Sonrisa horizontal	TT_TT	Llanto
:-P	Gesto travieso, con lengua de fuera	-.-	Molestia o frustración
<3	Corazón	OTL	Decepción de uno mismo

estado anímico de una persona.

URL

Es una práctica común en línea compartir enlaces o URL. Esto es particularmente cierto en una red social o foro de discusión, como *reddit*, donde se llega a tener una conversación sobre un enlace en particular (por ejemplo, un artículo o video). Así pues, los enlaces suelen ser usados como referencias o como tema central de la conversación.

Para fines de procesamiento de lenguaje natural, las URL también pueden considerarse como características. Para poder utilizar las URL, en este trabajo se hizo una sustitución de la URL completa por el nombre de dominio (por ejemplo, la URL `https://www.youtube.com/watch?v=gHYBLM3OdL4` se deja como www-youtube-com), y se encontró la preferencia de ciertos sitios web para una clase en específico.

Corrección ortográfica

Como se mencionó al principio de esta sección, las computadoras no tienen la misma facilidad que los humanos para comprender que una palabra bien escrita es la misma que cuando se encuentra mal escrita. Sorprendentemente, las faltas de ortografía en un texto pueden resultar beneficiosas para el estudio de los usuarios. Es posible encontrar que distintos grupos de personas escriben con distinta cantidad de fallas de ortografía, así como fallas dis-

tintas [24]. Esto puede facilitar la asociación de un usuario con algún grupo en particular, sobre todo en el caso de clasificación por edad.

2.1.2 Tokenización

Tokenización es el proceso de separar un texto en entidades individuales (llamadas *tokens*). A nivel de programación, consiste en separar una cadena en una lista de subcadenas. Existen varias librerías y métodos para ello, además de ser fácil de implementar manualmente. La primera consideración que hay que tomar para tokenizar es establecer qué es considerado como un token. Lo más sencillo es separar por espacios, pero esto no maneja los símbolos lingüísticos adecuadamente. Por ejemplo, la cadena "Por ejemplo, la cadena" quedaría como "Por", "ejemplo,", "la", "cadena". En el segundo token, "ejemplo,", se puede ver que la coma se tomó como parte de la palabra, lo cual va a evitar que ese token sea reconocido como el mismo que "ejemplo".

Para evitar este problema, el trabajo presente separa los símbolos lingüísticos (como comas, paréntesis, comillas y otros) de las letras que conforman una palabra. De esta manera, estos símbolos pasan a formar sus propios tokens. Así pues, la cadena "¿Qué es un token?" queda como "¿", "Qué", "es", "un", "token", "?".

Concatenación de negativos

Concatenación de negativos es el nombre que recibe el proceso de crear n-gramas de términos que niega con los términos que están negando. Es decir, se considera "don't want" como un concepto distinto a solamente "want", y mediante este proceso se asegura de que se tomen como tokens distintos. En la tabla 2.2 se encuentran los términos negativos considerados para este proceso. La idea detrás de este proceso es la de garantizar una entrada individual para cada concepto y otra para su negación, inclusive cuando no se utilizan otros tipos de n-gramas (explicada en la sección siguiente).

N-gramas

La técnica de n-gramas es muy común en el área de procesamiento de lenguaje natural. Consiste en tomar la lista de tokens resultante del proceso de tokenización y generar una

Tabla 2.2: Términos negativos usados para la concatenación de negativos

no, not, never, bad, don't, doesn't, didn't
won't, can't, wouldn't, couldn't, isn't
ain't', 'aren't, dont, doesnt, didnt, wont
cant, wouldnt, couldnt, isnt, aint, arent

nueva lista agrupando tokens consecutivos. La n en n-gramas se refiere al número de palabras consecutivas que se agrupan. Se utiliza para capturar secuencias de palabras, algo que resulta particularmente útil en conjunto con la bolsa de palabras las cuales ignoran la secuencialidad de las palabras. Cabe señalar que mientras más grandes sean los n-gramas, más infrecuentes serán. Esto es porque mientras más grande se el número de palabras que conforman al n-grama, la probabilidad de que se repita la combinación específica de palabras en un documento disminuye. Se sugiere no usar n-gramas mayores a 3 palabras.

La literatura habla sobre la existencia de otro tipo de n-gramas, que en lugar de modelar secuencias de palabras completas modelan secuencias de caracteres. Algunos les llaman q-gramas [25], mientras que otros les llaman n-gramas de letras [24]. Indistintamente de su nomenclatura, este tipo de n-gramas permite capturar raíces de palabras y conjugaciones por separado, permitiendo capturar el uso de tiempos gramaticales, diminutivos, entre otras cosas. También puede capturar similitudes de palabras, permitiendo que los errores ortográficos no eviten asociar dos palabras, a la vez que capturan las diferencias entre ellas.

2.2 Representación vectorial de texto

Tan importante como procesar el texto es poder representarlo mediante un vector. La necesidad de convertir un texto en un vector viene del hecho de que las computadoras únicamente saben trabajar con números y operaciones aritméticas. Por ejemplo, una cadena de texto existe en la memoria de la computadora como una cadena de números. Esto significa que el texto ya está representado como un vector, y pudiera parecer que esto ya es suficiente para realizar operaciones con él. Pero además es necesario que los vectores de todos los

documentos tengan la misma longitud para que el clasificador puede trabajar con ellos.

Entonces, ¿cómo se toman distintos textos, con distintas palabras y longitudes, y se transforman a una sola representación vectorial que sea capaz de representarlos a todos de la manera más fiel posible? La forma más sencilla, y más usada, es la bolsa de palabras.

2.2.1 Bolsa de palabras

Una de las primeras referencias a la bolsa de palabras puede encontrarse en el artículo *Distributional Structure* [26]. Consiste en utilizar un histograma de las palabras más utilizadas en los documentos para construir una "bolsa" con las N palabras más comunes. En su forma más sencilla, la utilización de la bolsa de palabras para construir una representación vectorial se logra construyendo un vector de longitud N, donde cada dimensión representa una palabra de la bolsa, y se anota en ésta la cantidad de veces que aparece dicha palabra en el documento que se está representando.

Por naturaleza, la bolsa de palabras como representación vectorial de textos tiene la limitación de no poder capturar la información secuencial y relacional entre las palabras. Por ejemplo, la oración "El gato se comió al ratón" tiene la misma representación en una bolsa de palabras que la oración "El ratón se comió al gato". Esto representa un problema si se busca capturar características semánticas y gramaticales. Para minimizar este problema se suele hacer uso de n-gramas para construir la bolsa de palabras, en lugar de las palabras individuales.

La fortaleza de la bolsa de palabras está en poder capturar la información temática de los documentos. En el caso de este trabajo ayudó a identificar que, en este *dataset*, los usuarios con depresión hablan más sobre ellos mismos y sus emociones, mientras que los usuarios sin depresión hablan más sobre política y economía.

Finalmente, una consideración que hay que hacer al usar bolsas de palabras es que son propensas a tener una gran dimensionalidad. Por ejemplo, una bolsa de palabras puede conformarse de 10000 términos, mientras que los documentos que representan pueden no exceder las 1000 palabras, algo particularmente común en el caso de documentos de redes sociales. En casos así, el vector para cada documento estaría compuesto predominantemente por ceros, mientras que en sus pocas dimensiones distintas de cero pueden tener números que no

alcancen las dos cifras.

Es por esto que cuando se construya una bolsa de palabras hay que tomar en cuenta dos cosas: pueden llegar a consumir mucha RAM, y, al ser tan dispersas, pueden meter ruido en el clasificador que entorpezca su desempeño. Ambos problemas se pueden ver aliviados con el siguiente tema: ganancia de información.

Selección de características usando ganancia de información

La medida estadística de ganancia de información se puede usar para estimar la relevancia de cada palabra con cada una de las clases. De esta manera es posible seleccionar únicamente las palabras que sean más significativas para cada una de las clases para la creación de la bolsa de palabras, dejando fuera las palabras que no sirven para diferenciar las clases. En la fórmula 2.1 puede verse cómo se calcula.

$$IG(X, Y) = E(X) - E(X|Y) \tag{2.1}$$

$$E(X) = -\sum p(X) \log p(X) \tag{2.2}$$

Para calcular la ganancia de información IG se requiere calcular la entropía E, cuya fórmula se puede encontrar en 2.2. Donde X es la matriz de datos (los vectores de representación de todos los documentos) y Y es el vector de características (en este caso, la bolsa de palabras). De este proceso se obtiene un valor de ganancia de información para cada término en el documento, el cual va de 0 a 1, donde el valor 1 significa que el término aporta mucha información para discriminar las clases, mientras que el valor 0 significa que el término no aporta nada de información para discriminar las clases.

En este trabajo se utilizó la implementación de Weka [27] de ganancia de información para calcular las palabras que aportan mayor información. Se seleccionaron todas las palabras que tuvieran ganancia de información distinta a 0, pues aplicar cualquier otro umbral de selección resultaba en una lista de palabras de dimensión muy reducida. El uso de la medida de ganancia de información redujo la dimensionalidad de las bolsas de palabras de 30000

términos a alrededor de 2500 términos, a costo del resultado de la clasificación. Sin embargo, fue imprescindible reducir la dimensionalidad de los vectores de esta manera para poder trabajar con los datos en memoria en experimentos subsecuentes.

