Detección de depresión en redes sociales considerando información del perfil de los usuarios

por

José de Jesús Titla Tlatelpa

Tesis sometida como requerimiento parcial para obtener el grado de

Maestro en Ciencias Computacionales

Instituto Nacional de Astrofísica, Óptica y Electrónica

Octubre, 2020

Tonantzintla, Puebla, México

Supervisada por:

Dr. Manuel Montes y Gómez
Coordinación de Ciencias Computacionales
INAOE

Índice General

Índice de Figuras

Índice de Tablas

Agradecimientos

Al Consejo Nacional de Ciencia y Tecnología (CONACYT), por el apoyo otorgado a través de la beca No. 905721. Así como al INAOE por las facilidades proporcionadas durante mis estudios y todos los profesores que transmitieron su conocimiento para alcanzar esta meta.

Al Dr. Manuel Montes y Gómez, profesor y amigo, que además de ser director de esta tesis me ha brindado un apoyo constante tanto personal como profesional. Gracias por sus comentarios y dedicación enfocados al mejoramiento de mi investigación.

A mis sinodales: Dra. Kelsey Alejandra Ramírez Gutiérrez, Dr. Aurelio López López y Dr. Luis Villaseñor Pineda, por sus comentarios y observaciones .

A mi familia por haberme forjado como la persona que soy hoy, brindarme su tiempo y mostrarme el camino para la superación. Muchos de mis logros entre los que se incluye este, se los debo a ustedes.

A la Dra. Sanchez Rinza Bárbara Emma, profesora y amiga que me dió el empujón para lograr una meta mas, también por brindarme su apoyo personal.

A mis compañeros y amigos, por su amistad y apoyo en estos años, y por todos los momentos divertidos y de alegría que pasamos juntos.

Dedicatoria

A mi padres,
quienes en todo momento me brindaron su amor
y palabras de aliento para alcanzar mis sueños.

Para mi hermano, por su comprensión y motivación,
a Rayo, compañero de desveladas y demás momentos,
me acompañó gran parte del camino y me siento feliz por ello.

Para todos los que me apoyaron y creen en mí.

Resumen

La depresión es un trastorno mental que afecta a una cantidad significativa de personas. Estas personas presentan cambios en su estado de ánimo y fatiga durante todo el día, ignorando que tienen signos de depresión. Últimamente, las redes sociales han sido un medio a través del cual las personas comparten sus vivencias y experiencias diarias. Esto abre la oportunidad de usar este contenido para hacer la detección de personas que sufren de depresión a partir de la información de su perfil.

Hacer detección de depresión a partir del contenido en las redes sociales es un problema complejo. Diferentes tipos de personas manifiestan su depresión de manera diferente y usando un tono emocional distinto. Motivados por este problema, en este trabajo de investigación se explora el uso de información del perfil de los usuarios junto a la polaridad emocional de sus publicaciones.

Para el método propuesto en esta tesis se consideran dos enfoques, en el primero se trabaja con la información de perfil considerando dos alternativas para usar esta información: como atributos adicionales en la representación y como clasificadores independientes; además, se utilizan dos representaciones de texto diferentes, una basada en palabras y otra en emociones. En el segundo enfoque se integra la polaridad de las publicaciones realizadas por los usuarios para construir una nueva representación enriquecida con esa información. Como última variante se explora un esquema de clasificación que combine los dos enfoques anteriores.

La evaluación se realiza sobre dos conjuntos de datos, el primero es una colección de Reddit y la segunda de Twitter. Para la evaluación se utiliza el valor F1 en la clase positiva (correspondiente a usuarios que sufren depresión) para medir los resultados obtenidos en la clasificación. Los resultados sugieren que la información

del perfil y de la polaridad del contexto de uso de las palabras son útiles para la tarea en cuestión. Asimismo, los resultados indican que hombres y mujeres, jóvenes y adultos que sufren depresión difieren tanto en los temas como en las emociones que comparten a través de sus comunicaciones.

Cabe mencionar que, según nuestro conocimiento, el método presentado en este trabajo es el primero en considerar la información de perfil junto a la polaridad de las publicaciones para la detección de usuarios que sufren depresión en redes sociales.

Abstract

Depression is a common mental disorder that affects a significant number of people. These people show changes in their mood and fatigue throughout the day, ignoring that they have signs of depression. Ultimately, social media had become a means through which people share their life-situations and daily experiences. This opens up the opportunity to use this content for the detection of people who suffer from depression from their profile information.

Detecting depression from social media content is a complex problem. Different types of people manifest their depression differently and using a different emotional tone. Motivated by this problem, in this research work we explore the use of profile information together with the emotional polarity of their publications.

For the proposed method in this thesis, two approaches are considered, in the first one we work with the profile information considering two alternatives to use this information: as additional attributes in the representation and as specific classifiers; In addition, two different text representations are used, one based on words and the other based on emotions. In the second approach, the polarities of the publications posted by the users are integrated to build a new enriched representation with that information. As a last variant, a classification scheme that combines the two previous approaches is explored.

The evaluation is performed on two datasets, the first is a collection on Reddit and the second on Twitter. For the evaluation, the F1 value over the positive class (corresponding to users who suffer from depression) is used to measure the results obtained in the classification. The results suggest that the information of the profile and the polarity of the context of use of the words are useful for the task in question.

Likewise, the results indicate that men and women, young and adult people who suffer from depression differ both in the topics and in the emotions they share through their communications.

It is worth mentioning that, according to our knowledge, the method presented in this work is the first to consider the profile information together with the polarity of emotions for the detection of users suffering from depression in social networks.

Capítulo 1

Introducción

1.1. Motivación

La salud mental es parte integral de la salud y el bienestar. La organización mundial de la salud (OMS) (O.M.S, 2013) concibe la salud mental como un estado de completo bienestar físico, mental y social, y no solamente la ausencia de afecciones o enfermedades. Los trastornos mentales tienen una alta prevalencia y son factores que contribuyen de manera importante a la morbilidad, la discapacidad, las lesiones y la mortalidad prematura, además de aumentar el riesgo de padecer otras condiciones de salud. Muchos factores de riesgo son comunes a los trastornos mentales y a otras enfermedades no transmisibles, entre ellos: bajo nivel socioeconómico, consumo de alcohol y el estrés; *no hay salud sin salud mental* (Prince et al., 2007). Entre los trastornos mentales más frecuentes que afectan la salud mental de las personas se encuentra la depresión. Este trastorno puede llegar a ser crónico o recurrente y dificultar el desempeño en el trabajo o la escuela, así como la capacidad para afrontar la vida diaria.

La depresión es un problema de salud que afecta gravemente a nuestra sociedad, es más que solo sentirse triste o tener un mal día. Millones de personas sufren de depresión cada año y solo algunas reciben un tratamiento adecuado; a pesar de que la depresión es una de las principales contribuyentes a la carga global de enfermedades, este no es uno de sus resultados más trágicos ya que cobra la vida del 15-20 % de todos sus pacientes a través del suicidio (Sadeque et al., 2016). De acuerdo con la Organización Mundial de la Salud (OMS), cerca de 800,000 personas se suicidan cada año, siendo esta la segunda causa de muerte en el grupo etario de

15 a 29 años. Aunque hay tratamientos eficaces para la depresión, más de la mitad de los afectados en todo el mundo (y más del 90 % en muchos países) no los reciben ya sea por falta de recursos o de personal sanitario capacitado[1]. Por esta razón la Organización Panamericana de la Salud (OPS) junto a la Organización Mundial de la Salud (OMS) adoptaron un plan de acción para el periodo 2015-2020 que tiene entre sus objetivos promover el bienestar mental para reducir la morbilidad, discapacidad y mortalidad (OPS and OMS, 2014).

Para poder ofrecer ayuda a quienes la necesitan, primero deben identificarse. En los servicios de salud mental, la detección de depresión clínica realizada por psicólogos se basa en herramientas estandarizadas para medir su evaluación. Estas herramientas incluyen el uso de cuestionarios y entrevistas enfocadas en el comportamiento de los pacientes. Un reto que se presenta en el diagnóstico de la depresión es que al contarse solo con lo reportado por el paciente, por vergüenza o prejuicios no reportan todos los síntomas. Esto supone una barrera debido a que con una evaluación errónea las personas con depresión no suelen ser diagnosticadas correctamente. En la población general hay evidencia de que la tasa de prevalencia para depresión es mayor en mujeres que en hombres; los hombres que no buscan ayuda para problemas de salud mental pueden experimentar sufrimiento innecesario que afecta su propio bienestar y el de los demás. Aunque las mujeres tienen el doble de probabilidades de ser diagnosticadas con depresión, los hombres tienen cuatro veces más probabilidades de morir por suicidio, lo que sugiere que muchos hombres tienen problemas de salud mental no diagnosticados o no los reportan (Call and Shafer, 2018).

Hoy en día internet es considerado una herramienta de comunicación y una fuente de información, dependiendo del enfoque con el que se use. Es un medio ubicuo tanto para los negocios como para el entretenimiento, pero podría decirse que ha tenido el mayor impacto como medio de comunicación interpersonal. Últimamente, el uso de las distintas plataformas y redes sociales se ha vuelto un medio a través del cual las personas tienden a expresar sus emociones y opiniones. Las publicaciones en estos sitios se realizan en un entorno natural y en el curso de actividades y acontecimientos diarios. Como tal, las redes sociales proporcionan un medio para capturar atributos de comportamiento que son relevantes para el pensamiento, el estado de ánimo, la

[1] https://www.who.int/news-room/fact-sheets/detail/depression

comunicación, las actividades y la socialización de un individuo (Choudhury et al., 2013). Debido a la gran presencia de actividad en estos medios se ha generado una gran cantidad de datos disponibles para realizar la observación de diversos usuarios en distintas plataformas. Aprovechando las pistas que los usuarios dejan sobre su comportamiento se han usado técnicas de clasificación de textos para la detección de este trastorno mental.

1.2. Problemática

En los últimos años la detección de depresión se ha tratado como un problema de clasificación supervisada basado en un gran cantidad de información textual para este proceso. Mucha de esta ha sido compartida por usuarios en redes sociales, por ejemplo Reddit. Técnicas de Procesamiento del Lenguaje Natural han sido aplicadas a esta información textual para examinar el impacto de las redes sociales en la detección de personas con depresión.

La detección automática de depresión en las redes sociales se ha logrado mediante la creación de modelos predictivos, que utilizan características o variables extraídas de los usuarios. Las características son tratadas como variables independientes en un algoritmo (Máquinas de Vectores de Soporte o Árboles de Decisión, por ejemplo). Una estrategia común para abordar este problema es utilizar representaciones basadas en palabras, desde una simple Bolsa de Palabras hasta un histograma de categorías de LIWC (Linguistic Inquiry and Word Count) o incluso una Red Neuronal alimentada por Word Embeddings (Wolohan et al., 2018; Resnik, Garron, y Resnik, 2013; Yates, Cohan, y Goharian, 2017; Trotzek, Koitka, y Friedrich, 2020). La disponibilidad de los mensajes de redes sociales ha permitido extraer especialmente información de salud que hace posible rastrear enfermedades, síntomas y medicamentos. Una desventaja en este enfoque es que se ha tratado por igual a todos los usuarios.

Se sabe de trabajos de psicología que hombres y mujeres, jóvenes y adultos manifiestan la depresión de distinta manera y usando un tono emocional distinto. Soportados por esta idea, en esta tesis se explora el uso de información de perfil de los usuarios junto a la polaridad emocional de sus publicaciones. Varios trabajos que han

estudiado la relación entre los atributos del perfil de los pacientes han encontrado claras diferencias entre los diferentes tipos de personas, particularmente entre hombres y mujeres. Más recientemente, también desde una perspectiva psicológica, algunos trabajos han analizado la relación entre el uso de las redes sociales y la depresión, así como las diferencias en su uso por parte de usuarios deprimidos femeninos y masculinos (Angst et al., 2002; Nolen-Hoeksema, 2001; Call and Shafer, 2018; McCrae, Gettings, y Purssell, 2017).

A pesar de que se han realizado un gran número de trabajos en el área, aún no se logra explicar completamente la diferencia en el comportamiento de depresión entre los diferentes grupos de personas. A continuación se presentan los objetivos propuestos para la presente investigación basados en la problemática expuesta anteriormente.

1.3. Objetivos

1.3.1. Objetivo general

Diseñar e implementar un método para la detección de usuarios de redes sociales que sufren depresión considerando información de su perfil así como los sentimientos y emociones en sus mensajes.

1.3.2. Objetivos específicos

- Proponer un método para la clasificación de usuarios con depresión que integre aspectos de su perfil, particularmente su género y edad.

- Diseñar una nueva representación de los usuarios de redes sociales que distinga la información positiva y negativa de sus mensajes.

- Evaluar, comparar y analizar el desempeño del método y representación propuestos en la detección de usuarios con depresión de al menos dos tipos

diferentes de redes sociales.

1.4. Organización de la tesis

Esta tesis está conformada por los siguientes capítulos:

- Capítulo 2. Marco teórico: En este capítulo se presentan los conceptos relacionados con la clasificación de textos y los algoritmos de aprendizaje que son utilizados en el trabajo, además de conceptos para describir y entender el método propuesto.

- Capítulo 3. Trabajo relacionado: Se presenta el trabajo más relevante relacionado con el tema de tesis y el método propuesto. Se hace una revisión de la detección de depresión enfocada en la clasificación de textos, se habla de foros de evaluación en esta tarea además de trabajos que usan distintas representaciones y tipos de atributos.

- Capítulo 4. Métodos y representaciones propuestas: Se explican detalladamente los métodos propuestos y las representaciones junto a los atributos que serán utilizados en el proceso de clasificación. Para dar cumplimiento a los objetivos, se describe como intervienen los atributos de perfil y la polaridad de sus publicaciones.

- Capítulo 5. Experimentos y resultados: Se exponen los experimentos derivados de los métodos propuestos y la evaluación de los mismos en el conjunto de datos de Reddit y Twitter. Se incluye un análisis de los resultados obtenidos por el mejor método.

- Capítulo 6. Conclusiones y trabajo futuro: Se exponen las conclusiones obtenidas y el trabajo futuro.

Capítulo 2
Marco Teórico

En el presente capítulo se describen los conceptos principales de esta tesis. Primero, relacionados con la clasificación de textos y algoritmos de clasificación, luego con el perfilado de autor y análisis de sentimientos, temas que fueron usados durante el desarrollo de esta investigación.

2.1. Clasificación de textos

La clasificación de textos es el proceso de asignar documentos a categorías predefinidas basándose en su contenido. La clasificación es binaria cuando solo se tienen dos clases, si se tienen más de dos clases se denomina multiclase. En algunos casos se puede pertenecer a más de una clase a la vez por lo que se dice entonces que es multietiqueta.