2.2.2 Pesado

Hasta ahora se ha hablado de cómo lograr convertir un documento a un vector, mediante la metodología de bolsa de palabras. La manera en que se ha abordado ha sido simplemente utilizando conteos del número de veces que aparecen los términos en un documento, algo que se conoce como conteos en bruto. Existen también otras formas de representar los conteos para lograr distintos fines, algo que se conoce como pesado. La finalidad de pesar los conteos es la de relativizar la presencia de los términos en relación a la extensión del documento. A continuación se abordarán dos estilos de pesado: normalización y TF-IDF.

Normalización

La normalización como forma de pesado consiste en dividir cada dimensión del vector de bolsa de palabras entre el número de términos del documento. Una consecuencia de la normalización es que la suma de todas las dimensiones de cada vector es igual a 1. Son dos las ventajas que supone el uso de la normalización:

- Pone en perspectiva el peso de cada término para el documento en que se encuentra
- Permite comparar la presencia de cada término entre todos los documentos

La normalización es la forma más sencilla de pesado y siempre es una opción a considerar cuando se realice una vectorización, no solamente en el dominio del análisis de texto.

TF-IDF

El TF-IDF (frecuencia de términos - frecuencia inversa de documentos, por sus siglas en inglés) es una forma más avanzada de pesado que consiste en tomar en cuenta la frecuencia con la que aparecen cada término en todos los documentos y relativizarla según la frecuencia con la que aparece en el documento en particular que se esté trabajando.

A continuación se presentan las fórmulas para calcularlo.

$$tfidf(t_k, d_j) = tf(t_k, d_j) \cdot idf(t_k) \tag{2.3}$$

$$tf(t_k, d_j) = \begin{cases} 1 + \log \#(t_k, d_j) & \text{si } \#(t_k, d_j) > 0 \\ 0 & \text{en cualquier otro caso} \end{cases} \tag{2.4}$$

$$idf(k) = \log \frac{|D|}{\#_D(t_k)} \tag{2.5}$$

Donde donde t_k representa el término que se está evaluando y d_j representa el documento que se está vectorizando en ese momento. Además, $\#_D(t_k)$ representa la cantidad de documentos en la colección D donde el término t_k ocurre al menos una vez.

Hay que tomar en cuenta, de elegir usar TF-IDF como esquema de pesado, que para poder vectorizar los documentos de esta manera es necesario primero calcular el valor de IDF para cada término del histograma, lo cual puede consumir mucho tiempo y memoria.

La utilidad de TF-IDF está en que resalta la importancia de los términos en el contexto de cada documento, dando poco valor a términos muy comunes en muchos documentos, y dando mucho valor a términos vistos en pocos documentos. De esta forma, TF-IDF también puede funcionar como filtro de términos con poca relevancia o que pueden meter ruido, como artículos y otros términos muy usados del idioma.

2.2.3 Algoritmos de clasificación

Una vez convertido el texto a vectores y aplicado el esquema de pesado deseado se puede proceder a la clasificación. La clasificación tiene dos fases: entrenamiento y prueba. Es necesario apartar los vectores en dos grupos: uno para entrenamiento y otro para prueba. La fase de entrenamiento consiste en alimentar con los vectores reservados para el entrenamiento al algoritmo de clasificación. De esta manera se le da al algoritmo la oportunidad de aprender y encontrar patrones que distinguen a ambas clases. Durante fase de pruebas se da la oportunidad al algoritmo de clasificación de demostrar su efectividad sobre el grupo de vectores

reservados para pruebas, y se evalúa el éxito del clasificador. Es importante utilizar dos grupos distintos de vectores en ambas fases para probar el clasificador sobre datos que nunca había visto antes.

A continución se explicará el algoritmo de clasificación utilizado para este trabajo.

Naïve-Bayes

Para entender la clasificación por Naïve-Bayes primero es necesario comprender cómo funciona el teorema de Bayes y cómo se utiliza.

El teorema de Bayes, llamado así por su creador Thomas Bayes, permite calcular la probabilidad de que un evento A ocurra en el futuro dado que un evento B ya ocurrió en el pasado. Es decir, permite calcular la probabilidad condicional de un evento, y se puede ver la fórmula para calcularlo en 2.6.

$$P(A \mid B) = \frac{P(B \mid A)\, P(A)}{P(B)} \tag{2.6}$$

Donde $P(A)$ es la probabilidad del evento A, y se le conoce como *apriori*. Representa el conocimiento que se tiene sobre el evento A en el pasado. $P(B)$ es la probabilidad del evento B, y se le conoce como evidencia. Significa cuán probable era que el B ocurriera en el presente. $P(B|A)$ es la probabilidad del evento B dada la probabilidad del evento A. Es decir, en cuántas de las ocasiones pasadas en que ha ocurrido A también ocurrió B. A esto se le conoce como *posteriori*.

Adaptado al problema actual, se clasifica cada usuario calculando la probabilidad de que un usuario pertenezca a la clase C dada la probabilidad de la presencia de la característica X, como se puede ver en la fórmula 2.7.

$$P(C \mid X) = \frac{P(X \mid C)\, P(C)}{P(X)} \tag{2.7}$$

Pero antes de poder realizar esta clasificación con el teorema de Bayes es necesario hacer las siguientes suposiciones:

- Que los atributos o características son independientes entre ellos y que tienen igual o similar distribución de probabilidad.

- Que es posible estimar la probabilidad de cada característica dado su presencia en cada clase dentro de los datos de entrenamiento.

En muchos casos, ambas suposiciones no serán ciertas. Sin embargo, resulta útil utilizar este modelo para hacer predicciones, y es considerado *baseline* en el área de NLP. Debido al uso de estas suposiciones es que se le conoce como *Naïve-Bayes* (lit. “Bayes ingenuo”).

El uso de un NBC (clasificador de Näive-Bayes, por sus siglas en inglés) en NLP es muy común por tener la ventaja de necesitar pocos datos para realizar el entrenamiento. Inclusive en la época moderna en la que hay mucha facilidad para recolectar datos gracias a las redes sociales, es difícil encontrar *datasets* especializados. Además, los *datasets* existentes pueden no estar correctamente etiquetados, tener pocas instancias individuales para cada clase o estar muy desbalanceados en cuanto a cantidad de instancias por clase. NBC permite hacer una clasificación acertada aún bajo este tipo de circunstancias. Otra ventaja del NBC específica para NLP es que se alinea muy bien con el funcionamiento de las bolsas de palabras. Las bolsas de palabras representan textos eliminando las relaciones e interdependencias de los términos, y esto significa que la primera suposición de Näive-Bayes es cumplida.

3 Metodología

Se utilizó el *dataset* de *reddit* de [1] para realizar las siguientes pruebas de clasificación y de impacto de características textuales sobre el resultado de clasificación.

El *dataset* consiste de dos conjuntos de datos, uno de entrenamiento y otro de pruebas. El conjunto de pruebas está separado en 10 unidades llamada *chunks*, las cuales tienen cada uno un 10 % de los datos de cada usuario de pruebas. El *dataset* se conforma de la siguiente manera:

Entrenamiento:

- 486 usuarios.
  - 83 % no deprimidos.
  - 17 % deprimidos.
- 292109 publicaciones.

Pruebas:

- 401 usuarios.
  - 87 % no deprimidos.
  - 13 % deprimidos.
- 236344 publicaciones.

Los *chunks* en que se subdivide el *dataset* de pruebas están ordenados cronológicamente. Para fines de este trabajo se evaluó el desempeño del experimento sobre cada *chunk* y se promediaron los 10 resultados.

Se muestra a continuación la tabla 3.1, la cual alista todos los experimentos realizados.

Tabla 3.1: Experimentos realizados

Tipo	Experimento
Base	Bolsa de palabras
Filtrado	Ganancia de información
Pesado	Normalización
	TF-IDF
Textuales	Números
	Minúsculas
	URL
	Negaciones
	n-gramas
	Todas juntas
Temporales	Actividad por hora
	Actividad por periodo del día
	Actividad semanal por hora
Mixtos	Clasificación - actividad por periodos del día
	Clasificación - actividad por periodos de la semana
	Análisis de sentimiento - polaridad por periodos del día
	Análisis de sentimiento - 5 dimensiones por periodos de la semana

3.1 Exploración de configuraciones de procesamiento de texto

Para la exploración de características textuales se combinaron todas las publicaciones de cada usuario y fue sobre este nuevo documento generado que se construyó una bolsa de palabras como representación. Las bolsas de palabras para cada experimento fueron construidas con las mismas características:

- Se construye un histograma de 10000 términos a partir de todos los documentos.
- Se utilizó ganancia de información para seleccionar los términos más relevantes de entre esos 10000. Se eligieron todos los términos que tuvieran ganancia de información mayor a 0.
- Se consideró como token cada palabra individualmente, excepto en el caso del experimento de bi-gramas y tri-gramas.

Finalmente se realiza la clasificación entrenando y utilizando un clasificador de näive-bayes. A estas condiciones se le denominan "experimento base" (o *baseline*), y cada característica explorada es una variación de este experimento base más la característica que se está evaluando.

Las características exploradas son:

- Experimento base con filtrado de números.
- Experimento base con todo en minúsculas.
- Experimento base con filtrado de URL (por ejemplo, la URL `https://www.youtube.com/watch?v=gHYBLM3OdL4` pasa a ser "www-youtube-com").
- Experimento base que considera la negación de un concepto como un concepto individual (ej: "don't want" se considera un solo *token*).
- Experimento base con bi-gramas y tri-gramas, además de términos individuales.
- Experimento base con todas las modificaciones citadas anteriormente.