El enfoque principal usado en este trabajo es el de aprendizaje supervisado. Es decir, dado un conjunto de documentos etiquetados $\mathcal{D} = \{(\mathbf{d}_i, y_i)\}_{i \in 1,\ldots,|\mathcal{D}|}$, formado por pares de representaciones de documentos $\mathbf{d}_i = \langle a_1, a_2, \ldots, a_n \rangle$ y etiquetas $y_i \in \{0, 1\}$, donde 0 y 1 indican la clase negativa y positiva respectivamente, el objetivo es encontrar una función $f : \mathbb{R}^n \to \{0, 1\}$ para clasificar documentos no etiquetados mapeando documentos a categorías, p.e., $y_j = f(\mathbf{u}_j)$.

2.1.1. Representaciones de textos

Generalmente, la entrada para un clasificador debe ser un documento representado como un vector de características. El objetivo es representar de forma

numérica los documentos para hacerlos matemáticamente computables (Yan, 2009). A continuación se describen algunas representaciones usadas en la clasificación de textos.

Bolsa de palabras

Una de las técnicas más utilizadas para representar el texto es la Bolsa de Palabras (BoW, por sus siglas en inglés). En esta representación los documentos son considerados un conjunto de palabras, sin que interese el orden en que están ubicadas dentro del texto, aunque si la frecuencia con que ocurren.

En el modelo BoW se transforman/codifican los documentos en una representación vectorial donde cada entrada indica la presencia/ausencia de un término del vocabulario. La representación tiene las siguientes características:

- El vocabulario está compuesto por todas las palabras diferentes en los documentos, éste es considerado como la base para la representación del vector
- Los documentos son representados por el conjunto de palabras que contienen (vocabulario).
- El orden de las palabras no es capturado por la representación.
- Al no ser capturado el orden de las palabras, no hay intentos por entender el contenido.

En la tabla 2.1 se muestra una representación BoW con un conjunto de documentos $D = \{d_1, d_2, ..., d_i\}$ y un vocabulario $V = \{t_1, t_2, t_j, ..., t_{|V|}\}$. Donde cada documento puede ser representado en el espacio vectorial con una dimensión $|V|$ y cada entrada del documento d_i tiene un peso W.

El peso $W_{i,j}$ corresponde al término t_j en el documento d_i, cuando el término no se encuentra en el documento entonces $W_{i,j} = 0$. Los distintos tipos de peso se explican más adelante.

	t_1	t_j	...	$t_{\|V\|}$
d_1	$W_{1,1}$			
d_2	$W_{2,1}$			
...				
d_i		$W_{i,j}$		

Tabla 2.1: Representación de una BoW.

N-gramas

Los n-gramas son ampliamente usados en tareas de procesamiento del lenguaje natural, son un conjunto de secuencias de palabras en una ventana dada. La n determina el tamaño de la ventana con el que se generará la secuencia de palabras, mientras más grande sea n más elementos del texto serán capturados. Los elementos capturados por los n-gramas comúnmente son caracteres y palabras, las tablas 2.2 y 2.3 muestran ejemplos de n-gramas.

Texto	Lorem ipsum dolor sit amet consectetur adipiscing elit
n = 1	Lorem, ipsum, dolor, sit, amet, consectetur, adipiscing, elit
n = 2	Lorem ipsum, ipsum dolor, dolor sit, sit amet, amet consectetur, consectetur adipiscing, adipiscing elit
n = 3	Lorem ipsum dolor, ipsum dolor sit, dolor sit amet, sit amet consectetur, amet consectetur adipiscing, consectetur adipiscing elit

Tabla 2.2: Ejemplo de n-gramas de palabras.

De esta forma se puede capturar algo de orden en las palabras pero a cambio de que la dimensionalidad de la representación crezca. Es decir, si la BoW se representa con n-gramas de palabras con n(1-2) el vocabulario será de tamaño 15 (8 uni-gramas

+ 7 bi-gramas).

Texto	Lorem_ipsum
n = 3	Lor, ore, rem, em_ ,m_i, _ip, ips, psu, sum
n = 4	Lore, orem, rem_, em_i, m_ip, _ips, ipsu, psum
n = 5	Lorem, orem_, rem_i, em_ip, m_ips, _ipsu, ipsum

Tabla 2.3: Ejemplo de n-gramas de caracteres.

Para conocer el número de n-gramas que podemos encontrar en un texto calculamos:

$$Ngramas_X = P - (N - 1) \tag{2.1.1}$$

Donde P es el número de palabras en el texto y N el tamaño de la secuencia. En el ejemplo de n-gramas de palabras el texto tiene 8 palabras, para n = 3 el resultado sería 6 secuencias de tri-gramas.

Etiquetado POS

Las palabras pertenecen a una categoría léxica (gramatical). Estas categorías nos indican el papel que las palabras desempeñan en una oración, cuando esta categoría es mayor a uno, se dice que la palabra es ambigua, esto puede pasar en la mayoría de los idiomas. Para determinar su papel en una oración se recurre al contexto, las categorías que rodean la palabra, representando esto un papel fundamental. Por ejemplo, considerando las oraciones *voy a volar mañana hacia Chile, hoy noté el volar del colibrí*; la palabra *volar* puede asumir la función de verbo o sustantivo. Así, las categorías gramaticales resultan de utilidad por la información que dan acerca de la palabra y su alrededor. Saber si una palabra es un verbo o un sustantivo nos ayuda a interpretarlas y sirve para encontrar su etiqueta más apropiada.

Las partes de la oración (POS, por sus siglas en inglés) explican cómo es usada cada palabra en una oración, hacer etiquetado POS es la tarea de asignar a cada palabra una etiqueta con la función que cumple. El idioma inglés cuenta con

un conjunto de etiquetas estándar conocido como el conjunto de etiquetas Penn Treebank Project. La tabla 2.4 muestra algunas etiquetas.

Número	**Etiqueta**	**Descripción**
1	CC	Conjunción coordinada
2	CD	Número cardinal
3	DT	Determinante
...	**...**	**...**
34	WP	Wh-pronombre
35	WP$	Wh-pronombre posesivo
36	WRB	Wh-adverbio

Tabla 2.4: Lista de algunas etiquetas POS usadas en el Proyecto Penn Treebank.

Con el etiquetado POS se puede hacer una representación similar a la BoW en donde el vocabulario pasa a ser representado por un par palabra-etiqueta. La tabla 2.5 muestra dos ejemplos de oraciones con etiquetado POS y la tabla 2.6 muestra la representación.

Entrada	Oración con etiquetado POS
d_1	The/DT book/NN is/VBZ on/IN the/DT table/NN
d_2	I/PRP love/VBP to/TO play/VB with/IN my/PRP$ dog/NN

Tabla 2.5: Ejemplo de oraciones con etiquetado POS.

El peso $W_{i,j}$ corresponde al par término,etiqueta $w_k t_j$ en el documento d_i.

	w_1t_1	**(refuse,VBP)**	**(refuse,NN)**	**...**	w_kt_j
d_1	$W_{1,1}$				
d_2	$W_{2,1}$				
...					
d_i					$W_{i,j}$

Tabla 2.6: Representación usando etiquetado POS.

2.1.2. Pesado de los atributos

En la representación de los textos cada atributo puede aportar información para la clasificación. Aunque no todos los atributos son útiles para describir un documento, hacer que su vector aporte peso a la representación general es importante ya que de eso dependerá el rendimiento del clasificador. El peso con el que se ponderará cada atributo según su documento varía dependiendo de la situación y/o el problema a abordar. A continuación se presentan tres de las variantes más usadas.

Booleano

Es la ponderación más simple que se puede asignar, en este pesado solo se considera si el término aparece o no en el documento. Para aplicar este pesado solo basta si el término t_j aparece en el documento d_i y asignar $w_{i,j} = 1$ si lo contiene, de otra forma se asigna un 0.

Frecuencia del Término

El pesado por frecuencia del término (TF, por sus siglas en inglés) mide que tan frecuente es la ocurrencia de cada palabra del vocabulario en un documento. Este tipo de pesado no hace distinción en aquellas palabras que son muy frecuentes en cualquier documento, y las que son muy frecuentes en unos documentos en particular. Las palabras que sean muy frecuentes en todos los documentos tendrán un tf alto, y no servirían para distinguir a los documentos de acuerdo a lo que tratan. De forma

simple la ecuación 2.1.2 define el pesado tf:

$$tf_{i,j} = f_{i,j} \tag{2.1.2}$$

Donde la frecuencia $f_{i,j}$ es el número de ocurrencias del término t_j en el documento d_i. Si la frecuencia $f_{i,j}$ es normalizada por la longitud del documento la ecuación es transforma a:

$$tf_{i,j} = \frac{f_{i,j}}{|d_i|} \tag{2.1.3}$$

Donde $|d_i|$ es el número de términos en dicho documento.

Frecuencia del Término - Frecuencia Inversa del Documento

La importancia que tienen ciertas palabras de un documento en comparación con todos los documentos puede ser medida con la Frecuencia del Término - Frecuencia Inversa del Documento (TF-IDF, por sus siglas en inglés). Éste es otro tipo de pesado utilizado en las representaciones del texto y combina dos elementos.

IDF mide que tan importante es un término, compara el número de todos los documentos con el número de documentos que contienen el término que se está analizando. IDF es calculado de la forma siguiente:

$$idf_j = log\frac{N}{df_j} \tag{2.1.4}$$

Donde N es el número de documentos y df_t es la cantidad de documentos que contienen el término j. Así la fórmula de Frecuencia del Término - Frecuencia Inversa del Documento es mostrada en la ecuación.

$$tfidf_{i,j} = tf_{i,j} * idf_j \tag{2.1.5}$$

2.1.3. Algoritmos de aprendizaje

La clasificación es un enfoque de aprendizaje computacional en el que un algoritmo aprende de los datos proporcionados y luego asigna un valor de salida a una muestra de datos. En esta sección se explican algunos algoritmos que fueron utilizados en los experimentos de este trabajo.

Máquinas de Soporte Vectorial

Las máquinas de soporte vectorial (SVM, por sus siglas en inglés) fueron introducidas por (Cortes and Vapnik, 1995), este algoritmo de aprendizaje puede ser utilizado para clasificación binaria, multiclase y regresión. Las SVM encuentran un hiperplano h que separa los ejemplos de ambas clases con un *margen máximo*. El hiperplano es construido con los ejemplos frontera en el margen los cuales se denominan *vectores de soporte*.

Suponiendo que tenemos un conjunto D de documentos etiquetados para el entrenamiento $(d_i, y_i)_{i\in 1,\ldots,|D|}$ donde $d_i \in \mathbb{R}^n$ y pertenece a una de las dos posibles clases en $y_i \in \{-1, 1\}$. Asumimos que los datos son linealmente separables, es decir, podemos dibujar una línea en una gráfica separando a d_1 de d_2. El hiperplano puede describirse por $\mathbf{w^*x+b = 0}$, los **vectores de soporte** son los ejemplos más cercanos al hiperplano de separación. El objetivo de las máquinas de vectores de soporte (SVM) es orientar este hiperplano de manera que esté lo más lejos posible de los miembros más cercanos de ambas clases.

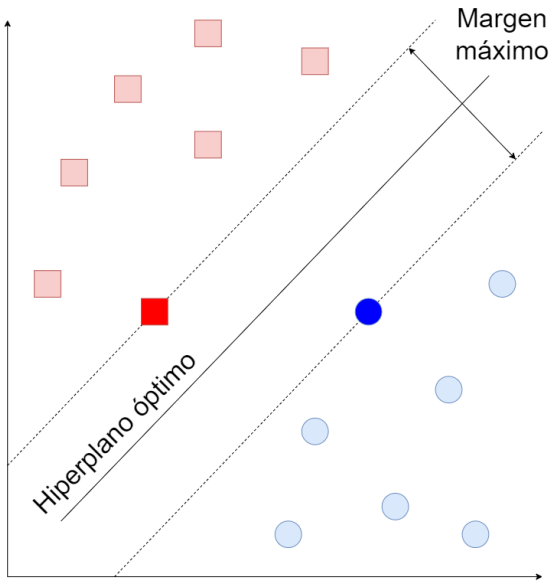


Figura 2.1: Hiperplano y margen máximo

Árboles de Decisión

Un árbol de decisión es una herramienta predictiva que se construye siguiendo la estrategia "divide y vencerás". Gráficamente es similar a un árbol con nodos internos que representan el atributo y nodos terminales que representan la clase. Las aristas representan el camino entre dos nodos y contiene el valor de la decisión.

Los árboles de decisión implican un proceso de división interno para tomar una decisión. Al principio se selecciona el atributo más importante para ser la raíz. En cada nodo el conjunto se divide y se construye un nuevo árbol de decisión. Para determinar que atributo debe elegirse a continuación, se selecciona en cada nivel el más discriminativo. Uno de los algoritmos más conocidos es el ID3.

Algoritmo 1 *ID3*

Entrada: I: instancias, T: atributo_objetivo, A: atributos

Si todas las instancias tienen a T perteneciente a la misma clase c **entonces**
    Regresar el árbol con el nodo raíz con la etiqueta c
Si A está vacío **entonces**
    Regresar el árbol con el nodo raíz con la etiqueta más común de T en I
si no
    a = el mejor atributo en A
    Para cada valor v de a **hacer:**
        sea I(v) el subconjunto de instancias con valor v para a
        Si $I(v)$ está vacío **entonces**
            Regresar un nodo con el valor más común de T en I
        si no
            Regresar ID3(I(v), T, A-a)

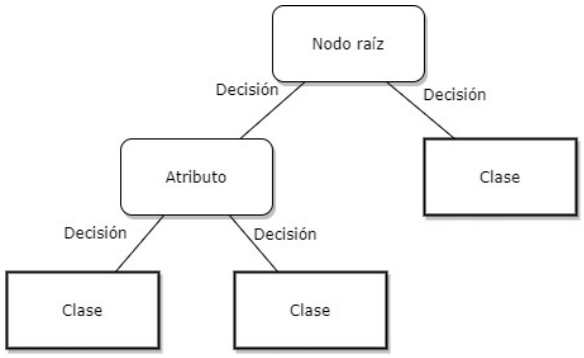


Figura 2.2: Árbol de decisión

Algoritmos de ensamble

En las tareas de clasificación no existe un algoritmo de aprendizaje dominante. Se tienen que probar diferentes clasificadores para conocer cuál es el que da el mejor resultado, y así escoger con cual tendremos mejor rendimiento según el conjunto de datos. Un enfoque de ensamble de clasificadores, consiste en combinar más de un modelo de clasificación para producir un mejor rendimiento predictivo que utilizar uno solo. El principio se basa en que un grupo de clasificadores débiles puede volverse fuerte. Es importante que entre los miembros del ensamble exista diversidad en las predicciones.

Bagging (Contracción de *Boostrap Aggregating*, por sus siglas en inglés) es un método de ensamble simple y poderoso que genera clasificadores con varias muestras de los ejemplos. Además funciona especialmente para algoritmos de aprendizaje que cambian su estructura con los ejemplos (Árboles de Decisión, por ejemplo). Cuando se hace Bagging el objetivo es reducir la varianza de un árbol de decisión creando varios subconjuntos de datos a partir de una muestra de entrenamiento elegida al azar con reemplazamiento. El resultado final es un conjunto de diferentes modelos con sus respectivas etiquetas (que es más robusto que un solo árbol de decisión). Para obtener la predicción final se utiliza la clase que fue obtenida más veces, es decir, se hace un método de votación simple.