En el caso del experimento con n-gramas se construye la bolsa de palabras con los 10000 mono-gramas más comunes, más los 10000 bi-gramas más comunes, más los 10000 tri-gramas más comunes, dando un total de 30000 términos en la bolsa previo a la selección por ganancia de información. Se calcula por separado el histograma para cada n-grama debido a que la cantidad de veces que aparecen los tri-gramas son menores que los bi-gramas, y estos a su vez aparecen menos veces que los mono-gramas, por lo que seleccionar los 30000 términos más comunes resultaría en una distribución desigual de mono-gramas, bi-gramas y tri-gramas. De esta manera, se seleccionan n-gramas uniformemente, permitiendo a más n-gramas de mayor dimensión a ser evaluados por el algoritmo de ganancia de información.

3.2 Exploración de características temporales

Estos experimentos únicamente exploran características temporales. La exploración de características temporales se realizó a partir del conteo del número de publicaciones que pertenecen a distintos periodos de tiempo. Hay distintos conjuntos de periodos de tiempo establecidos y en cada experimento se explican cuales fueron utilizado. La representación vectorial del usuario consiste en una dimensión para cada periodo de tiempo establecido, los cuales varían según el experimento en cuestión. Finalmente se realiza la clasificación entrenando y utilizando un clasificador de näive-bayes.

Los experimentos realizados son:

- Actividad por hora (número de publicaciones a cada hora del día, indistintamente del día de la semana)
- Actividad por periodo del día (número de publicaciones a cada mañana, tarde o noche de cada día de la semana)
- Actividad semanal por hora (número de publicaciones a cada hora de cada día de la semana)

Se determinaron estos experimentos, al igual que los periodos del día, tras analizar los patrones de actividad para ambos tipos de usuarios. Este análisis se realizó mediante mapas de calor de actividad de ambos tipos de usuario, mostrados en la figura 3.1. Estos mapas se

realizaron sumando la actividad de todos los usuarios de cada clase y normalizando los datos. Se puede observar que hay diferencias significativas en la actividad a ciertas horas de ciertos días. Más aún, llama la atención la actividad homogénea entre semana para los usuarios sin depresión, a diferencia de los usuarios con depresión.

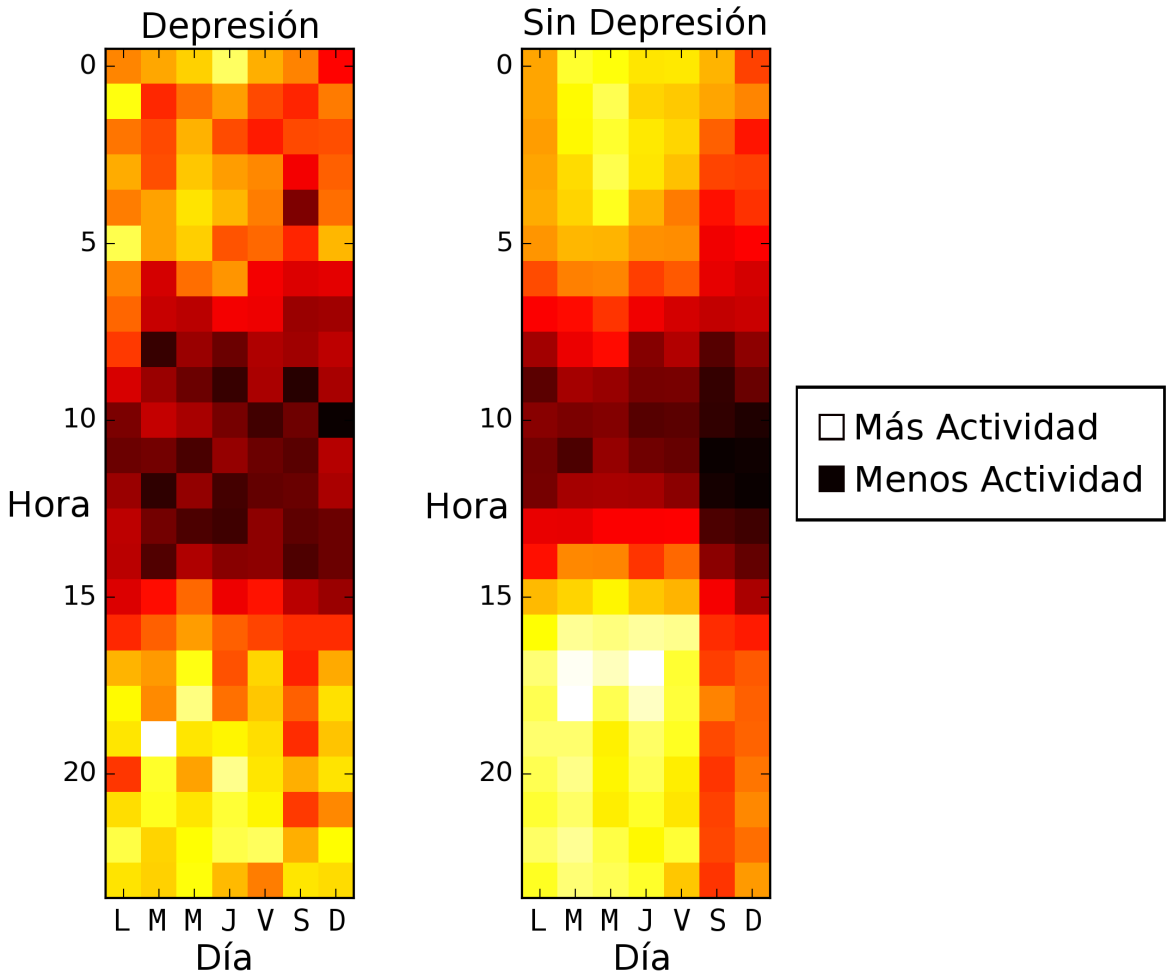

Figura 3.1: Actividad de ambos tipos de usuario durante la semana. Cada casilla representa una hora.

Los experimentos en esta sección se diseñaron para estudiar la actividad de usuarios a través de dos ciclos comunes en nuestras vidas: el diurno y el semanal.

3.2.1 Actividad por hora

El experimento de actividad por hora es el más sencillo de esta categoría. En él se representa a los usuarios con un vector de 24 dimensiones, una por cada hora del día. En este experimento se busca explorar la posibilidad de distinguir entre ambos tipos de usuarios según la actividad que tienen a cada hora del día, sin importar otros factores. Es con este vector que se representan a los usuarios y que se entrena el clasificador.

3.2.2 Actividad por periodo del día

Para el experimento de actividad por periodo del día se establecieron tres periodos del día, mostrados en la tabla 3.2. Estos periodos del día son reutilizados en experimentos posteriores.

Tabla 3.2: Periodos del día establecidos para experimentos

Periodo	Horas
Mañana	6 am a 2 pm
Tarde	2 pm a 10 pm
Noche	10 pm a 6 am

Estos periodos se establecieron en base a la observación de la actividad en la figura 3.1 y en la idea de que es común que la vida de las personas gire en torno a estos periodos del día en los cuales suelen trabajar, dormir, o realizar alguna otra actividad. Este experimento representa los usuarios con un vector de tres dimensiones, cada una correspondiente a estos periodos del día. Se llenaron las casillas del vector con conteos de las publicaciones que cada usuario realizó en esa hora y día en particular. Es con este vector con el que se entrena el clasificador.

3.2.3 Actividad semanal por periodos del día

El experimento de actividad semanal por periodos del día es el más complejo de esta categoría. Consiste en crear una matriz para cada usuario donde cada columna represente un día de la semana y cada fila represente uno de los periodos del día establecidos en 3.2. La

matriz resultante es de 7 por 3 dimensiones. Las casillas de la matriz se llenan con conteos de las publicaciones que cada usuario realizó en es periodo del día y día de la semana en particular. Finalmente, aplana la matriz en un vector unidimensional y esto constituye la representación final del usuario, con la que se entrena el clasificador.

3.3 Exploración de características mixtas en dos pasos

Los experimentos de esta sección combinan las características textuales y temporales para la evaluación de usuarios. Esto se logra mediante el procesamiento de la información a dos pasos. El primer paso consiste en la representación y clasificación de publicaciones individuales de cada usuario, y el segundo paso toma los resultados del primer paso para crear una representación para los usuarios y con esto clasificarlos. Los detalles específicos varían de experimento a experimento y se detallen en sus secciones correspondientes.

La idea detrás de estos experimentos viene de pensar que cada publicación individual es una ventana a la mente del usuario en un punto específico en el tiempo, y que al procesar estos puntos individuales quizá se pueda dar seguimiento a los sentimientos del usuario a través del tiempo.

3.3.1 Clasificación de publicaciones mediante características temporales y textuales mixtas

En estos experimentos, el primer paso busca representar las publicaciones individuales mediante una bolsa de palabras para luego clasificarlas mediante el clasificador de Näive-Bayes . Se tomó la mejor configuración de características textuales para la construcción de la bolsa de palabras. Además de esta bolsa de palabras, el vector de características para las publicaciones incluye dos dimensiones temporales que se eligieron para caracterizar mejor cada publicación: la hora del día a la que fue publicado el texto, y un campo binario que representa si la publicación fue realizada entre semana o en fin de semana (valor 0 o 1). No se exploraron más características temporales, y estas características se conservaron por mejorar el éxito del clasificador.