Algoritmo 2 *Bagging*

Entrada: T: número de iteraciones, S: conjunto de datos etiquetados
Salida: C_t: clasificadores con $t = 1$, ..., T

1: **Para** $t = 1$ *hasta* T **hacer:**
2: S_t = submuestra de S con reemplazamiento
3: C_t = se construye clasificador con S_t
4: $t+ = 1$
5: **Fin para**

La imagen 2.3 muestra de forma gráfica como funciona el método de ensamble Bagging presentado en el algoritmo 2.

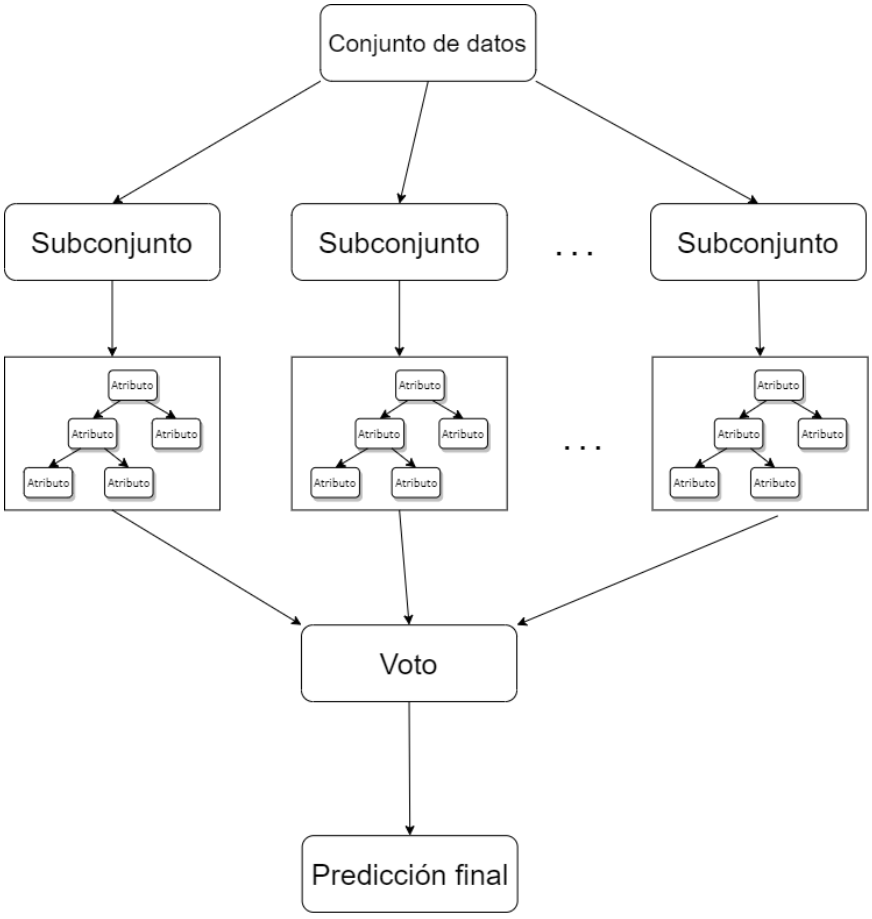


Figura 2.3: Ensamble con Bagging

2.1.4. Métricas de evaluación

Debido a que trabajamos con tareas de clasificación es importante medir el desempeño de cada clasificador y cada representación con el fin de poder hacer una comparación entre cada una de ellas. A continuación se describe precisión, recuerdo y medida F1 (Olson and Delen, 2008), métricas utilizadas en el presente trabajo.

Precisión

La precisión es una medida que nos indica la proporción de documentos clasificados correctamente para una clase respecto a las predicciones totales de la clase.

$$P(C_i) = \frac{TP}{TP + FP} \tag{2.1.6}$$

True Positives: son casos donde la clase actual es *True* y la predicción también es *True*. **False Positives**: son casos donde la clase actual es *False* y la predicción es *True*.

Recuerdo

El recuerdo es una medida que indica la proporción de los documentos positivos clasificados correctamente, entre el número total de documentos positivos reales.

$$R(C_i) = \frac{TP}{TP + FN} \tag{2.1.7}$$

True Positives: son casos donde la clase actual es *True* y la predicción también es *True*. **False Negatives**: son casos donde la clase actual es *True* y la predicción es *False*.

Medida F1

En algunos problemas puede darse el caso de querer dar prioridad a tener mejor recuerdo aunque eso puede llevar a tener poca precisión o viceversa. Por lo tanto, es mejor si podemos obtener una única medida que represente tanto a la precisión como al recuerdo. Podemos combinar ambos aplicando una media armónica.

$$F1(C_i) = 2 * \frac{P(C_i) * R(C_i)}{P(C_i) + R(C_i)} \tag{2.1.8}$$

En la ecuación 2.1.8 P es la precisión y R el recuerdo de la clase C_i.

2.2. Perfilado de autor y análisis de sentimientos

La tarea de extraer toda la información posible del autor de un documento es muy conocida en Procesamiento del Lenguaje Natural. El **Perfilado de Autor** (AP, por sus siglas en inglés) tiene como objetivo analizar el texto escrito para determinar atributos generales o demográficos del autor. La tarea de AP también ha sido considerada con un enfoque en la clasificación de textos; en este escenario el foco de la investigación es seleccionar las mejores características textuales para modelar el perfil del usuario en base a lo que escribe. Esta tarea se basa en la idea de que personas con características de perfil en común también comparten similitudes lingüísticas, esto en parte por su ambiente cultural o social. Existen trabajos en psicología que han motivado estudios en computación para el AP; estos trabajos han estudiado cómo se comparte el lenguaje entre las personas, y han establecido una relación entre el uso del lenguaje y los rasgos de personalidad. (Pennebaker, Mehl, y Niederhoffer, 2003).

Hay dos tipos de características textuales que han sido clave en la tarea, las *características basadas en contenido* (n-gramas y tópicos, por ejemplo) y las *características basadas en el estilo* (signos de puntuación y emoticones, por ejemplo) (Ortega-Mendoza et al., 2018). Herramientas accesibles para predecir algunas variables demográficas pueden mejorar substancialmente la utilidad de las redes sociales para diversas aplicaciones. El lexicón pesado diseñado por (Sap et al., 2014) es un recurso léxico que considera palabras asociadas a géneros y edades. La predicción de género y edad se obtiene aplicando una suma de todas las frecuencias relativas de palabras pesadas sobre un documento; la edad del usuario corresponde al resultado de la suma, mientras que el género se indica con el signo de la suma, un resultado positivo indica un usuario femenino y un resultado negativo un usuario masculino.

El **Análisis de Sentimientos** (Sentiment Analysis, en inglés) es el proceso de determinar el tono emocional detrás de una serie de palabras, es usado para obtener entendimiento de actitudes, opiniones y emociones expresadas en un texto (Hassan et al., 2017). Debido a que los datos suelen contener ruido, hacer este análisis no es una tarea trivial. Por ejemplo, el usar palabras o abreviaturas de algunas expresiones puede hacer que la evaluación del lenguaje no pueda determinar correctamente si la

expresión es *positiva, negativa o neutral*. Gran cantidad de información disponible en redes sociales no tiene una estructura definida. La habilidad de obtener los sentimientos detrás de opiniones ha permitido que las aplicaciones sean amplias y poderosas, esto debido a que varias organizaciones adoptan este enfoque para darle un uso práctico.

Una herramienta para hacer análisis de sentimientos es SentiWordNet, un lexicón de opiniones derivado de la base de datos WordNet (Baccianella, Esuli, y Sebastiani, 2010). Cada término se asocia con tres puntuaciones indicando objetividad e información positiva y negativa. Dado un documento, las oraciones se dividen en términos a consultar en SentiWordNet; la polaridad corresponde al signo de la suma, un signo positivo indica positividad y un signo negativo indica negatividad.

Capítulo 3

Trabajo Relacionado

En el presente capítulo se expone el trabajo relacionado más relevante para el desarrollo de esta tesis de acuerdo al enfoque propuesto. Se incluyen trabajos en el área de computación que abordan la detección de depresión como una tarea de clasificación de textos. En la sección 3.1 se presentan las diferentes soluciones que se le han dado a la tarea usando solamente información textual. Posteriormente en la sección 3.2 se muestran enfoques que han considerado información adicional a la extraída de los textos de los usuarios. Por último la sección 3.4 presenta una discusión entre el método propuesto y el trabajo que ya se encuentra desarrollado, introduciendo la idea de nuestro trabajo.

3.1. Detección de depresión con enfoque de clasificación de textos

Un medio eficaz para examinar el desorden depresivo es a través del análisis del lenguaje, el uso del lenguaje puede estar vinculado a información importante sobre el comportamiento de las personas y sus estados psicológicos (Pennebaker, Mehl, y Niederhoffer, 2003). En los últimos años la detección de depresión se ha abordado a través del aprendizaje automático considerándola como una tarea de clasificación de textos.

3.1.1. Criterios para la recolección de datos

Obtener el conjunto de datos con el que se va a trabajar es importante para la tarea, de él depende el rendimiento y el comportamiento que se obtenga en la clasificación. Las redes sociales, específicamente para la detección de depresión, han sido medios que han permitido la adquisición de estos datos (Guntuku et al., 2017). En los siguientes párrafos se presentan algunos enfoques que han sido utilizados para la creación de conjuntos de datos que han sido utilizados en esta tarea.

Contar con psicólogos profesionales que ayuden a hacer una validación de los datos según su criterio ha sido de gran utilidad en el enfoque de ***aplicación de cuestionarios***. Estos cuestionarios tienen un grado alto de validez y fiabilidad siendo solamente superados por las entrevistas clínicas (Guntuku et al., 2017). Los cuestionarios consisten en una serie de preguntas sobre las actividades cotidianas, presencia de algún síntoma y comportamiento en algunas situaciones. Las herramientas más usadas son la escala de depresión del centro de estudios epidemiológicos (CES-D, por sus siglas en inglés) y los síntomas definidos por el manual diagnóstico y estadístico de los trastornos mentales (DSM-5, por sus siglas en inglés). Trabajos que han usado este enfoque para la construcción de su conjunto de datos son los de (Choudhury et al., 2013; Schwartz et al., 2014; Reece et al., 2016). Sin embargo, la cantidad de recursos y el tiempo necesario para logra el objetivo es una desventaja de este enfoque.

Las personas usan cada vez más las redes sociales para compartir pensamientos y opiniones con sus conocidos. Como tal, las redes sociales proporcionan un medio para capturar atributos de comportamiento relevantes de un individuo (Choudhury et al., 2013). Utilizar un enfoque basado en la ***auto-declaración del diagnóstico clínico de depresión*** consiste en usar datos de fuentes públicas como las redes sociales en busca de frases específicas. Varios trabajos reportan haber utilizado frases como *I'm diagnosed with depression, I have been diagnosed with depression, I was diagnosed with depression* (Coppersmith et al., 2015; Yates, Cohan, y Goharian, 2017; Shen et al., 2017). Una vez que se identifican los usuarios se procede a descargar el historial de sus publicaciones, (Losada and Crestani, 2016) adaptaron este enfoque para la construcción del conjunto de datos utilizado en eRisk. Aunque este enfoque es una forma fácil para construir el conjunto de datos, tiene como desventaja incluir

usuarios que realmente no padezcan la enfermedad.

Las redes sociales no son el único lugar en el que personas con alguna enfermedad mental acuden para expresarse o buscar ayuda. Los foros en línea y sitios web de discusión son la segunda fuente de datos pública a la que se puede acudir respecto a estos temas. Enfocarse en la ***pertenencia a foros y grupos de soporte web*** consiste en buscar tableros o grupos por tema y conseguir el historial de publicaciones de usuarios auto-declarados con depresión. (Losada and Crestani, 2016) adoptaron este enfoque en Reddit, al ser una plataforma donde los contenidos están organizados por áreas de interés *subreddits* y se pueden encontrar temas referentes a diversas condiciones médicas. Algo similar sucede en los grupos de apoyo en HealthBoards donde los usuarios pueden iniciar un hilo o responder al hilo iniciado por otro usuario; *7 Cups of Tea* es una comunidad de soporte para el estrés emocional en donde puedes comunicarte con terapistas expertos e incluso volverte en un oyente voluntario (Sadeque et al., 2016; Losada and Crestani, 2016; Loveys et al., 2018).

El último enfoque que se abordará consiste en la ***búsqueda por palabra clave***. Se usa para buscar usuarios o publicaciones de usuarios que pertenezcan a un grupo en específico y posteriormente obtener su historial de publicaciones. En el trabajo de (Jamil et al., 2017) se usa un conjunto de palabras claves, como lo pueden ser síntomas, para asegurar que los usuarios pertenezcan al grupo de depresión. Por otro lado (Hiraga, 2017) busca entradas en blogs ordenados por popularidad que incluyan la palabra "*depression*", posteriormente toma en cuenta otros criterios para formar el grupo final de usuarios deprimidos.

3.1.2. Representaciones y enfoques comunes

Una manera de abordar la detección de depresión es tratarla como un problema estándar de clasificación de textos. De esta forma, los usuarios representan la clase a discriminar. Por lo tanto, distintos enfoques tradicionales pueden ser empleados. Existen trabajos recientes enfocados en construir modelos para intentar detectar enfermedades mentales; el análisis automatizado de la información en redes sociales proporciona métodos para la detección temprana (Guntuku et al., 2017).

BoW y N-gramas de palabras

Las palabras usadas en los textos reflejan los temas y la forma en que las personas se expresan. Representarlas mediante una bolsa de palabras (BoW, por sus siglas en inglés) para clasificarlos con una máquina de soporte vectorial (SVM, por sus siglas en inglés) ha sido un enfoque utilizado para detección de depresión. Este tipo de representación se construye con características extraídas de textos representando a los documento con un vector de características, asignando valores a cada uno. Por ejemplo, en una BoW construida con vocabulario como atributos, los valores en cada vector puede ser booleano (1 o 0) o un valor entero positivo que indique la frecuencia de una palabra.

En un análisis sobre algunos clasificadores y el tipo de representación que más se usa para hacer líneas base (Wang and Manning, 2012) opinan que usar uni-gramas o bi-gramas puede beneficiar en la clasificación dependiendo de la tarea. Por ejemplo, usar bi-gramas puede mejorar el rendimiento obtenido.

Las representaciones de Bolsas de Palabra han sido muy utilizadas para hacer detección de depresión a nivel-usuario y a nivel-documento. Sin embargo, uno de los problemas con este tipo de representaciones es la alta dimensionalidad, esto requiere un gran número de recursos computacionales para llevar a cabo la clasificación. Esto podría ser aún más costoso si además se desea comparar la utilidad de usar uni-gramas y n-gramas tanto de palabras como de caracteres en ambos niveles. Por ejemplo, el trabajo de (Hiraga, 2017) donde se llega a utilizar n-gramas con valor de 1-10.