La clasificación de las publicaciones se realiza heredando la etiqueta del usuario a sus pu-

blicaciones, bajo la idea de entrenar un clasificador que reconozca publicaciones escritas por personas etiquetadas con depresión (a esto se le denominó publicaciones con estado aparente depresivo, o EAD) o escritas por personas etiquetadas sin depresión (a esto se le denominó publicaciones con estado aparente no depresivo, o EAN). Es importante recalcar que esta clasificación de publicaciones no significa que se estén etiquetando publicaciones como "depresivas" o "no depresivas", sino como "escritas por un usuario que aparenta depresión" o "escritas por un usuario que no aparenta depresión". A esto se le llama "clasificar publicaciones según el estado aparente del usuario". La idea tras esta clasificación de publicaciones surge de la observación documentada de que las personas con depresión tienen fluctuaciones en sus estados anímicos, y no siempre parecen tener depresión.

A groso modo, el segundo paso en estos experimentos toma la clasificación de cada publicación para cada usuario y lo asocia al tiempo de la publicación. Los detalles del segundo paso son distintos en cada experimento y se describen a continuación.

Seguimiento de actividad por periodos del día

Con la clasificación de cada publicación en el primer paso se genera un vector para cada usuario de longitud N, donde N es el número de publicaciones de cada usuario. Este primer vector, llamado "vector de estados aparentes", representa las veces que el usuario aparenta tener depresión y las veces que no lo aparenta, ordenadas secuencialmente. En la tabla 3.3 se pueden encontrar los tokens que conforman este vector.

Tabla 3.3: Tokens que conforman el vector de estados aparentes.

Token	Qué representa
0	Estado aparente no depresivo (EAN)
1	Estado aparente depresivo (EAD)

Con el vector de estados aparentes se genera un segundo vector con bi-gramas, llamado "vector de bi-gramas". El vector de bi-gramas intenta capturar las veces en que el usuario presenta el mismo estado aparente, así como las veces en que el usuario presentó un cambio

en su estado aparente. Los tokens que conforman este vector se pueden encontrar en la tabla 3.4.

Tabla 3.4: Tokens que conforman el vector de bi-gramas.

Token	Qué representa
[0, 0]	EAN seguido de EAN
[0, 1]	EAN seguido de EAD
[1, 0]	EAD seguido de EAN
[1, 1]	EAD seguido de EAD

Del vector de estados aparentes se genera un tercer vector, esta vez asociando la etiquetas de estado aparentes a los periodos del día en que fueron publicados. Estos periodos son los mismos utilizados en el experimento actividad por periodo del día y pueden verse en la tabla 3.2. Estos periodos del día se establecieron en base a las horas del día en que hay más diferencia en actividad entre ambos tipos de usuario según la figura 3.1, ignorando las diferencias entre los días de la semana. Además de considerar la actividad durante los periodos del día, para este experimento se incluyeron en este vector tokens que representan periodos de inactividad entre periodos de actividad. Es decir, periodos del día consecutivos en los que no hubo ninguna publicación. Con esto se busca modelar cuánto tiempo pasa entre publicación y publicación. En la tabla 3.5 se pueden observar todos los tokens que se generan para este vector, llamado "vector de actividad".

Tabla 3.5: Tokens que conforman el vector de actividad.

	EAD	EAN	Inactividad
Mañana	m1	m0	mx
Tarde	t1	t0	tx
Noche	n1	n0	nx

Finalmente, se construye la representación del usuario en base a estos tres vectores. La

representación final del usuario se conforma del número de apariciones de cada token en cada uno de los vectores que se describieron anteriormente, dado en porcentajes. En la tabla 3.6 se muestra cada dimensión del vector junto a la característica que representa.

Tabla 3.6: Dimensiones que conforman vector de representación de usuario, los tokens que cada característica representa y el vector del que provienen

Vector de origen	Dimensión	Token asociado
Vector de estados aparentes	Porcentaje de EAD	1
	Porcentaje de EAN	0
Vector de bigramas	Porcentaje de EAD seguido de EAD	[1, 1]
	Porcentaje de EAD seguido de EAN	[1, 0]
	Porcentaje de EAN seguido de EAD	[0, 1]
	Porcentaje de EAN seguido de EAN	[0, 0]
Vector de actividad	Porcentaje de EAD por la mañana	m1
	Porcentaje de EAD por la tarde	t1
	Porcentaje de EAD por la noche	n1
	Porcentaje de EAN por la mañana	m0
	Porcentaje de EAN por la tarde	t0
	Porcentaje de EAN por la noche	n0
	Porcentaje de inactividad por la mañana	mx
	Porcentaje de inactividad por la tarde	tx
	Porcentaje de inactividad por la noche	nx

Con esto se construye un vector final de representación de usuario que consta de 15 dimensiones. Es con este vector que se entrena el clasificador de Näive-Bayes.

Seguimiento de actividad por periodos de la semana

Este experimento, similar al anterior, utiliza para el segundo paso el mismo vector de estados aparentes y el mismo vector de bi-gramas. Además utiliza también un vector de acti-

vidad, pero este vector de actividad es distinto al experimento anterior. En este experimento se analiza la actividad del usuario por periodos de la semana, en lugar de por periodos del día. Para construir este vector de actividad se asocian las etiquetas de estados aparentes a los periodos de la semana en que fueron publicados. Los periodos de la semana pueden verse en la tabla 3.7. Igual que en el experimento anterior, estos periodos se establecieron según la actividad de la figura 3.1, pero esta vez observando grupos de tiempo en toda la semana donde la actividad difiera más entre ambas clases. Se hace una distinción entre la actividad de los días entre semana y los días del fin de semana, juntando toda la actividad del fin de semana en un solo periodo y separando la actividad entre semana en periodos con mayor especificidad.

Tabla 3.7: Periodos del día establecidos para experimentos

Periodo	Horas
Transición 1	5 am a 7 am
Mañana	8 am a 1 pm
Transición 2	2 pm a 5 pm
Noche	6 pm a 6 am
Fin de semana	Todo el fin de semana

Los tokens que representan el los estados aparentes de los usuarios asociados a los periodos de la semana se encuentran en la tabla 3.8. Cabe señalar que para este experimento no se tomó a consideración los periodos de inactividad descritos en el experimento anterior.

Finalmente, se construye la representación del usuario en base a estos tres vectores. La representación final del usuario se conforma del número de apariciones de cada token en cada uno de los vectores que se describieron anteriormente, dado en porcentajes. En la tabla 3.9 se muestra cada dimensión del vector junto a la característica que representa.

El vector final que representa al usuario tiene en total 16 dimensiones. Es con este vector que se entrena el clasificador de Näive-Bayes.

Tabla 3.8: Tokens que conforman el vector de actividad.

	EAD	EAN
Transición 1	t1	t0
Mañana	m1	m0
Transición 2	tt1	tt0
Noche	n1	n0
Fin de semana	w1	w0

Tabla 3.9: Dimensiones del vector de representación de usuario, el token de la característica que representa y el vector del que provienen

Vector de origen	Dimensión	Token asociado
Vector de estados aparentes	Porcentaje de EAD	1
	Porcentaje de EAN	0
Vector de bigramas	Porcentaje de EAD seguido de EAD	[1, 1]
	Porcentaje de EAD seguido de EAN	[1, 0]
	Porcentaje de EAN seguido de EAD	[0, 1]
	Porcentaje de EAN seguido de EAN	[0, 0]
Vector de actividad	Porcentaje de EAD en la primera transición	t1
	Porcentaje de EAD por la mañana	m1
	Porcentaje de EAD en la segunda transición	tt1
	Porcentaje de EAD por la noche	n1
	Porcentaje de EAD en fin de semana	w1
	Porcentaje de EAN en la primera transición	t0
	Porcentaje de EAN por la mañana	m0
	Porcentaje de EAN en la segunda transición	tt0
	Porcentaje de EAN por la noche	n0
	Porcentaje de EAN en fin de semana	w0

3.3.2 Análisis de sentimiento

Los experimentos de esta sección utilizan análisis de sentimiento y valores emocionales para procesar las publicaciones individuales de los usuarios. El primer experimento, como los experimentos anteriores, también requiere dos pasos para la representación y clasificación de los usuarios. El segundo experimento construye una matriz con valores emocionales para varios periodos de tiempo, y es directamente con esta información que representa al usuario.

Los valores emocionales de cada publicación fueron extraídos mediante la herramienta SenticNet [21]. SenticNet asigna 5 valores a un texto dado, llamados dimensiones emocionales. Estas dimensiones son: polaridad, amenidad, consideración, sensibilidad y aptitud. SenticNet consiste en una base de datos para más de 10000 términos en inglés, para los cuales se tienen asignado un valor entre el rango 0 a 1 en cada una de las 5 dimensiones emocionales (excepto polaridad, la cual tiene un rango que va de -1 a 1). El valor de cada publicación se calcula a partir del el promedio de los valores individuales de cada término que conforma a la publicación. Cabe señalar que es necesario convertir los valores de polaridad sentimental al rango 0 a 1 para poder utilizar Näive-Bayes como clasificador, pues dicho clasificador no es compatible con números negativos.