Por otra parte, un segundo problema en estas representaciones ocurre en escenarios donde el vocabulario es amplio, pero los datos y las clases se encuentran desbalanceados. A causa de esto la representación tiende a favorecer la clase mayoritaria, aunque el documento a clasificar pertenezca a la otra clase. En consecuencia, se tendría que aplicar técnicas de re-muestro (submuestreo y sobremuestreo) a los datos para balancear ambas clases. (Jamil et al., 2017) se enfrentó a este problema y en su trabajo concluyó que lo mejor es hacer un submuestreo.

LIWC y LDA

El uso de recursos para el análisis de textos como LIWC (Linguistic Inquiry and Word Count) ha sido factible en los análisis temáticos. LIWC es capaz de calcular cómo las personas usan diferentes categorías de palabras a través de su forma de expresión; esto está basado en la idea de que las palabras que una persona usa revela información acerca de su estado psicológico. LIWC ha sido usado extensamente para distintos propósitos. En (Sadeque et al., 2016) identifican las clases psicolingüísticas que están más asociadas a los usuarios en un foro. La idea es encontrar cambios a través del tiempo en el lenguaje de los usuarios que abandonaron un grupo de apoyo. También se han realizado análisis en usuarios que hablan libremente sobre su condición y usuarios que retienen información; donde a pesar de cubrir sus huellas lingüísticas, cuando se categorizan en grupos temáticos siguen siendo más similares a los usuarios deprimidos que al grupo de control. En el trabajo de (Wolohan et al., 2018) esto fue de ayuda debido a usuarios deprimidos que se resistían a hablar públicamente sobre sus síntomas depresivos, pero que sin embargo sufren de depresión.

Otra forma interesante de usar representaciones LIWC consiste en usarlas junto a tópicos generados con LDA (Latent Dirichlet Allocation). LDA es un modelo probabilístico generativo en el que cada elemento de una colección se modela como una mezcla finita sobre un conjunto subyacente de temas (Blei, Ng, y Jordan, 2003). Por ejemplo, en (Loveys et al., 2018) se compara la forma en que cambia la expresión de depresión según la cultura haciendo un análisis de temas seleccionados con LIWC y un análisis de modelado de tópicos con LDA. Construir una representación mezclando más de un enfoque temático también ha sido una forma de afrontar la detección de depresión. Combinar atributos LIWC con un número de tópicos generados automáticamente ha mostrado un buen desempeño (Resnik, Garron, y Resnik, 2013). Esto también ha sido mostrado al explorar variaciones del modelo como lo hace el trabajo de (Resnik et al., 2015) al utilizar una versión supervisada de Latent Dirichlet Allocation (sLDA).

Representaciones basadas en emociones

Recientemente las representaciones basadas en emociones han expuesto el potencial de usar emociones discretas como atributos, en lugar de usar únicamente

características textuales. Para construir estas representaciones se consideran ocho emociones reconocidas básicas (Ekman and Davidson, 1994); se generan emociones finamente-granuladas usando un recurso léxico con el fin de calcular un histograma de frecuencias para estudiar la efectividad de las características basadas en emociones. El trabajo de (Chen et al., 2018) aprovecha la intensidad de las emociones para identificar depresión en usuarios de Twitter. (Aragón et al., 2019) va un paso más allá usando una Bolsa de Sub Emociones (BoSE), que representa las publicaciones de los usuarios mediante un histograma de emociones finamente-granuladas. Primero creando un diccionario de sub-emociones para posteriormente remplazar cada palabra en un texto con la etiqueta de su sub-emocion más cercana. Finalmente, a partir de este *texto enmascarado* se calcula el histograma de sub-emociones.

Representaciones aprendidas

Las redes neuronales se han utilizado para obtener resultados sobresalientes, especialmente en el área de clasificación de imágenes. Recientemente, estudios han demostrado que también se pueden usar de manera efectiva para tareas de clasificación de texto (Zhang and Wallace, 2017). Una arquitectura común consiste en utilizar una red neuronal convolucional (CNN) y después una red neuronal recurrente (RNN). El uso de *embeddings* se han convertido en una forma popular y eficiente de modelar palabras e interacciones entre ellas; usan varias neuronas para representar una palabra y dejan que cada neurona sea parte de la descripción de varias palabras. (Trotzek, Koitka, y Friedrich, 2018, 2020) definieron múltiples estrategias y consideraron un amplio rango de características para construir sus modelos. Por ejemplo, meta-datos lingüísticos a nivel de usuario, *embeddings* basados en GloVe y fastText; y usaron modelos basados en CNN y RNN. (Yates, Cohan, y Goharian, 2017) emplearon una arquitectura más sencilla usando CNN donde la entrada es cada publicación de un usuario. Cada publicación es procesada por la red convolucional y mezclada para crear una representación vectorial del usuarios; y con ese vector se llevará a cabo la clasificación.

3.2. Utilización de múltiples tipos de atributos

En detección de depresión existen trabajos que combinan dos o más tipos de atributos. Algunos trabajos han considerados características de los usuarios, por

ejemplo, género, edad, o incluso su personalidad. Otros han considerado atributos más estilísticos como uso de pronombres o emoticones. También hay trabajos que, además del texto, consideran información multimodal.

El trabajo de (Schwartz et al., 2014) tiene como objetivo estimar como cambia la depresión a través de las estaciones del año. En este trabajo la información sobre personalidad es evaluada junto a n-gramas de palabras, tópicos LDA y categorías LIWC. Los resultados muestran a la depresión como una estructura que cambia con el tiempo y no solo como una enfermedad que se tiene o no.

En (Mowery et al., 2016) clasifican publicaciones de los usuarios para conocer si contiene o no evidencia de depresión. Para ello, ellos usan el *género y edad* de los usuarios junto a n-gramas de palabras, etiquetado POS, emoticones y categorías LIWC. Los resultados obtenidos no fueron concluyentes en cuanto a su relevancia para esta tarea, ya que en todos los experimentos se utilizaron junto con el resto de los atributos.

Otro ejemplo es el trabajo de (Shen et al., 2017) donde evalúan el lenguaje usado por los usuarios para diferenciar entre deprimidos y no deprimidos. Proponen un enfoque multimodal que combinan emociones, información personal, tópicos, palabras específicas del dominio e información de la imagen del perfil del usuario. Los resultados muestran que los usuarios deprimidos expresan más emociones negativas en redes sociales.

3.3. Foros de evaluación

Las redes sociales se han vuelto una plataforma atractiva para desarrollar enfoques relacionados a la salud mental que den soporte a usuarios con alguno de estos problemas, es por ello que existen algunos foros de evaluación.

eRisk

La Conferencia y Laboratorios del Foro de Evaluación (CLEF, por sus siglas en inglés) consiste en una conferencia independiente *peer-review* sobre una amplia gama de temas en los campos de la evaluación de acceso a la información multilingüe y

multimodal, y un conjunto de laboratorios y talleres diseñados para probar diferentes aspectos de los sistemas de recuperación de información *mono* y *cross-lingüe*. Early Risk (eRisk) es un taller que surge en el año 2017 con el objetivo de explorar la metodología de evaluación, las métricas de efectividad y las aplicaciones prácticas (particularmente las relacionadas con la salud y la seguridad) de la detección temprana de riesgos en internet. Su objetivo principal es crear una nueva área de investigación interdisciplinaria potencialmente aplicable a una amplia variedad de situaciones y perfiles personales diferentes. Los ejemplos incluyen posibles pedófilos, acosadores, individuos que podrían caer en manos de organizaciones criminales, personas con tendencias suicidas o personas susceptibles a la depresión.

En eRisk 2018 el laboratorio tuvo dos tareas principales: 1) Detección temprana de signos de depresión y, 2) Detección temprana de signos de anorexia. Para la tarea en detección temprana de signos de depresión el conjunto de entrenamiento proveniente de Reddit se fue liberando secuencialmente en partes; cada parte representó el historial del usuario en un periodo de tiempo. El reto consistió en preprocesar esas piezas de evidencia y detectar indicios de depresión lo antes posible. Los resultados obtenidos en el foro fueron similares a los logrados en el del año pasado. Aunque el rendimiento aún es modesto, esto sugiere que diferenciar a los usuarios deprimidos de los no deprimidos sigue siendo un desafío.

CLPsych

El taller de Lingüística Computacional y Psicología Clínica (CLPsych, por sus siglas en inglés) ha acogido tareas compartidas y no compartidas durante varios años. El objetivo de estas tareas es proveer comparaciones de varios enfoques para modelar lenguaje relevante para la salud mental en redes sociales.

En el año 2015 la tarea compartida consistió en hacer una clasificación de acuerdo a tres categorías de salud mental: usuarios auto-diagnosticados con depresión, usuarios auto-diagnosticados con trastorno por estrés postraumático (PTSD, por sus siglas en inglés) y usuarios de control que no reportaron nada. Los datos usados para la tarea provienen de Twitter. Esta tarea compartida sirvió como oportunidad para que una variedad de equipos compararan técnicas y enfoques para extraer señales lingüísticas relevantes para la salud mental de los usuarios.

3.4. Discusión

A continuación se describen los criterios considerados en la tabla 3.1 para comparar los trabajos de detección de depresión:

- **Múltiples tipos de atributos:** considera si usan más de un tipo de atributo en la construcción de la representación de los usuarios.
- **Atributos demográficos:** si considera, dentro del conjunto de atributos, algunos relacionados con el perfil del usuarios, como género o edad.
- **Atributos de sentimiento y/o emoción:** si considera, dentro del conjunto de atributos, algunos relacionados con sentimientos y/o emociones.
- **Clasificadores especiales por tipo de usuario:** es decir, si se construyeron clasificadores especiales para ciertos tipos de usuarios.

Varios trabajos cuentan con información de los atributos demográficos de los usuarios pero muy pocos son los que la consideran para construir la representación; a pesar de que algunos se enfocan en explorar la diferencia entre los usuarios deprimidos y no deprimidos por la forma en que se expresan. Los trabajos que si la consideran, no le dan mucha relevancia ya que en los experimentos la utilizan junto con el resto de los atributos. El trabajo propuesto se diferencia en especializar la clasificación e integrar su información del perfil como atributos extra; por otro lado, el uso de múltiples tipos de atributos para construir la representación nos ayuda a obtener una representación enriquecida. Por esta razón en este trabajo se usará información de sentimientos y emociones además de la textual. En otras palabras, al considerar la información de perfil junto a la polaridad de sus publicaciones se puede mejorar la detección de usuarios que sufren de depresión en redes sociales.

Con la revisión de los trabajos presentados en este capítulo se lograron identificar puntos importantes para integrarlos en este trabajo. El método propuesto se caracteriza por usar de dos formas alternativas la información de perfil, como atributos extras y como clasificadores independientes para ciertos usuarios. Además de usar una representación basada en la polaridad de las publicaciones y otra basada en emociones.

Trabajo	Múltiples tipos de atributos	Atributos demográficos	Atributos de sentimiento y/o emoción	Clasificadores especiales por tipo de usuario
(Jamil et al., 2017)	Sí	No	Sí	No
(Sadeque et al., 2016)	Sí	No	Sí	Sí
(Wolohan et al., 2018)	Sí	No	Sí	No
(Loveys et al., 2018)	Sí	Sí	No	No
(Resnik, Garron, y Resnik, 2013)	Sí	No	No	No
(Chen et al., 2018)	Sí	No	Sí	No
(Aragón et al., 2019)	No	No	Sí	No
(Trotzek, Koitka, y Friedrich, 2018)	Sí	No	No	No
(Yates, Cohan, y Goharian, 2017)	Sí	No	Sí	No
(Schwartz et al., 2014)	Sí	No	Sí	No
(Mowery et al., 2016)	Sí	Sí	Sí	No
(Shen et al., 2017)	Sí	No	Sí	No
Propuesta	Sí	Sí	Sí	Sí

Tabla 3.1: Algunos trabajos relevantes.

Capítulo 4

Métodos y representaciones propuestas

En el presente capítulo se presentan los enfoques propuestos para la detección de depresión de usuarios en redes sociales. Éstos consideran las siguientes ideas principales: i) el uso de los atributos de los usuarios y ii) la integración de la polaridad y emociones de sus publicaciones. En las primeras secciones se explican los enfoques mencionados y en la última sección el método con la combinación de ambos enfoques.

4.1. Género y edad en la detección de depresión

El análisis de lenguaje, usado por usuarios con alguna enfermedad mental a través del contenido que generan en redes sociales, ha sido escasamente explorado cuando involucra como factor a atributos demográficos. El trabajo presentado por (Preotiuc-Pietro et al., 2015) realiza un análisis del lenguaje de las personas deprimidas usando los atributos demográficos de *género y edad*. Para realizar dicho análisis se hace una predicción de los usuarios con una regresión logística usando un atributo ya sea género o edad. En los resultados obtenidos por las predicciones de usuarios, se concluye que usar solo género o edad como único atributo no es muy útil en la clasificación; para mejorar los resultados de la clasificación se considera hacer una predicción combinando ambos atributos de usuario y se obtiene una mejoría del 15 %, ilustrando que entre ellos existe información complementaria. Este resultado sugiere que estos dos atributos de usuarios tienen alguna información relevante para esta tarea y deben ser tomados en cuenta. Por esta razón se decide obtener estos dos atributos demográficos de los usuarios en los conjuntos de datos y buscar la manera

de trabajar con ellos incorporándolos en la representación. En la descripción de los siguientes métodos se asume que el género y la edad ya son conocidos.

4.1.1. Género y edad como atributos extra

Este enfoque consiste simplemente en incluir los atributos de edad y género en la representación de los usuarios. El objetivo de este enfoque es explorar el efecto de usar la información del perfil como dos atributos extra, si es que se enriquece la representación y ayudan en el proceso de clasificación. La figura 4.1 muestra de forma gráfica esta idea.

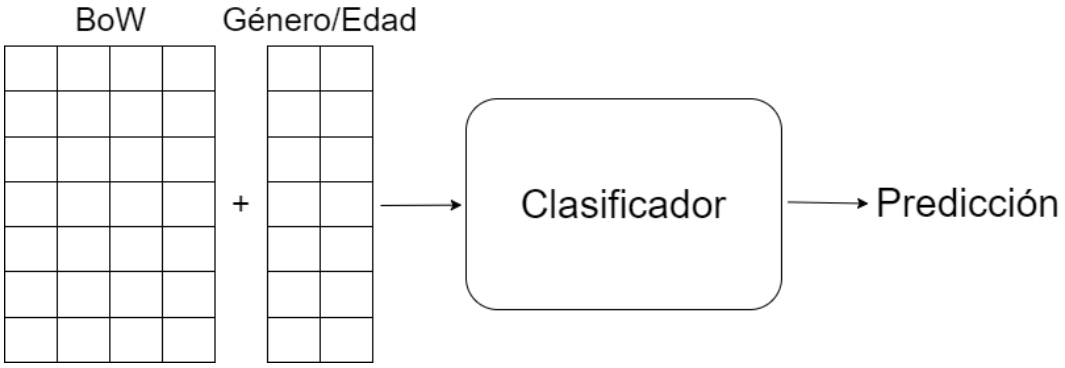


Figura 4.1: Género y edad como atributos extras

En la representación de bolsa de palabras se usan uni-gramas de palabras con un pesado TF-IDF. Antes de incluir los atributos de género y edad se normalizan sus *valores estimados*[1] en un rango de 0 a 1 con la ecuación 4.1.1.

$$Z_i = \frac{X_i - min(X)}{max(X) - min(X)} \tag{4.1.1}$$

Donde X_i es el valor de género o edad del i-ésimo usuario, $min(X)$[2] es el valor mínimo y $max(X)$[3] el valor máximo de entre todos los usuarios.