Se realizaron dos experimentos con análisis de sentimiento: seguimiento de polaridad por periodos del día y seguimiento de 5 dimensiones emocionales por periodos de la semana.

Seguimiento de polaridad por periodos del día

La idea tras este experimento fue la de identificar la polaridad de sentimiento que experimentan los usuarios durante los periodos del día establecidos en la tabla 3.2 y determinar si se podía encontrar un patrón de sentimiento que diferenciara a ambos tipos de usuario.

Para realizar este experimento se comenzó asignando un valor de polaridad emocional a cada publicación de cada usuario mediante la librería SenticNet. De aquí se obtiene un valor 0 o 1 para cada publicación, lo cual resulta idéntico a la clasificación por NBC usada en el experimento “seguimiento de actividad por periodos del día”. Este vector se denomina “vector de polaridad sentimental”, y sus tokens pueden encontrarse en la tabla 3.10. Al terminar de construir el vector de polaridad sentimental se concluye el primer paso.

Tabla 3.10: Tokens que conforman el vector de polaridad sentimental.

Token	Qué representa
0	Polaridad positiva (PP)
1	Polaridad negativa (PN)

Comenzando el segundo paso, el experimento es casi idéntico al experimento "seguimiento de actividad por periodos del día". Primero se construye un vector de bi-gramas a partir del vector de polaridad sentimental. Los tokens que conforman este vector se pueden encontrar en la tabla 3.11.

Tabla 3.11: Tokens que conforman el vector de bi-gramas.

Token	Qué representa
[0, 0]	PP seguido de PP
[0, 1]	PP seguido de PN
[1, 0]	PN seguido de PP
[1, 1]	PN seguido de PN

A continuación se asocia la polaridad de las publicaciones al periodo del día en que se publicaron. Los periodos del día son los mismos que se definieron en la tabla 3.2. Para ello se construye un vector de actividad, conformado por los tokens descritos en la tabla 3.12. Para este experimento no se utilizaron los tokens de inactividad descritos en el experimento seguimiento de actividad por periodos del día.

Finalmente, se construye la representación del usuario en base a estos tres vectores. La representación final del usuario se conforma del número de apariciones de cada token en cada uno de los vectores que se describieron anteriormente, dado en porcentajes. En la tabla 3.13 se muestra cada dimensión del vector junto a la característica que representa.

Es con esta información que se construye un vector final de representación de usuario, el cual consta de 15 dimensiones. Con este vector se entrena el clasificador de Näive-Bayes.

Tabla 3.12: Tokens que conforman el vector de actividad.

	PN	PP
Mañana	m1	m0
Tarde	t1	t0
Noche	n1	n0

Tabla 3.13: Dimensiones del vector de representación de usuario, el token de la característica que representa y el vector del que provienen

Vector de origen	Dimensión	Token asociado
Vector de polaridad sentimental	Porcentaje de PN	1
	Porcentaje de PP	0
Vector de bigramas	Porcentaje de PN seguido de PN	[1, 1]
	Porcentaje de PN seguido de PP	[1, 0]
	Porcentaje de PP seguido de PN	[0, 1]
	Porcentaje de PP seguido de PP	[0, 0]
Vector de actividad	Porcentaje de PN por la mañana	m1
	Porcentaje de PN por la tarde	t1
	Porcentaje de PN por la noche	n1
	Porcentaje de PP por la mañana	m0
	Porcentaje de PP por la tarde	t0
	Porcentaje de PP por la noche	n0

Seguimiento de 5 dimensiones emocionales por periodos de la semana

La idea detrás de este experimento es la de observar y analizar las emociones de los usuarios por semana. Es decir, ¿cómo se sienten los usuarios durante cada momento de la semana? ¿Cómo son los cambios emocionales durante la semana promedio para alguien con depresión y para alguien sin depresión? Utilizando la librería de SenticNet se intentó responder a estas preguntas mediante este experimento.

Se comienza extrayendo los valores de las 5 dimensiones emocionales para cada publicación de cada usuario, y con ellas llenar la tabla 3.14 que construyó para representar la semana. En cada casilla se suman todos los valores de todos las publicaciones que corresponden a ese periodo del día de ese día en particular. Dado que cada casilla contiene 5 valores distintos, realmente esta representación es una matriz de 3 dimensiones.

Nótese la columna y fila "Total". En la columna Total se suman los valores de cada periodo del día. Así mismo, la fila Total suma todos los valores de cada día de la semana. La intersección de ambas columnas representa la suma de todos los valores de la tabla, o lo que es lo mismo, todos los valores del usuario. Para finalizar la construcción de esta tabla, se normalizan todos los valores, dividiendo entre la cantidad de publicaciones que corresponden a cada casilla.

Tabla 3.14: Representación semanal

	Lunes	Martes	Miércoles	Jueves	Viernes	Sábado	Domingo	Total
Mañana								
Tarde								
Noche								
Total								

Una vez construida esta matriz se aplanó en un vector unidimensional y se calculó ganancia de información para cada valor. De esta manera se redujo la dimensionalidad de la representación, y en el proceso se pudo encontrar los días, periodos y valores emocionales más significativos de ambas clases. Y es entonces que se procede a entrenar un clasificador

de Näive-Bayes.

3.4 Propuesta para concurso del foro eRisk

Como fue mencionado al principio de este capítulo, el *dataset* utilizado para los experimentos fue construido el concurso del foro eRisk del CLEF 2017. Este concurso requirió una propuesta metodológica para la detección de depresión. El *dataset* fue dividido en 10 *chunks* para la evaluación. Cada semana se iba liberando un nuevo *chunk* de pruebas y se debía evaluar el desempeño del algoritmo de clasificación sobre estos nuevos datos. El algoritmo, además, debía de ser capaz de determinar si ya tenía la seguridad suficiente para afirmar si cada usuario tiene o no depresión. Una vez que se toma esa decisión ya no se puede reevaluar ese usuario. El concurso permitía entregar hasta 5 participaciones distintas.

Debido a restricciones de tiempo, la experimentación para desarrollar el método ocurrió durante estas semanas de entrega de datos, por lo que no fue posible participar semanalmente. Sin embargo, se incorporaron cuatro algoritmos de toma de decisión y se entregaron cuatro resultados en la décima semana del concurso. Los cuatro algoritmos de toma de decisión se basan en clasificar a cada usuario en cada *chunk* y son los siguientes:

- Clasificación más común en los últimos 3 *chunks*
- Clasificación más común en los últimos 5 *chunks*
- Clasificación más común en los primeros 3 *chunks*
- Clasificación más común en los primeros 5 *chunks*

Se puede ver que hay dos tipos de toma de decisión, una basada en los últimos x *chunks* y otra basada en los primeros x *chunks*. En ambos casos, se decide no tomar una decisión si no se cuenta con la correspondiente cantidad de chunks. Los algoritmos que toman los primeros x *chunks* fueron pensados para clasificar lo más rápidamente posible, dado que el concurso trataba sobre detección temprana de depresión. Desafortunadamente, este enfoque no ve toda la información del usuario. Dada la división cronológica de información en chunks, los primeros bloques de información corresponden a la información más antigua, y no a la

más reciente. Es por esto que se incorporaron los algoritmos que toman en cuenta los últimos x *chunks*. Aunque no cuenten para el concurso como detección temprana, sí toman en cuenta la información más reciente del usuario.

4 Experimentación y Resultados

A continuación se detalla la manera en que fueron evaluados los experimentos y se disponen los resultados de los mismos. Se pueden encontrar tres tipos de resultados: el desempeño de la clasificación de los usuarios, las listas de palabras relevantes que se generan para cada clase y el resultado de la participación en el foro eRisk de CLEF 2017.

4.1 Métricas de desempeño

En el presente trabajo se utilizan dos métricas de desempeño: valor-F y ERDE. Valor-F fue utilizado para comparar el desempeño de los distintos experimentos entre ellos. ERDE es la medida de desempeño propuesta para el concurso eRisk del CLEF 2017 y se utilizó para comparar la propuesta para el concurso con los otros participantes, junto al valor-F.

4.1.1 Valor-F

Para evaluar los resultados de desempeño se utilizó la medida de evaluación valor-F, conocida también como *F-score*. Esta medida de evaluación se eligió por ser útil para medir el desempeño de predicciones binarias. Para calcularse es primero necesario calcular la "matriz de confusión", la cual se puede encontrar en la tabla 4.1.

Tabla 4.1: Matriz de confusión y sus partes

Clase real	Clase Predicha	
	Negativa	Positiva
Negativa	TN	FP
Positiva	FN	TP

Donde TP significa *True Positive* (verdadero positivo), TN significa *True Negative* (verdadero negativo), FP significa *False Positive* (falso positivo), y FN significa *False Negative* (falso negativo). Estos parámetros representan el tipo de resultado de clasificación posible. La matriz de confusión se llena con los conteos correspondientes para cada tipo de resultado. El uso de la matriz de confusión permite analizar los resultados de clasificación en términos de cuántas instancias fueron asignadas correctamente a sus clases y cuántas incorrectamente, así como el porcentaje de éxito de clasificación para cada clase.

Teniendo esta información es posible calcular el valor-F. En la fórmula 4.1 se puede ver que para calcular el valor-F se necesitan las medidas de rendimiento Precisión (ver fórmula 4.2) y Exhaustividad (ver fórmula 4.3).

$$F_1 = 2 \times \frac{\text{Precisión} \times \text{Exhaustividad}}{\text{Precisión} + \text{Exhaustividad}} \tag{4.1}$$

$$\text{Precisión} = \frac{TP}{TP + FP} \tag{4.2}$$

$$\text{Exhaustividad} = \frac{TP}{TP + FN} \tag{4.3}$$

En las tablas de resultados presentadas en las secciones subsecuentes se hace un desglose del valor-F para cada clase con tal de observar el impacto sobre la clasificación por separado de las clases, algo que es importante dado que la tarea se trata de identificar correctamente a los usuarios de la clase depresiva.

4.1.2 ERDE

La métrica de desempeño ERDE (*early risk detection error*, lit. “error de detección temprana de riesgo”) fue propuesta en [1] junto al *dataset* usado en el presente trabajo con la finalidad de analizar el desempeño de los clasificadores para detectar el riesgo de depresión de manera temprana, a la vez de estudiar las limitaciones de la métrica misma. Esta métrica penaliza al clasificador mientras más datos requiera para tomar una decisión.

Para realizar el cálculo se requiere de una toma de decisión binaria d y de un retraso k. Este retraso se refiere a la cantidad de publicaciones de un usuario que el sistema necesitó analizar para poder tomar la decisión sobre dicho usuario. Por ejemplo, si un usuario tiene 25 publicaciones en cada *chunk*, y el sistema necesita de 2 *chunks* para tomar la decisión de clasificación, se dice que se tiene un retraso k de 50.

Entonces, ERDE se evalúa 4 casos mediante la fórmula 4.4.

$$ERDE_0(d,k) = \begin{cases} c_{fp} & \text{si } d = \text{positivo y } \textit{ground truth} = \text{negativo (FP)} \\ c_{fn} & \text{si } d = \text{negativo y } \textit{ground truth} = \text{positivo (FN)} \\ lc_0(k) \cdot c_{tp} & \text{si } d = \text{positivo y } \textit{ground truth} = \text{negativo (TP)} \\ 0 & \text{si } d = \text{negativo y } \textit{ground truth} = \text{positivo (TN)} \end{cases} \tag{4.4}$$

Para esta tarea, el valor de c_{fn} corresponde a 1. El valor de c_{fp} corresponde a la proporción de usuarios para cada clase, lo que equivale a 0.1296. El valor de c_{tp} es igual a c_{fn}, por lo que equivale a 1. La función $lc_o(k)$ es la responsable de penalizar el tiempo de evaluación y se calcula con la fórmula 4.5.

$$lc_o(k) = 1 - \frac{1}{1 + e^{k-o}} \tag{4.5}$$

Esta función $lc_o(k)$ recibe como parámetro el valor de o, el cual establece la cantidad de publicaciones al rededor del cual el costo de decisión aumenta. Así pues, la penalización ocurre entorno a una cantidad fija de publicaciones para cada usuario, y no con el porcentaje de publicaciones utilizadas para cada usuario.

El concurso fue evaluado con dos tipos de ERDE: ERDE5 y ERDE50. Esto se refiere al valor o utilizado en cada una, siendo en ERDE5 $o = 5$ y en en ERDE50 $o = 50$. Es decir, ERDE5 comienza a penalizar al clasificador si requiere más de 5 publicaciones para tomar una decisión, mientras que ERDE50 penaliza al clasificador si requiere 50 o más publicaciones para tomar una decisión.

4.2 Resultado del experimento base (bolsa de palabras)

A continuación se presenta el resultado que corresponde al experimento denominado "Experimento base". Consiste en una bolsa de palabras construida a partir de 10000 términos sin procesamiento alguno, clasificados con un Clasificador Naïve-Bayes. Se utiliza como experimento de referencia al ser el más sencillo que se puede realizar en clasificación de textos. En posteriores tablas de esta sección se hace referencia a este experimento con el nombre "Base".

Tabla 4.2: Resultado de valor-F para experimento de bolsa de palabras o experimento base

Experimento	Clase No Depresiva	Clase Depresiva
Experimento base	0.7840	0.4229

De utilizar el valor-F general de la clasificación, sin el desglose por clases, el resultado sería 0.6034. No sería posible observar el desbalanceo de las clases, ni el desempeño específico sobre la clase de interés, que es la depresiva.

Se puede observar que, si bien se cuenta con moderado éxito de clasificación para todos los usuarios, la clase depresiva se resiste a ser bien caracterizada.

Finalmente, al explorar los subsecuentes experimentos se requirió que se mejorara la clasificación sobre el experimento base.

4.3 Resultados del uso de ganancia de información en bolsa de palabras

El resultado de este experimento demuestra que la selección de palabras con mayor ganancia de información entorpece la clasificación, como se ve en la tabla 4.3. A pesar de contar con las palabras más relevantes, se pierde mucha información en el proceso, y se teoriza que es por esta razón que disminuye el éxito de clasificación.

Sin embargo, el filtrado de términos por ganancia de información fue utilizado en algunos experimentos subsecuentes para poder trabajar con el volumen de información generado dentro del límite de la memoria RAM.

Tabla 4.3: Resultados de valor-F para el experimento de uso de ganancia de información

Experimento	Clase No Depresiva	Clase Depresiva
Experimento base	0.7840	0.4229
Ganancia de información	0.7961	0.4156

4.4 Resultados del uso de esquemas de pesado

A continuación se presentan los resultados del uso de los esquemas de pesado. La tabla 4.4 muestra que el uso de la normalización introduce ruido en el clasificador, evitando que pueda identificar usuarios de la clase depresiva. Al mismo tiempo, el uso de TF-IDF sí permite hacer distinción entre ambos tipos de usuarios, pero no supera al experimento base.

Tabla 4.4: Resultados de valor-F para experimentos de esquemas de pesado

Experimento	Clase No Depresiva	Clase Depresiva
Experimento base	0.7840	0.4229
Normalización	0.9307	0.0000
TF-IDF	0.7722	0.4172

Llama la atención que tanto filtrado por ganancia de información como TF-IDF no hayan mejorado el éxito de las clasificación, pues ambas suelen ser muy usadas en tareas de NLP y tener buenos resultados. Esto sugiere que la naturaleza de esta tarea es distinta de otras tareas de NLP, y que tanto eliminar ciertas características aparentemente ambiguas como cambiar el pesado de los términos en la bolsa de palabras quita información útil para distinguir a ambos grupos de usuarios.

4.5 Resultados de clasificación en experimentos temporales, textuales y mixtos

En la tabla 4.5 se puede observar los resultados del uso de características textuales sobre el impacto de la clasificación con Naïve-Bayes en una bolsa de palabras con ganancia de información. La diferencia en el resultado de clasificación es mínima, pero sí se encuentran mejoras. Así mismo, el mejor resultado no se obtuvo al juntar todas las características, sino al usar la concatenación de negaciones.

Tabla 4.5: Resultados de valor-F para experimentos de características textuales

Experimento	Clase No Depresiva	Clase Depresiva
Experimento base	0.7840	0.4229
Números	0.7730	0.4118
Minúsculas	0.7818	0.4251
URL	**0.7853**	0.4241
Negaciones	0.7746	**0.4255**
n-gramas	0.7775	0.4159
Todas juntas	0.7748	0.4105

Solo tres experimentos no superaron al experimento base: quitar números, n-gramas y todas las alteraciones juntas. Llama la atención que el uso de n-gramas entorpezca al clasificador. Los n-gramas incrementan la información disponible para el clasificador al incorporar secuencialidad de las palabras. Esto puede significar que confunde al clasificador, en lugar de enriquecerlo. Al mismo tiempo, limpiar las URL y quitar las minúsculas mejoran el resultado, pese a remover información en ambos casos. Sin embargo, quitar los números sí reduce el éxito del clasificador, lo que sugiere que sí aportan información útil. Finalmente, la concatenación de negaciones es la característica que más beneficia al clasificador. Aunque solo sean un tipo particular de n-gramas, la información específica que reúne sí enriquece al clasificador.

La tabla 4.6 muestra los resultados de experimentos temporales. Se puede observar que, por sí mismas, las características temporales no fueron capaces de diferenciar a los dos tipos de usuarios.

Tabla 4.6: Resultados de valor-F para experimentos de características temporales

Experimento	Clase No Depresiva	Clase Depresiva
Experimento base	0.7840	0.4229
Actividad por hora	**0.9307**	0.0000
Actividad por periodo del día	0.9300	0.0000
Actividad semanal por periodos del día	0.0000	**0.2296**

A continuación se muestra la tabla 4.7 con los resultados de experimentos con características mixtas. Se puede observar que el experimento de clasificación con actividad por periodos del día mejora los resultados del experimento base para ambas clases. Así mismo, se puede observar que el uso de análisis de sentimiento para representar las publicaciones no da mejores resultados que el experimento base, y es incapaz de identificar usuarios de la clase depresiva.

Tabla 4.7: Resultados de valor-F para experimentos de características mixtas

Experimento	Clase No Depresiva	Clase Depresiva
Experimento base	0.7840	0.4229
Clasificación - actividad por periodos del día	0.8852	**0.4758**
Clasificación - actividad por periodos de la semana	**0.8923**	0.4426
Análisis de sentimiento - polaridad por periodos del día	0.7331	0.2642
Análisis de sentimiento - 5 dimensiones por periodos de la semana	0.6272	0.2327

El que los experimentos sobre los periodos de tiempo por sí mismos no hayan podido clasificar a los usuarios depresivos, pero que el experimento mixto sobre esos mismos periodos de tiempo sí haya superado a la bolsa de palabras resulta muy interesante. La combinación

de la información temática, capturada por la bolsa de palabras, con la información temporal, capturada por los periodos del día, parece caracterizar mejor a los usuarios para distinguir entre ellos.