[1]Los valores obtenidos en la suma al aplicar el lexicón de referencia en el texto de los usuarios.
[2]En el caso del género el valor mínimo es un número con signo negativo.
[3]En el caso del género el valor máximo es un número con signo positivo.

4.1.2. Clasificadores específicos por tipo de usuario

Basándose en la predicción de género y edad este enfoque consiste en entrenar clasificadores especializados para cada tipo de usuario, y agregar un atributo extra. Básicamente, las muestras de género y edad se dividen en dos subconjuntos: hombres y mujeres, jóvenes y adultos; $\mathcal{U}_H$ y $\mathcal{U}_M$ siendo $\mathcal{U}_H \cup \mathcal{U}_M = \mathcal{U}$, y $\mathcal{U}_J$ y $\mathcal{U}_A$ siendo $\mathcal{U}_J \cup \mathcal{U}_A = \mathcal{U}$ respectivamente. La idea de forma general es, dado un usuario a clasificar se determina su género o edad, y luego dependiendo de éste atributo la instancia se dirige a su clasificador correspondiente; ya sea el clasificador de hombres o de mujeres, o si se usó la edad al clasificador de jóvenes o adultos. La figura 4.2 muestra las ideas anteriores.

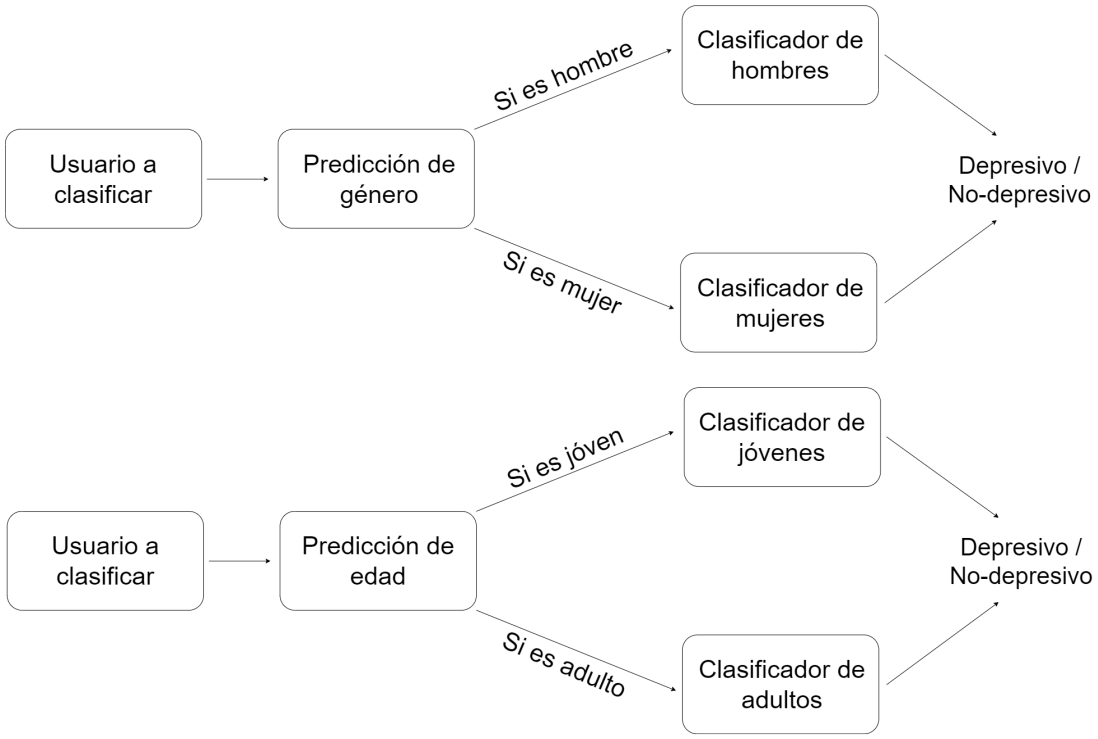


Figura 4.2: Clasificadores específicos de depresión por tipo de usuario.

Los valores de género y edad pueden ser conocidos de antemano, o pueden determinarse automáticamente. La predicción automática de estos atributos se describe en la sección 5.3.1. Este enfoque de clasificación propuesto es independiente de la representación de los usuarios, puede usarse cualquiera. En este trabajo en particular, y

con el interés de analizar las diferencias temáticas y emocionales, entre los distintos grupos de usuarios depresivos, usamos un representación basada en palabras y otra en emociones.

Representación con Bolsa de Palabras

En el enfoque temático se usa una representación de bolsa de palabras con unigramas de palabras y pesado TF-IDF para cada género. A esta representación se le agrega únicamente como atributo extra el género o la edad (previamente normalizado). Esto es así porque el género ya está siendo usado de forma explícita en los clasificadores específicos de hombre y mujer, por otra parte cuando se trabaja con edad, se usa de forma explícita joven y adulto.

Representación con Bolsa de Sub Emociones

Para el enfoque emocional se utiliza la representación basada en sub-emociones propuesta en (Aragón et al., 2019). En esta representación cada palabra se asocia a una sub emoción específica, por ejemplo, el texto *"Weekly Wednesday happiness thread* se representa como *"positive438 positive36 joy279 anticipation317"*. Posteriormente se hace el mismo procedimiento para construir una bolsa de palabras, en este caso es una bolsa de sub emociones con uni-gramas; este es uno de los primeros acercamientos en este trabajo para la exploración de usar información de perfil junto a emociones.

4.2. Polaridad de las publicaciones en la detección de depresión

Los sentimientos son un tipo de información importante en la comunicación de las personas, según como se estudie puede relacionarse con el estado de ánimo. En años recientes el contenido léxico, representado por texto, ha sido de los más estudiados. Sin embargo, no solo las palabras son las que diferencian a deprimidos y no-deprimidos, sino la polaridad asociada a éstas es lo que en muchas ocasiones es lo más relevante. Por ejemplo, cuando se mencionan palabras relacionadas a asuntos familiares o laborales en contextos positivos o negativos.

4.2.1. Nueva representación

Se aplica un método para obtener la polaridad de cada publicación en el historial de usuario y separar las publicaciones por polaridad, es decir, ahora para cada usuario se tiene un historial de publicaciones positivas, negativas y neutras. En la descripción de los siguientes métodos se asume que para cada post ya se conoce su polaridad, en la sección 5.3.2 se explica cómo se determinaron éstas en nuestros experimentos.

Representación de polaridad positiva y negativa

Después de obtener la separación de publicaciones por polaridad se construyen dos bolsas de palabras, una con las publicaciones positivas y otra con las publicaciones negativas. En las representaciones se usa un pesado TF, de esta forma tenemos los vocabularios con sus valores para cada polaridad. Para las publicaciones con polaridad neutra no se construye una representación porque se desea comparar la diferencia de las palabras en contextos positivos y negativos.

Unión y normalización de las polaridades

El siguiente paso consiste en concatenar ambas bolsas de palabras con polaridades para hacer una comparación entre ellas. Después de concatenar las dos bolsas de palabras se realizó una normalización de sus valores. Esta normalización se hizo de la siguiente manera:

$$t\widehat{f}^{+}_{i,j} = \frac{f^{+}_{i,j}}{f^{+}_{i,j} + f^{-}_{i,j}} \tag{4.2.1}$$

Donde $f_{i,j}$ es el número de ocurrencias del término t_j en el documento d_i, $f^{+}_{i,j}$ es el número de ocurrencias del término en dicho documento en la bolsa de palabras con polaridad positiva y $f^{-}_{i,j}$ en la bolsa de palabras con polaridad negativa.

De forma complementaria, el nuevo cálculo del valor TF para la bolsa de palabras con polaridad negativa se hace igual a la ecuación anterior quedando 4.2.2.

$$t\widehat{f}^{-}_{i,j} = \frac{f^{-}_{i,j}}{f^{-}_{i,j} + f^{+}_{i,j}} \tag{4.2.2}$$

En la figura 4.3 se muestra un ejemplo para el cálculo en ambas representaciones utilizando la palabra *familia*.

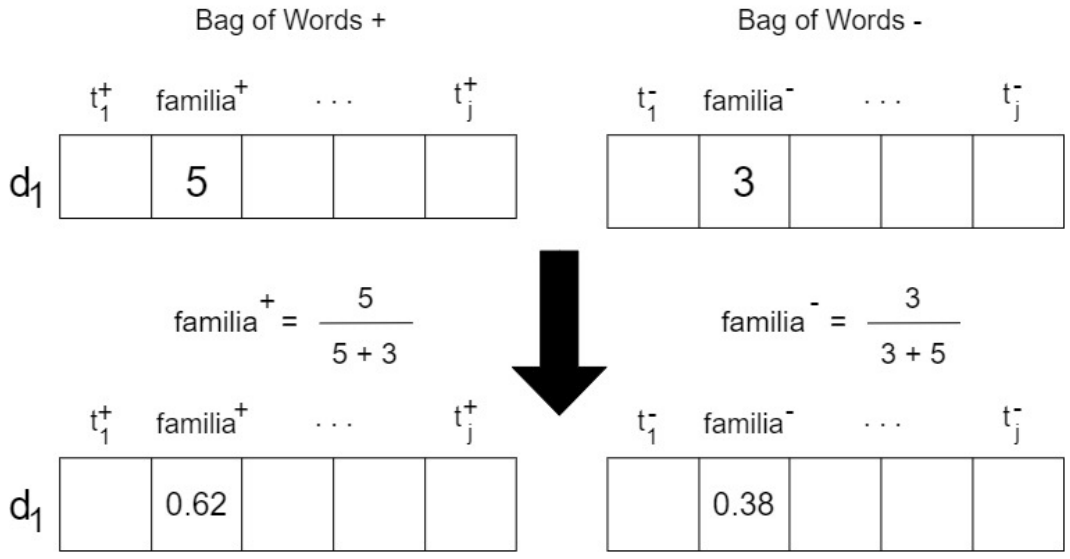


Figura 4.3: Cálculo del nuevo valor TF.

Después de calcular los nuevos valores TF en cada bolsa de palabras, la nueva representación será conocida a partir de ahora como Bolsa de Polaridades (BoP). Esta nueva representación será usada para sustituir a la bolsa de palabras.

4.3. Combinación de información de perfil con la polaridad y emociones de las publicaciones

La idea de combinar la información basada en palabras junto a sus polaridades tiene como objetivo comparar su uso en contextos positivos y negativos.

4.3.1. Clasificación combinando ambos enfoques

Los enfoques presentados en la sección 4.1 también pueden ser usados con la representación de Bolsa de Polaridades (BoP). Ambos enfoques se mantienen igual y solo la presentación se cambia, incluso el uso de los atributos del perfil se mantienen igual. La configuración de la nueva representación se presenta a continuación:

Bolsa de Polaridades con multi-atributos

Se considera nuevamente la idea de agregar información adicional como atributos extra en la representación. Para enriquecer la representación de Bolsa de Polaridades (BoP) además de agregar la información del perfil se agregan los porcentajes por polaridad en el historial de cada usuario.

Los tres atributos de porcentaje de polaridad son calculados al momento de hacer la separación de *post* por polaridad. Al inicio se acumulan haciendo un simple conteo de *posts*, posteriormente se convierten en porcentajes y al final los porcentajes son convertidos en un número entre 0 y 1. De esta forma, tanto la información de perfil como la información de polaridad queden representados en la misma escala.

Cabe mencionar que en los porcentajes de *post* por polaridad se consideran las tres polaridades: positiva, negativa y neutral. La figura 4.4 muestra de forma gráfica la representación de Bolsa de Polaridades + Multi-atributos.

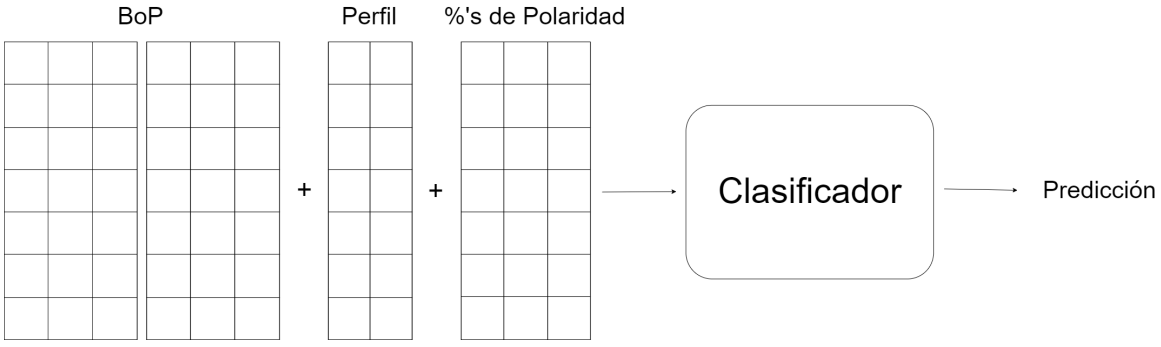


Figura 4.4: Bolsa de Polaridades + Multi-atributos.

4.3.2. Bolsa de Sub Emociones + Bolsa de Emociones

Se tiene la idea de comparar la información basada en palabras (contenido temático) y la información basada en el contenido emocional. El objetivo es comprobar que hombres y mujeres que sufren de depresión se diferencian tanto en los temas que mencionan como en el tipo de emociones que transmiten.

De la misma forma que la sección anterior, se presenta la nueva configuración de la representación siendo ahora Bolsa de Sub Emociones. La idea es agregar como atributos extra las nueve emociones de las que parte BoSE. Con las nueve emociones se hace un cálculo similar a bolsa de palabras por lo que pasa a ser llamada Bolsa de Emociones (BoE). La figura 4.5 muestra de forma gráfica la representación BoSE + BoE.