4.6 Resultados de palabras más relevantes según el procesamiento de texto

En esta sección se mostrarán y compararán los primeros 25 términos con mayor ganancia de información según experimento textual, tal como se detalló en la sección 3.1. Se eligieron 25 términos por cuestiones de espacio, pero las listas completas son de entre 1000 y 3700 términos.

La tabla 4.8 muestra todos los experimentos con sus respectivos términos. Si bien es posible encontrar ciertas similitudes entre las listas, al cambiar la configuración de características textuales se pueden encontrar términos que no se encuentran en otras listas.

Como es de esperarse, los experimentos que usan n-gramas cuentan con más información por capturar secuencias de palabras. Pero se encontraron hallazgos inesperados en términos como “me_.”, el cual muestra que no solo es relevante el término “me”, sino que es importante cuando se usa para finalizar la oración. Esto sugiere que, aunque por sí mismos no permitan clasificar mejor a los usuarios, sí aportan información útil para analizar los patrones de habla de los usuarios.

Otro hallazgo que llama la atención se encuentra en experimentos que no procesan las URLs, ya que muestran que términos que pudieran ser considerados como ruido aportan ganancia de información, como “http”. No es posible saber si esto es consecuencia de algún problema en la construcción del *dataset*, pero sí es posible interpretarse que la presencia de enlaces es relevante para distinguir ambos tipos de usuario.

Finalmente, resalta la semántica de los primeros 25 términos. Indistintamente del procesamiento de texto, se observa que hablar sobre uno mismo, sentimientos y relaciones interpersonales son lo que más información aporta a la hora de discriminar usuarios. Así mismo, al ver la lista completa de palabras se encuentran muchos términos políticos, como nombres de países o mandatarios políticos. Cabe señalar que no es posible determinar si la presencia

de algún tema en particular es consecuencia de la metodología de la construcción del *dataset* o si realmente es representativo de las diferencias entre ambos tipos de usuarios en cualquier contexto.

4.7 Resultados del concurso y evaluación ERDE

Para la competencia eRisk del CLEF 2017 se presentó el método "Clasificación - actividad por periodos del día" de la tabla 4.7 por ser el que mejor resultados obtuvo para clasificar a los usuarios depresivos. En base a ese clasificador se presentaron las cuatro propuestas de toma de decisión al concurso:

- CHEPEA - Clasificación más común en los últimos 3 *chunks*
- CHEPEB - Clasificación más común en los últimos 5 *chunks*
- CHEPEC - Clasificación más común en los primeros 3 *chunks*
- CHEPED - Clasificación más común en los primeros 5 *chunks*

En total participaron 8 equipos: GPL, FHDO-BCSG, UArizona, LyR, UNSL, UQAM, NLPIS, y CHEPE (nosotros). En la tabla 4.9 se muestra los resultados de valor-F, de ERDE5 y de ERDE50.

Cabe señalar que esta tarea se realizó como proyecto piloto para comenzar a recaudar información sobre la caracterización de usuarios depresivos en internet, y que aún falta mucho trabajo que realizar en esta área. Con esto en mente, se puede observar que el rango de resultados para el valor-F es bastante amplio para todos los participantes. A pesar de esto, el resultado más alto fue de 0.64, lo que sugiere que aún no se encuentra cómo atacar adecuadamente la complejidad de esta tarea.

Respecto a los valores ERDE, la brecha entre los equipos no es tan amplia. Particularmente llama la atención que, para ambos casos, la propuesta del presente trabajo se encontró a 3 % o menos del primer lugar, a pesar de que fuimos el último equipo en entregar resultados. Esto sugiere que la métrica de evaluación penaliza muy prematuramente, y que pasado este umbral de penalización se vuelve indistinto en qué momento se decida clasificar al usuario.

Tabla 4.8: 25 palabras con mayor ganancia de información para los experimentos de características textuales

Base	Números	Minúsculas	URLs	Negación	n-gramas	Todas
my	my	/	feel	feel	my	me
/	I'm	feel	myself	com	/	i'm
I'm	/	://	:)	myself	I'm	like
The	The	com	friends	:)	The	i_was
feel	feel	http	depression	depression	I_was	and_i
://	://	myself	new	friends	I_have	feel
com	com	:)	pain	new	and_I	but_i
http	http	new	talk	pain	feel	._the
myself	myself	depression	its	its	2	._i'm
:)	:)	friends	tried	talk	but_I	i_can
A	A	www	girl	girl	._The	that_i
=	friends	its	feeling	tried	://	myself
friends	depression	sometimes	sometimes	feeling	com	me_.
depression	new	talk	kids	sometimes	._com	if_i
new	www	pain	sleep	kids	I_can	i_had
www	In	girl	anxiety	sleep	http	of_my
In	pain	tried	felt	anxiety	http_://	:)
pain	talk	feeling	$	felt	that_I	game
talk	friend	sleep	alone	$	com_/	new
friend	its	kids	relationship	relationship	._com_/	for_me
its	tried	anxiety	show	show	me_.	depression
tried	girl	she's	stay	stay	myself	i_feel
girl	feeling	felt	leave	she's	I_just	because_i
feeling	sometimes	black	she's	bed	I_had	friends

Tabla 4.9: Resultados del CLEF eRisk 2017 ordenados por valor-F, ERDE5 y ERDE50

Equipo	valor-F
FHDO-BCSGA	0.64
FHDO-BCSGE	0.60
UNSLA	0.59
FHDO-BCSGD	0.57
FHDO-BCSGC	0.56
FHDO-BCSGB	0.55
UQAMA	0.53
UQAMB	0.48
CHEPEA	**0.48**
GPLD	0.47
CHEPEB	**0.47**
GPLC	0.46
CHEPEC	**0.46**
UArizonaD	0.45
UArizonaE	0.45
CHEPED	**0.45**
UQAMC	0.42
UArizonaA	0.40
UQAME	0.39
UQAMD	0.38
GPLA	0.35
UArizonaC	0.34
GPLB	0.30
ArizonaB	0.30
LyRB	0.16
LyRC	0.16
LyRD	0.15
NLPISA	0.15
LyRA	0.14
LyRE	0.08

Equipo	ERDE5
FHDO-BCSGB	12.70 %
FHDO-BCSGA	12.82 %
FHDO-BCSGD	13.04 %
UArizonaB	13.07 %
UQAMD	13.23 %
FHDO-BCSGC	13.24 %
UQAMC	13.58 %
UNSLA	13.66 %
UQAME	13.68 %
LyRE	13.74 %
UQAMB	13.78 %
UQAMA	14.03 %
GPLC	14.06 %
FHDO-BCSGE	14.16 %
GPLD	14.52 %
UArizonaA	14.62 %
UArizonaD	14.73 %
CHEPEA	**14.75 %**
CHEPEB	**14.78 %**
CHEPEC	**14.81 %**
CHEPED	**14.81 %**
UArizonaE	14.93 %
LyRD	14.97 %
NLPISA	15.59 %
LyRA	15.65 %
LyRC	16.14 %
LyRB	16.75 %
GPLA	17.33 %
UArizonaC	17.93 %
GPLB	19.14 %

Equipo	ERDE50
UNSLA	9.68 %
FHDO-BCSGA	9.69 %
UArizonaD	10.23 %
FHDO-BCSGB	10.39 %
FHDO-BCSGD	10.53 %
FHDO-BCSGC	10.56 %
UArizonaB	11.63 %
UQAMD	11.98 %
UArizonaE	12.01 %
GPLC	12.14 %
CHEPEA	**12.26 %**
UQAMA	12.29 %
CHEPEB	**12.29 %**
FHDO-BCSGE	12.42 %
CHEPEC	**12.57 %**
CHEPED	**12.57 %**
UArizonaA	12.68 %
UQAME	12.68 %
UArizonaC	12.74 %
GPLD	12.78 %
UQAMB	12.78 %
UQAMC	12.83 %
LyRE	13.74 %
LyRD	14.47 %
LyRA	15.15 %
LyRC	15.51 %
NLPISA	15.59 %
LyRB	15.76 %
GPLA	15.83 %
GPLB	17.15 %

4.8 Hallazgos sobre el *dataset*

Los resultados presentados en esta sección muestran el resultado de clasificar a todos los usuarios del apartado de pruebas del *dataset*. Sin embargo, el *dataset* permite hacer pruebas más detalladas. A continuación, en la tabla 4.10, se muestran los resultados del algoritmo más exitoso, "Clasificación - actividad por periodos del día", sobre cada uno de los 10 *chunks* por separado.

Tabla 4.10: Resultados de clasificación desglosados por *chunks*

Chunk	Clase No Depresiva	Clase Depresiva
1	0.8665	0.4557
2	**0.8992**	**0.5109**
3	0.8798	0.4552
4	0.8828	0.4690
5	0.8807	0.4730
6	0.8909	0.4930
7	0.8832	0.4615
8	0.8882	0.4714
9	0.8912	0.4857
10	0.8896	0.4823

Resalta el hecho de que en el segundo *chunk* se obtiene el mejor resultado, así como la varianza que hay en los resultados. Esto mismo fue observado por los otros participantes. Así mismo se señaló que en el *dataset* de entrenamiento se obtuvieron considerablemente mejores resultados por parte de todos los participantes. En la tabla 4.11 se contrasta el resultado obtenido en el *dataset* de entrenamiento contra el de pruebas, también para el algoritmo "Clasificación - actividad por periodos del día".

Se desconoce la razón por la que existe tanta discrepancia en los resultados. Es posible que haya algún sesgo en la construcción del *dataset* que haga que los usuarios entre ambos

Tabla 4.11: Resultados de clasificación en *dataset* de entrenamiento contra *dataset* de pruebas

Dataset	Clase No Depresiva	Clase Depresiva
Entrenamiento	**0.9243**	**0.7345**
Pruebas	0.8852	0.4758

conjuntos de datos sean distintos, pero los autores aseguraron durante la conferencia CLEF 2017 que fueron elegidos al azar.

5 Conclusiones

En este trabajo se exploraron los efectos de distintas características lingüísticas y temporales para la clasificación de usuarios con y sin depresión. Además, se estudia el uso de algunos esquemas de pesado, y de filtrado por ganancia de información. También se realizó análisis de sentimientos como características lingüísticas, y se estudiaron formas de combinar información lingüística y textual.

De esta exploración surge un método para la clasificación de usuarios a dos pasos que primero extrae información de las publicaciones individuales y luego realiza un metaanálisis de esta información para clasificar al usuario. Este método, llamado "procesamiento a dos pasos", se cree que tiene potencial para ser utilizado en otras áreas donde se requiera clasificar una instancia a partir de múltiples instancias, y no solamente en la tarea de detección de usuarios de redes sociales con depresión o en el área de procesamiento de lenguaje natural.

Como se señaló en la hipótesis, es posible encontrar patrones en las publicaciones de usuarios de redes sociales que sean capaces de diferenciar a usuarios con depresión de usuarios sin depresión. Estos patrones pueden encontrarse en características lingüísticas y textuales de manera conjunta, aunque también por sí mismas permitan hacer un análisis cualitativo.

La exploración de esta área es limitada por la información contenida en el *dataset* utilizado, pero es un área de oportunidad enorme. Surgen preguntas como: ¿Serán distintos los usuarios de otras redes sociales, como *facebook* y *twitter*, a los de *reddit*? ¿De qué manera se puede extraer y estudiar información de otras actividades en línea para estudiar usuarios? Por ejemplo, juegos en línea, hábitos de navegación, etc..

En lo que concierne al presente trabajo, se considera que se alcanzaron los objetivos y que la hipótesis fue demostrada.

5.1 Trabajo futuro

Como trabajo futuro se pueden considerar las siguientes opciones:

- Probar otras representaciones de texto, en lugar de bolsa de palabras. Particularmente se considerarían métodos como CSA [18] y TCOR [24].
- Probar otros clasificadores, en lugar de NBC.
- Encontrar o construir nuevos *datasets* de otras redes sociales.
- Probar otras librerías para el análisis de emociones.
- Utilizar la metodología de procesamiento de usuarios a dos pasos en otros dominios, como identificación de autoría.

Referencias

[1] D. E. Losada and F. Crestani, "A test collection for research on depression and language use," in *International Conference of the Cross-Language Evaluation Forum for European Languages*, pp. 28–39, Springer, 2016.

[2] C. for Disease Control, P. (CDC, *et al.*, "Current depression among adults—united states, 2006 and 2008.," *MMWR. Morbidity and mortality weekly report*, vol. 59, no. 38, p. 1229, 2010.

[3] M. Hu and B. Liu, "Mining opinion features in customer reviews," in *AAAI*, vol. 4, pp. 755–760, 2004.

[4] E. Kouloumpis, T. Wilson, and J. D. Moore, "Twitter sentiment analysis: The good the bad and the omg!," *Icwsm*, vol. 11, no. 538-541, p. 164, 2011.

[5] G. Coppersmith, M. Dredze, C. Harman, and K. Hollingshead, "From adhd to sad: Analyzing the language of mental health on twitter through self-reported diagnoses," *NAACL HLT 2015*, p. 1, 2015.

[6] T. Nguyen, D. Phung, B. Dao, S. Venkatesh, and M. Berk, "Affective and content analysis of online depression communities," *IEEE Transactions on Affective Computing*, vol. 5, no. 3, pp. 217–226, 2014.

[7] M. De Choudhury, S. Counts, E. J. Horvitz, and A. Hoff, "Characterizing and predicting postpartum depression from shared facebook data," in *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, pp. 626–638, ACM, 2014.

[8] S. Park, S. W. Lee, J. Kwak, M. Cha, and B. Jeong, "Activities on facebook reveal the depressive state of users," *Journal of medical Internet research*, vol. 15, no. 10, p. e217, 2013.

[9] M. J. Paul and M. Dredze, "You are what you tweet: Analyzing twitter for public health.," *Icwsm*, vol. 20, pp. 265–272, 2011.

[10] M. De Choudhury, M. Gamon, S. Counts, and E. Horvitz, "Predicting depression via social media.," in *ICWSM*, p. 2, 2013.

[11] M. Park, D. W. McDonald, and M. Cha, "Perception differences between the depressed and non-depressed users in twitter.," in *ICWSM*, 2013.

[12] M. Nadeem, "Identifying depression on twitter," *arXiv preprint arXiv:1607.07384*, 2016.

[13] G. B. Colombo, P. Burnap, A. Hodorog, and J. Scourfield, "Analysing the connectivity and communication of suicidal users on twitter," *Computer communications*, vol. 73, pp. 291–300, 2016.

[14] H. Almeida, A. Briand, and M.-J. Meurs, "Detecting early risk of depression from social media user-generated content," 2017.

[15] F. Sadeque, D. Xu, and S. Bethard, "Uarizona at the clef erisk 2017 pilot task: Linear and recurrent models for early depression detection," in *CEUR workshop proceedings*, vol. 1866, NIH Public Access, 2017.

[16] E. Villatoro-Tello, G. Ramırez-de-la Rosa, and H. Jiménez-Salazar, "Uam's participation at clef erisk 2017 task: Towards modelling depressed bloggers," 2017.

[17] I. A. Malam, M. Arziki, M. N. Bellazrak, F. Benamara, A. El Kaidi, B. Es-Saghir, Z. He, M. Housni, V. Moriceau, J. Mothe, *et al.*, "Irit at e-risk," 2017.

[18] M. P. Villegas, D. G. Funez, M. J. G. Ucelay, L. C. Cagnina, and M. L. Errecalde, "Lidic - unsl's participation at erisk 2017: Pilot task on early detection of depression," in *CLEF*, 2017.

[19] M. Trotzek, S. Koitka, and C. M. Friedrich, “Linguistic metadata augmented classifiers at the clef 2017 task for early detection of depression,” 2017.

[20] M. L. Errecalde, M. P. Villegas, D. G. Funez, M. J. G. Ucelay, and L. C. Cagnina, “Temporal variation of terms as concept space for early risk prediction,” 2017.

[21] E. Cambria, S. Poria, D. Hazarika, and K. Kwok, “Senticnet 5: discovering conceptual primitives for sentiment analysis by means of context embeddings,” in *AAAI*, 2018.

[22] P. Domingos, *The master algorithm: How the quest for the ultimate learning machine will remake our world*. Basic Books, 2015.

[23] Urban Dictionary, “Urban dictionary: Caps lock,” 2005.

[24] A. P. L. Monroy and L. V. Pineda, “Atribución de autoría utilizando distintos tipos de características a través de una nueva representación,” 2012.

[25] E. S. Tellez, S. Miranda-Jiménez, M. Graff, D. Moctezuma, O. S. Siordia, and E. A. Villaseñor, “A case study of spanish text transformations for twitter sentiment analysis,” *Expert Systems with Applications*, vol. 81, pp. 457–471, 2017.

[26] Z. S. Harris, “Distributional structure,” *Word*, vol. 10, no. 2-3, pp. 146–162, 1954.

[27] E. Frank, M. Hall, and I. Witten, “The weka workbench,” *Data mining: Practical machine learning tools and techniques. Burlington: Morgan Kaufmann*, 2016.

Curriculum Vitae

Formación

2009-2014
Ingeniería en Sistemas de Software
Universidad Autónoma de Chihuahua

2016-2017
Maestría en Ingeniería en Computación
Universidad Autónoma de Chihuahua

Logros

Participación en el Conference and Labs of the Evaluation Forum (CLEF) 2017 para el laboratorio de Early Risk Detection (eRisk).

Conocimientos

- Lenguajes de programación como Python, C, C++, Java, SQL, BASH, y otros.
- Plataformas de Virualización (KVM, QEMU, VirtualBox).
- Sistemas Unix (Ubuntu), Windows y MacOS.
- Administración de Sistemas Remotos.
- Desarrollo de sistemas de Machine Learning.
- Procesamiento de Lenguaje Natural.

Rasgos

- Creativo.
- Lógico.
- Solucionador de problemas.
- Pragmático.
- Organizado.
- Planificador.
- Autónomo.
- Confiable.
- Rápido Aprendizaje.

Idiomas

- Español (Lengua Materna).
- Inglés (Avanzado).
- Japonés (Certificación JLPT Nivel 5).