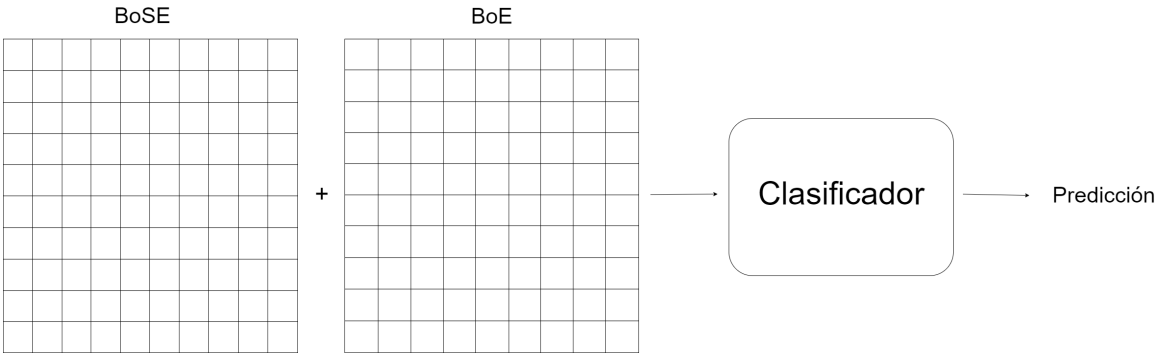


Figura 4.5: Bolsa de Sub Emociones + Bolsa de Emociones.

Capítulo 5

Experimentos y resultados

En el presente capítulo se describe la configuración experimental usada para evaluar los métodos propuestos. En las secciones se explica lo relacionado a los conjuntos de datos, desde cómo están compuestos, sus características, análisis que se hicieron, limpieza y preprocesamiento. También se expone el método para la predicción de género y edad junto a cómo se determinó la confiabilidad de esos datos. Posteriormente se presentan los resultados en la detección de usuarios con depresión de Reddit y Twitter, se incluye un análisis de la clasificación y de las características que fueron clave.

5.1. Conjuntos de datos

Para hacer la evaluación de los métodos propuestos se utilizaron dos conjuntos de datos en inglés, cada conjunto fue construido para la tarea de detección de depresión.

El primer conjunto de datos se obtiene de la plataforma de marcadores sociales y sitio web de discusión **Reddit**. Esta plataforma está compuesta por varias comunidades pequeñas llamadas *Subreddits*, los subreddits están dedicados a un tema en particular sobre el que las personas hacen publicaciones relacionadas. Entre los subreddits existentes se pueden encontrar algunos relacionados a diferentes condiciones médicas como la depresión. Este conjunto de datos proveniente de Reddit fue utilizado en la tarea compartida eRisk 2018 (Losada, Crestani, y Parapar, 2018). Este conjunto de datos contiene el historial de publicaciones de usuarios que cumplieran los criterios de selección. Para la construcción se usó el criterio de la *auto-declaración*

del diagnóstico clínico de depresión y la *pertenencia a foros y grupos de soporte web*. Los términos y condiciones de Reddit permitieron usar el contenido para motivos de la investigación usando la información de forma anónima, en el trabajo de (Losada and Crestani, 2016) se describe más a detalle la construcción del conjunto de datos.

El segundo conjunto de datos es obtenido de usuarios en la red social Twitter, este conjunto fue construido y utilizado en el trabajo de (Shen et al., 2017). Para obtener los datos se emplearon métodos heurísticos basados en reglas, obteniendo así dos conjuntos de datos bien etiquetados como referencia para usuarios con depresión y sin depresión en Twitter. Como las personas deben ser observadas un periodo de tiempo de acuerdo a la experiencia clínica, se obtiene un tweet de anclaje junto a todos los tweets publicados en un lapso de 30 días. Los usuarios son considerados como deprimidos si alguna de sus publicaciones contenía la expresión *"(I'm/ I was/ I am/ I've been) diagnosed with depression"*. Los usuarios son considerados como no deprimidos si nunca hicieron alguna publicación que contenga la palabra *"depress"*.

La tabla 5.1 muestra información de los conjuntos de datos y cómo se distribuye en el entrenamiento y prueba para cada clase.

Conjunto de datos	**Training**		**Test**	
	Dep	NDep	Dep	NDep
eRisk'18	135	752	79	741
Twitter	2626	5373		

Tabla 5.1: Conjunto de datos usados, cada conjunto de datos tiene las dos clases (Deprimido = Dep, No-Deprimido = N.Dep).

5.2. Limpieza de los conjuntos de datos

Hacer una limpieza de los conjuntos de datos antes de preprocesarlos, y el preprocesamiento mismo, es un paso importante que tiene como objetivo preparar la información para las siguientes etapas o para futuros cambios y manipulaciones que se

hagan sobre ellos. Todas estas etapas se verán reflejadas en el proceso de clasificación.

Cada conjunto de datos recibe una limpieza y preprocesamiento diferente, ambos se describen a continuación:

Reddit

Como se mencionó anteriormente, Reddit está compuesto de subreddits dedicados a un tema en particular sobre el que las personas hacen sus publicaciones.

Reposteo y publicación de spam o enlaces

Reddit siendo considerada como una red social cuenta con algunos detalles que al momento de extraer los datos para construir el conjunto de datos de forma automática son incluidos y pueden causar ruido. Por ejemplo, considerar a un *bot* como un usuario. En Reddit, obtener el historial del usuario no garantiza extraer solo las publicaciones de un subreddit en particular, es decir, aunque la persona fue elegida por pertenecer o seguir temas de desórdenes mentales, las publicaciones extraídas no son solo de ellos. Esto apoya a la diversidad de temas evitando que la clasificación sea basada en entradas relacionadas con un tema.

Reddit dispone de funciones comunes a cualquier red social, algunas funciones causan problemas al extraer el historial y por ello son considerados en la limpieza. El primero de ellos es publicar el mismo contenido en diferentes subreddits o grupos e incluso el compartir la publicación de alguien mas en el perfil propio. Hacer un **reposteo** de publicaciones puede ocupar mucho espacio en el historial del usuario dependiendo de la extensión de la publicación y el número de veces que fue publicada; además de ocupar entradas en el historial esta información se encuentra repetida y puede no ser útil al construir la representación. Un par de detalles extras también son tomados en cuenta junto al reposteo de publicaciones; la **publicación de spam** es algo que se quiere evitar porque es de interés el contenido que es escrito originalmente por el mismo usuario. Hacer publicaciones excesivas, extensas o en el que se compartan muchos **enlaces** puede ser señal de que la persona está haciendo una variante de reposteo publicando contenido de otra persona o simplemente

compartiendo muchos enlaces externos que no tienen sentido con un objetivo en común.

Para lidiar con estos problemas se usó la siguiente heurística considerando las entradas de las publicaciones en el conjunto de datos: se trata de reposteo cuando la publicación solo contiene el título, es decir, se trata de una publicación o publicaciones sin texto en las que el título es lo único que las identifica; se considera spam a las publicaciones que sin importar su extensión incluyan una gran cantidad de enlaces externos a cualquier otro sitio web.

Twitter

Auto-declaración del diagnóstico y publicación de spam o enlaces

Para el conjunto de datos de Twitter se hizo una limpieza un poco diferente a los datos de Reddit. Twitter a diferencia de Reddit es una red social más conocida y más usada por las personas donde no existen grupos de forma explícita siendo lo más cercano a ellos los *hilos*.

Al usarse el enfoque de ***auto-declaración del diagnóstico clínico de depresión*** y debido a los tweets de anclaje, un usuario con depresión puede ser identificado fácilmente provocando que la clasificación se haga de forma simple únicamente por esas entradas. De forma similar a como sucede en Reddit, la **publicación de spam** es algo que se desea evitar y esto está ligado con la publicación de **enlaces**, ya sean enlaces a páginas externas o hacia el mismo Twitter.

Remover publicaciones relacionadas con depresión ayudará a evitar el sesgo que se provoque por publicaciones muy relacionadas o relacionadas directamente con depresión. Para remover estas entradas se usa el patrón usado para recolectar a los usuarios deprimidos: si algunas de las publicaciones contenía la expresión *"(I'm/ I was/ I am/ I've been) diagnosed with depression"* es removida del historial del usuario; de igual forma se considera spam a las publicaciones que sin importar que contengan mucho texto (considerando que en Twitter está limitada la cantidad de caracteres por publicación) esta incluya varios enlaces para sitios web externos o del mismo Twitter.

5.3. Análisis de los conjuntos de datos

Antes de empezar a trabajar con los conjuntos de datos se hace un análisis de la información que se obtuvo después de aplicar la limpieza y extraer los atributos del perfil. Esto se hace así con el fin de preparar los datos para el preprocesamiento que se haga antes de la etapa de clasificación.

5.3.1. Método de predicción de atributos demográficos

La información de perfil, especialmente los atributos demográficos, han sido relacionados con algunos síntomas depresivos por la manera en que son expresados y por cómo se manifiestan a través de los usuarios. El género y la edad juegan un papel importante en los desórdenes de salud mental, para este fin, se busca hacer una predicción de ambos en cada usuario y usarlos de forma complementaria para mejorar la detección de depresión.

Predicción de género y edad

Para la predicción de género y edad de cada usuario en los conjuntos de datos se usan los recursos léxicos disponibles en el sitio web del Proyecto de Bienestar Mundial[1] (WWBP, por sus siglas en inglés). La predicción primero se hace a nivel publicación de forma que se acumulen los pesos de cada una, la suma final de los pesos de cada publicación representará el resultado de la predicción en el historial del usuario; el lexicón ponderado es aplicado como la suma de todas las frecuencias relativas de palabras ponderadas sobre un documento:

$$usage_{lex} = \sum_{word \in lex} w_{lex}(word) * \frac{freq(word, doc)}{long(doc)} \tag{5.3.1}$$

Donde $w_{lex}(word)$ es el peso de la palabra $word$ en el lexicón lex, $freq(word, doc)$

[1] http://wwbp.org/lexica.hmtl

es la frecuencia de la palabra en un documento/usuario y $long(doc)$ es el número total de palabras para ese documento/usuario. La edad del usuario corresponde al resultado de su suma, mientras que el género se indica con su signo, un resultado positivo indica un usuario femenino y un resultado negativo un usuario masculino.

Lexicones de atributos demográficos

Los lexicones que se desarrollaron en (Sap et al., 2014) son los adecuados para esta tarea, estos lexicones fueron derivados de palabras de uso en Facebook, blogs y Twitter usando modelos de regresión y clasificación para obtenerlo en forma de palabras con pesos asociados. Por lo tanto este lexicón contiene lo necesario para realizar una buena predicción. Este recurso se divide en dos lexicones, la tabla 5.2 muestra algunas palabras del lexicón de género y la tabla 5.3 muestra algunas palabras del lexicón de edad.

Término	**Peso**
girlfriend	-470.437717
barber	-267.0095954
grandfather	-145.7553358
comics	-97.47982951
...	**...**
pregnant	78.52096042
boyfriend	384.5768948
makeup	440.3430309
pedicure	575.3626421

Tabla 5.2: Lista de algunas palabras en el lexicón de género.

La longitud de los lexicones de género y edad es de 7137 y 10797 términos respectivamente. El lexicón de edad contiene términos y pesos más variados, una forma fácil de interpretar ambos lexicones es considerar el signo de los pesos. En género, un valor con signo positivo indica ser más femenino y negativo masculino, en edad mientras el valor es más grande indica que ese término es usado por personas de edad mayor.

Término	**Peso**
grandparents	-478.1726186
mummy	-366.6457657
highschool	-270.7439212
semester	-83.28277991
...	**...**
institute	56.08309549
wife	221.4001809
sons	642.2636269
grandson	1967.141828

Tabla 5.3: Lista de algunas palabras en el lexicón de edad.

Las figuras 5.1 y 5.2 muestran la distribución de género y edad en cada conjunto de datos. Para la edad se divide en dos grupos los usuarios, un usuario se considera Joven si su edad es menor o igual a 24 años, por otra parte, si es mayor a 24 años se considera Adulto.

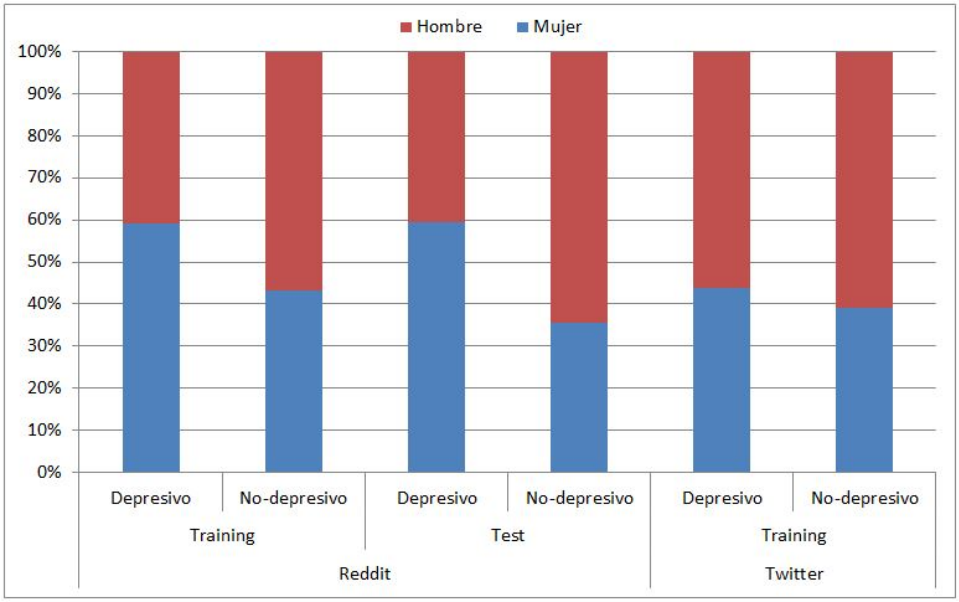


Figura 5.1: Distribución de género en ambos conjuntos de datos.

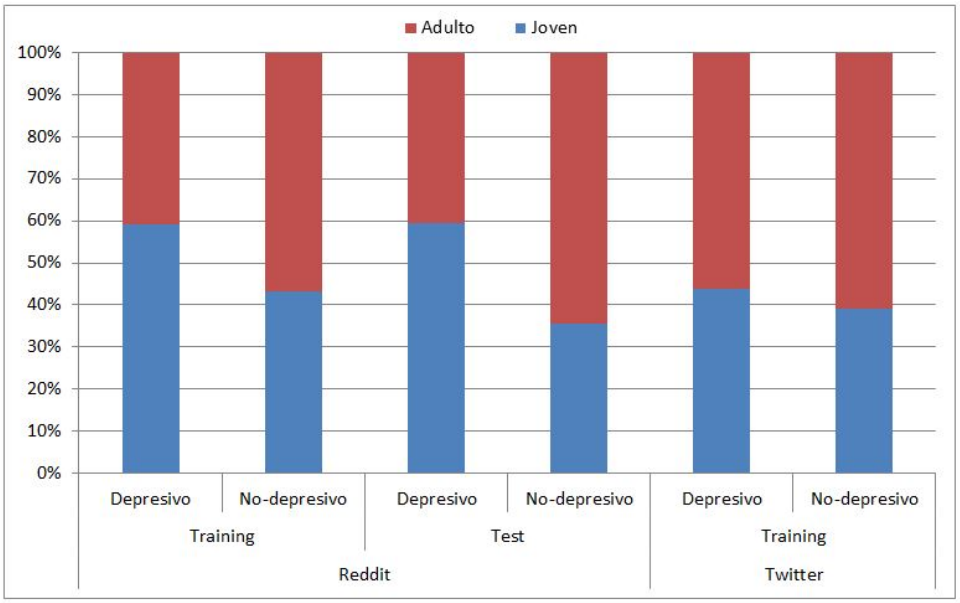


Figura 5.2: Distribución de edad en ambos conjuntos de datos.

5.3.2. Clasificación de post por polaridad

Actualmente en estudios de Procesamiento del Lenguaje Natural a menudo se definen los sentimientos usando escalas de valores. Por ejemplo, una escala de 5-puntos de Likert para representar desde fuertemente negativo a fuertemente

positivo; pero también como forma de medición a menudo se puede descomponer en dos aspectos: *la polaridad y la intensidad del sentimiento* (Tian, Lai, y Moore, 2018).

Para determinar el tono emocional en las publicaciones hechas por los usuarios en redes sociales se hace una predicción de la polaridad por cada post del usuario en su historial. Para realizar esta predicción y clasificar cada post de acuerdo a su polaridad se desarrolló el método que se describe a continuación.

Método para el cálculo de polaridad

Para cada historial de usuario en el conjunto de datos se obtienen una a una las publicaciones para hacer su cálculo de polaridad. Como primer paso se hace un etiquetado POS para obtener una lista con los pares (palabra, etiqueta) que más adelante serán comparados utilizando SentiWordNet. Antes de hacer la comparación se lematiza la palabra de acuerdo a su etiqueta POS, la etiqueta POS en todo el método juega un papel importante para no alterar la publicación. Los synset en SentiWordNet al igual que en WordNet están ordenados por frecuencia de uso (01 siendo el sinónimo más común, un número mayor indica menor uso común). Cada synset contiene la estructura **('palabra.etiqueta.uso')**, al comparar en SentiWordNet un par (palabra, etiqueta) se busca el sinónimo más usado que coincida con la etiqueta POS deseada y si se cumplan estos dos requisitos se acumula la puntuación positiva/negativa de la palabra. Una vez comparada la última palabra de esa publicación, el valor final se obtiene con la división de la suma de las puntuaciones entre el número de palabras que aportaron información. En la figura 5.3 se muestra de forma gráfica el cálculo de polaridad.

El algoritmo 3 explica de forma más detallada el cálculo de polaridad presentado en la figura 5.3. En el algoritmo también se puede notar cómo al momento de hacer el cálculo de polaridad se acumula el número de post en cada polaridad; más adelante al terminar de calcular todo el historial estos valores acumulados se convierten en porcentajes y serán usados también en la representación que se construya.

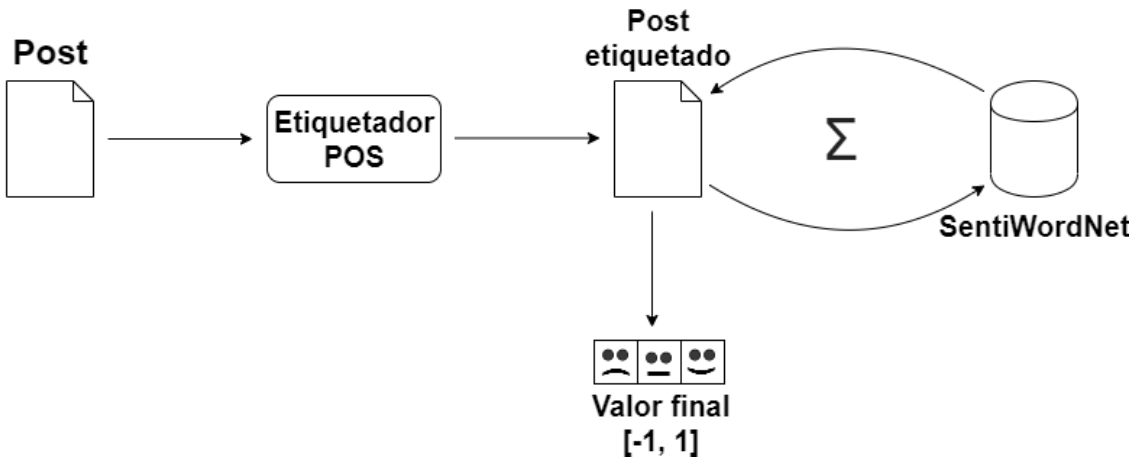


Figura 5.3: Cálculo de polaridad.

Algoritmo 3 *Cálculo de polaridad (extendido)*

Entrada: P: publicación del usuario
Salida: V: Valor entre [-1, 1], D: acumulador de posts por polaridad

L = etiquetado POS de la publicación P
Para $l = 1$ *hasta longitud_de_L* **hacer:**
    $lemma$ = obtener lema de la palabra de acuerdo a su etiqueta POS
    $synsets$ = obtener el synset del lema de acuerdo a su etiqueta POS
    Si $longitud\_synset > 0$ **entonces**
        $sinonimo = synset\_mas\_frecuente$
        $acumulador = sinonimo.valor\_positivo + sinonimo.valor\_negativo$
        $emparejados+ = 1$
    Fin si
    $l+ = 1$
Fin para
V = acumulador / emparejados
Si $V < 0$ **entonces**
    $D[negativos]+ = 1$
si no si $V == 0$
    $D[neutrales]+ = 1$
si no
    $D[positivos]+ = 1$
Fin si

Distribución de polaridad

Se hace un análisis de la distribución de las publicaciones según su valor de polaridad recordando que el valor que se asigna a cada publicación está dentro del

rango de [-1, 1]; indicando que -1 es una publicación muy negativa, 1 una publicación muy positiva y 0 indica polaridad neutra.

La figura 5.4 muestra cómo se distribuye la polaridad en el conjunto de entrenamiento y prueba para el conjunto de datos eRisk 2018, ambos tienen una distribución similar donde la polaridad menos positiva y neutral cubren la mayoría de publicaciones.

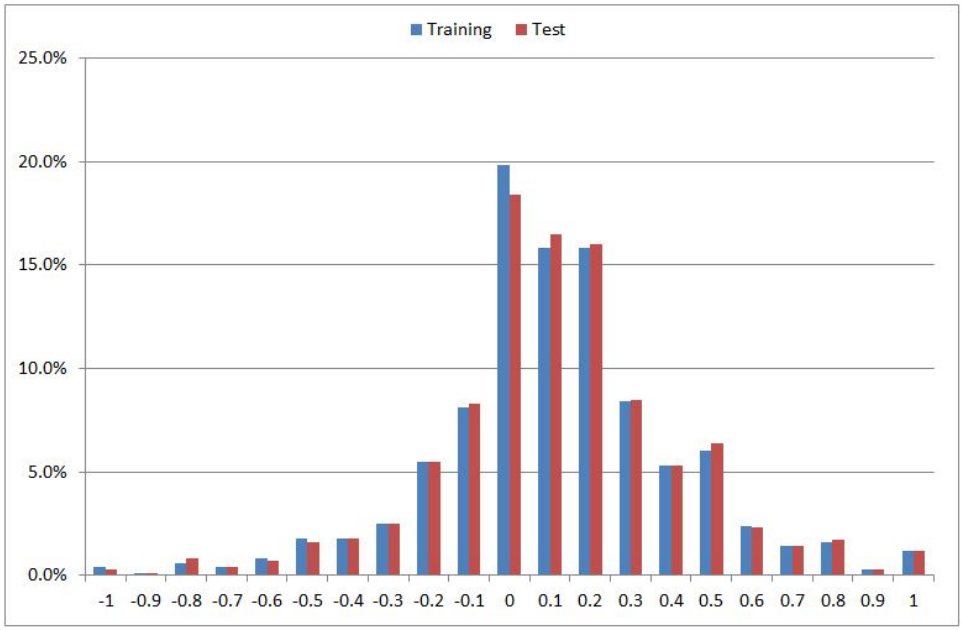


Figura 5.4: Distribución de polaridad en el conjunto de datos de Reddit.

En el conjunto de datos de Twitter las publicaciones se siguen acumulando entre los menos positivos y menos negativos además de la polaridad neutral, la figura 5.5 muestra esta distribución.

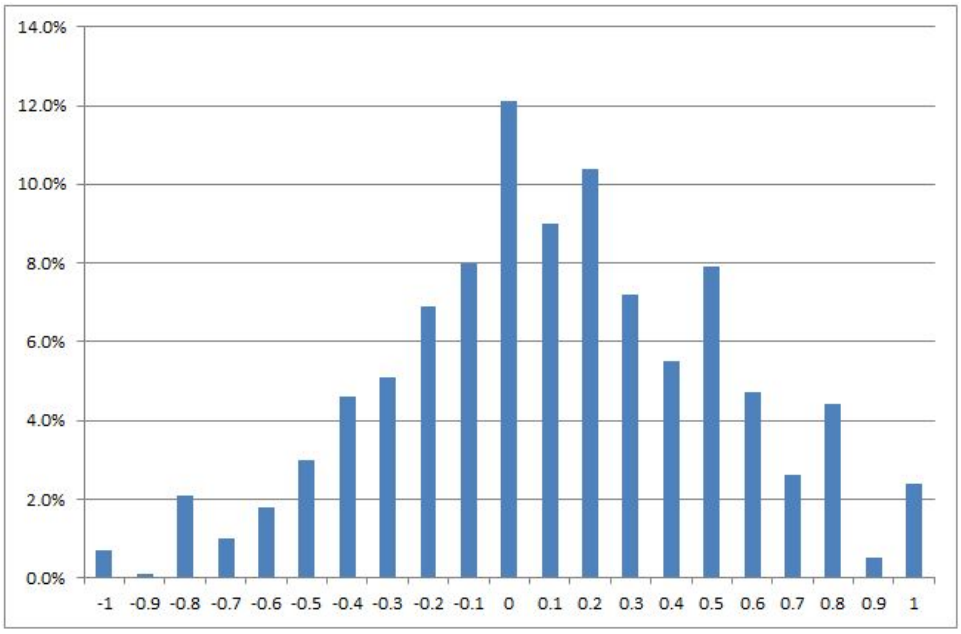


Figura 5.5: Distribución de polaridad en el conjunto de datos de Twitter.

5.4. Evaluación de los métodos

5.4.1. Configuración experimental

Preprocesamiento

El preprocesamiento de los conjuntos de datos tiene como propósito prepararlos para las siguientes etapas que se realizarán con ellos. De forma general en ambos conjuntos de datos se normaliza el texto con una expresión regular removiendo caracteres especiales y pasando a minúsculas todas las palabras. Para el conjunto de datos de Twitter adicionalmente se convierten los *emojis* de cada publicación a su equivalente en texto.

Clasificación

Para la etapa de clasificación se seleccionaron las características más discriminativas usando la distribución chi^2 (3,000 y 3,500 palabras para Reddit y Twitter respectivamente, y 500 sub-emociones en ambos casos). Se probaron los algoritmos

de clasificación *Máquina de Soporte Vectorial*, *Árboles de Decisión*, *Random Forest* y *Bagging*. Sin embargo, el algoritmo de ensambles de clasificadores Bagging con Árboles de Decisión fue el que mostró mejor rendimiento, por lo que los resultados mostrados en las demás secciones son los que se obtuvieron para este clasificador. En la clasificación del conjunto de datos eRisk 2018 se hicieron cinco repeticiones de Bagging (por ser un clasificador estocástico); para el conjunto de datos de Twitter al no estar dividido en entrenamiento y prueba se hizo una validación cruzada de 5 pliegues, estableciendo la partición del entrenamiento en 80 % y la de prueba en 20 %. Para evaluar las predicciones de los métodos y representaciones propuestas se usa como métrica el valor F1 sobre la clase positiva (depresiva), por ser la forma en que se realiza en trabajos previos y en el foro eRisk. Los experimentos fueron desarrollados en el lenguaje de programación *Python* mediante herramientas conocidas, entre ellas *NLTK 3.5*[2] y *Scikit-Learn*[3].

Baseline

Para evaluar la mejora que tiene cada método en la etapa de clasificación se crea un baseline, con el cual se compararan los resultados obtenidos. Para la construcción del baseline se usan dos representaciones, la primera es una Bolsa de Palabras con uni-gramas de palabras y la segunda es una Bolsa de Sub Emociones con uni-gramas de sub emociones (Aragón et al., 2019).

5.4.2. Evaluación del rol de la información de perfil en la detección de depresión

Para evaluar la importancia que tiene la información de perfil para la detección de usuarios con depresión se llevan a cabo dos experimentos diferentes. En el primero se evalúa/mide el efecto de agregar el **género y edad como atributos extra** a la representación de Bolsa de Palabras. En el segundo se evalúa el método de usar clasificadores específicos para cada tipo de usuario, por ejemplo, hombres y mujeres, o jóvenes y adultos. El objetivo de estos experimentos es conocer que tan beneficiosa

[2]https://www.nltk.org
[3]https://scikit-learn.org/stable

puede ser la información de perfil para la representación y la clasificación de los usuarios con depresión.

Género y edad como atributos extra

La tabla 5.4 muestra los resultados obtenidos en las dos colecciones usadas. Se compara el rendimiento de las dos representaciones (BoW y BoSE) usando género y edad como atributos extra.

Método	Conjunto de datos Reddit	Conjunto de datos Twitter
BoW	0.63 (+/- 0.04)	0.85 (+/- 0.02)
BoW + Atributos Extra	0.64 (+/- 0.05)	0.86 (+/- 0.01)
BoSE	0.59 (+/- 0.06)	0.82 (+/- 0.02)
BoSE + Atributos Extra	0.63 (+/- 0.02)	0.83 (+/- 0.03)

Tabla 5.4: Resultados de la metrica F1 sobre la clase positiva en la alternativa de usar atributos extra considerando Bolsa de Palabras (BoW) y Bolsa de Sub Emociones (BoSE)

Como se puede observar, agregar los atributos de género y edad en ambas representaciones (BoW y BoSE) mejora los resultados de clasificación del baseline. En ambos conjuntos de datos y representaciones es visible la mejora que aporta agregar estos dos atributos.

Método de clasificadores específicos por tipo de usuario

En la evaluación de esta alternativa dependiendo la información de perfil que se está usando, se agrega solo un atributo de perfil a la representación. Cada clasificador recibe y se especializa en un tipo de usuario. En la tabla 5.5 se observan los resultados obtenidos usando clasificadores específicos en género y edad.

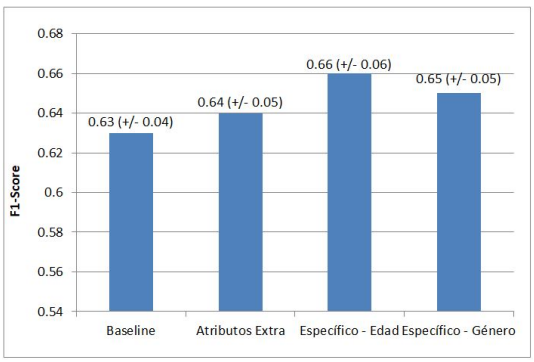
Cuando se trabaja con edad un clasificador se especializa en la predicción de usuarios Jóvenes y el otro en usuarios Adultos, añadiendo en la representación, ya sea BoW o BoSE, un atributo extra correspondiente al genero del usuario. En caso contrario, cuando se trabaja con género un clasificador se especializa en la predicción de usuarios Masculinos y el otro en usuarios Femeninos, agregando a la representación, ya sea BoW o BoSE, un atributo extra correspondiente a la edad del usuario.

Método	Conjunto de datos Reddit	Conjunto de datos Twitter
BoW	0.63 (+/- 0.04)	0.85 (+/- 0.02)
BoW + Clasificadores específicos (Edad)	**0.66 (+/- 0.06)**	0.86 (+/- 0.02)
BoW + Clasificadores específicos (Género)	0.65 (+/- 0.05)	**0.87 (+/- 0.02)**
BoSE	0.59 (+/- 0.06)	0.82 (+/- 0.02)
BoSE + Clasificadores específicos (Edad)	0.64 (+/- 0.04)	0.83 (+/- 0.02)
BoSE + Clasificadores específicos (Género)	**0.66 (+/- 0.03)**	**0.85 (+/- 0.02)**

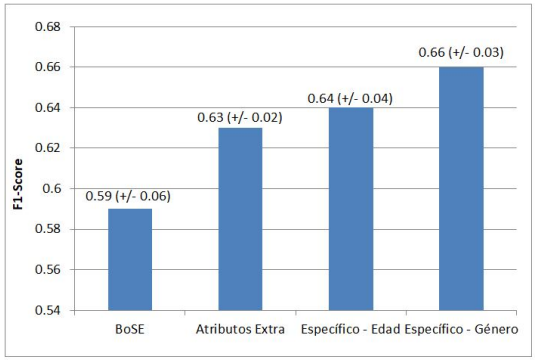
Tabla 5.5: Resultados de la métrica F1 sobre la clase positiva en la alternativa de usar clasificadores específicos considerando Bolsa de Palabras (BoW) y Bolsa de Sub Emociones (BoSE)

Observando los resultados con el uso de clasificadores específicos para la clasificación la diferencia es más notable. Esto muestra que aunque agregar la información de perfil como atributos extra ayuda a mejorar el rendimiento de la clasificación, hacer uso de clasificadores especializados es aún más útil.

A forma de resumen, la figura 5.6 muestra los resultados comparativos entre el baseline, el uso de atributos extra y el uso de clasificadores específicos obtenidos en el conjunto de datos de Reddit. La figura 5.7 muestra los mismos resultados comparativos para el conjunto de datos de Twitter.

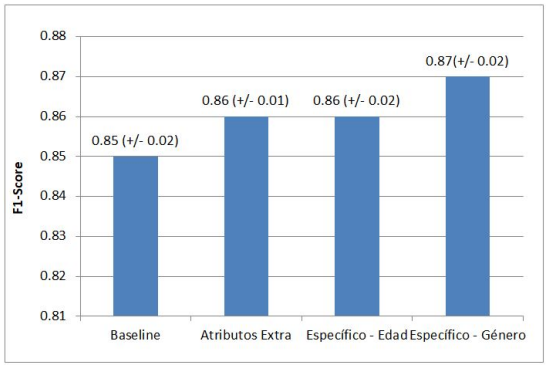


(a) Reddit - BoW

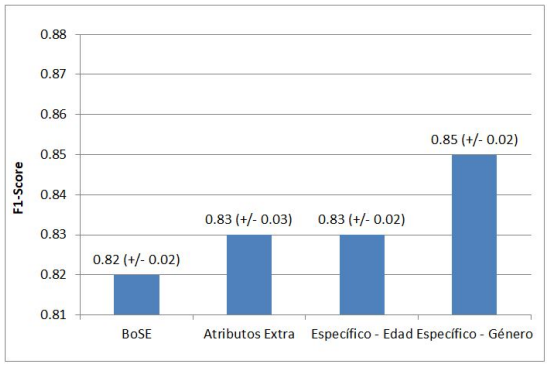


(b) Reddit - BoSE

Figura 5.6: Resultados obtenidos en la clasificación usando información del perfil en el conjunto de datos de Reddit.



(a) Twitter - BoW



(b) Twitter - BoSE

Figura 5.7: Resultados obtenidos en la clasificación usando información del perfil en el conjunto de datos de Twitter.

5.4.3. Evaluación del rol de la polaridad de las publicaciones en la detección de depresión

Considerando los resultados obtenidos usando la información del perfil, en esta sección se evalúa y compara el rendimiento de la nueva representación en la etapa de clasificación usando múltiples atributos. La tabla 5.6 muestra estos resultados. En el baseline se cambia la representación de Bolsa de Palabras por la representación de Bolsa de Polaridades (BoP). La representación de Bolsa de Sub Emociones se cambia por la representación de Bolsa de Sub Emociones+Bolsa de emociones (BoSE+BoE).

El objetivo de estos experimentos es comparar la información que aporta la nueva representación sin y con atributos de perfil además de agregar información porcentual de las polaridades.

Método	Conjunto de datos Reddit	Conjunto de datos Twitter
BoP	0.59 (+/- 0.04)	0.84 (+/- 0.01)
BoP + Multi-Atributos	0.65 (+/- 0.03)	0.85 (+/- 0.01)
BoSE + BoE	0.63 (+/- 0.02)	0.83 (+/- 0.01)
BoSE + BoE + Perfil	0.64 (+/- 0.04)	0.85 (+/- 0.02)

Tabla 5.6: Resultados de la métrica F1 sobre la clase positiva usando la representación BoP+Multi-Atributos y la representación BoSE + BoE

En estos experimentos se puede observar que usar la nueva representación junto a los 5 atributos adicionales (género, edad y los tres porcentajes de polaridad) iguala o mejora los resultados obtenidos en la sección anterior. El resultado de usar BoP+Multi-atributos en la representación es el mismo al uso de clasificadores específicos, de igual forma agregar las nueve emociones a la representación BoSE mejora los resultados.

5.4.4. Evaluación del método combinando ambos enfoques

Motivados por el buen desempeño obtenido en la mayoría de los experimentos se evalúa el último enfoque de clasificación propuesto. La tabla 5.7 muestra la comparación de los valores F1 obtenidos en la clasificación. En este enfoque se compara el rendimiento de usar la representación BoP+Multi-Atributos y la representación BoSE+BoE, usando clasificadores específicos por tipo de usuario.

Los resultados obtenidos en la tabla 5.7 indican que: *i*) incluir la información de perfil junto a la polaridad de las publicaciones contribuyen a mejorar la detección de usuarios con depresión; *ii*) usar el enfoque de clasificadores específicos por tipo de

usuario sigue siendo aún más útil.

Método	Conjunto de datos Reddit	Conjunto de datos Twitter
BoP	0.59 (+/- 0.04)	0.84 (+/- 0.01)
BoP + Multi-Atributos	0.65 (+/- 0.03)	0.85 (+/- 0.01)
BoP + Clasificadores específicos (Edad)	0.66 (+/- 0.08)	0.88 (+/- 0.03)
BoP + Clasificadores específicos (Género)	**0.68 (+/- 0.04)**	**0.88 (+/- 0.02)**
BoSE + BoE	0.63 (+/- 0.02)	0.83 (+/- 0.01)
BoSE + BoE + Perfil	0.64 (+/- 0.04)	0.85 (+/- 0.02)
BoSE + BoE + Clasificadores específicos (Edad)	0.66 (+/- 0.05)	0.85 (+/- 0.03)
BoSE + BoE + Clasificadores específicos (Género)	**0.67 (+/- 0.04)**	**0.86 (+/- 0.01)**

Tabla 5.7: Resultados de la métrica F1 sobre la clase positiva usando la representación BoP+Multi-Atributos y la representación BoSE+BoE en el enfoque de clasificadores específicos por tipo de usuario.

5.4.5. Comparación con el Estado del Arte

En las secciones anteriores se evaluaron los diferentes métodos de clasificación propuestos con las diferentes representaciones. Aunque el objetivo de esta tesis es un método para la detección de depresión que combine la información del perfil con los sentimientos y mensajes de las publicaciones, se comparan los resultados obtenidos con los del estado del arte. Las figuras 5.8 y 5.9 muestran la comparación de los mejores resultados reportados hasta el momento contra tres de nuestros resultados: baseline (BoW), mejor resultado considerando información del perfil de usuario, y mejor resultado considerando la representación basada en polaridad enriquecida con información del perfil. En el caso de la colección de Reddit el mejor resultado

reportado hasta el momento es de 0.64 (Trotzek, Koitka, y Friedrich, 2018) y en la colección de Twitter es de 0.85 (Shen et al., 2017).

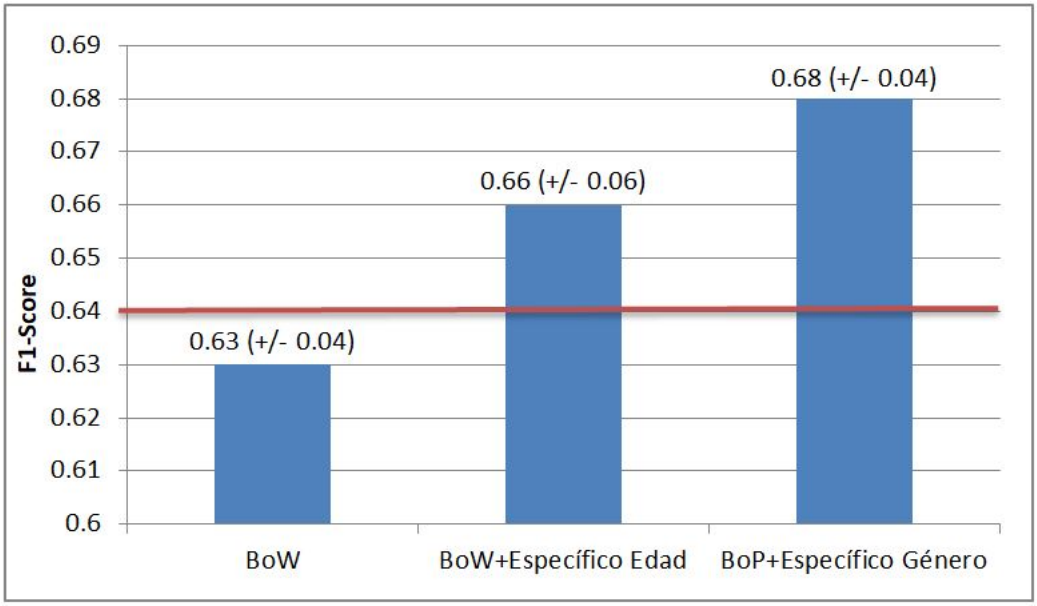


Figura 5.8: Comparación de resultados en eRisk 2018, la línea horizontal indica el resultado en el estado del arte obtenido por (Trotzek, Koitka, y Friedrich, 2018).

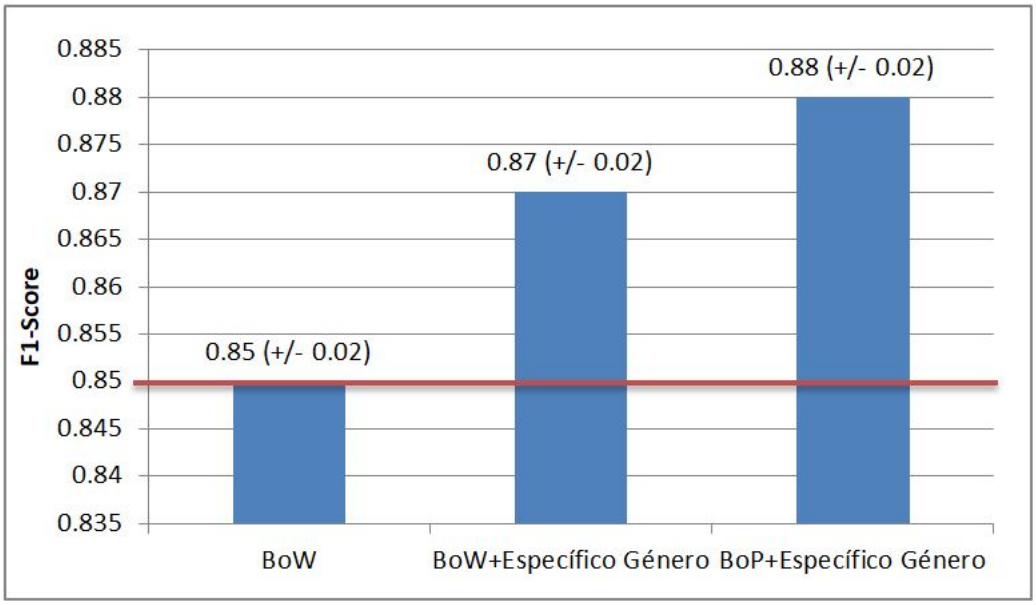


Figura 5.9: Comparación de resultados en Twitter, la línea horizontal indica el resultado en el estado del arte obtenido por (Shen et al., 2017).

Para el baseline se mantiene *BoW* por ser una de las representaciones más usadas en las tareas de clasificación de textos, *BoW+Específico Edad* muestra el resultado obtenido al usar clasificadores específicos en jóvenes y adultos, *BoW+Específico Género* muestra el resultado obtenido al usar clasificadores específicos en hombres y mujeres, y *BoP+Específico Género* muestra el resultado obtenido al usar clasificadores específicos en hombres y mujeres con una representación BoP+Multi-atributos.

5.5. Análisis de resultados

Para comprender más los resultados alcanzados por el método propuesto, se realizan diversos análisis tanto de los resultados como de la representación y la información extraída de los conjuntos de datos en la etapa de clasificación. Este análisis también tiene como objetivo encontrar diferencias en el vocabulario y en las emociones de los usuarios.

5.5.1. Errores en la predicción

Teniendo en cuenta los resultados mostrados en la sección anterior, para el análisis de errores en la predicción nos enfocamos en el método que obtuvo el mejor rendimiento en cada conjunto de datos. En las figuras 5.10 y 5.11 se muestra el análisis de errores según el **tamaño de historial de usuario** para el conjunto de datos de Reddit y Twitter respectivamente. La división se hizo asumiendo que los datos se comportan como una distribución normal. Para cada historial de usuario se obtiene el número de palabras totales por las que está compuesto y posteriormente para cada intervalo se obtienen el número usuarios cuyo historial corresponde a esa extensión.

En ambos conjuntos de datos se sigue el mismo comportamiento respecto a los usuarios que el método predice mal. Conforme el tamaño del historial va siendo más grande los errores en la predicción se van reduciendo debido a la cantidad de evidencia con la que se cuenta. En Twitter esto es más notable debido a la cantidad de usuarios que componen el conjunto de datos presentando un porcentaje de error menor al que se presenta en Reddit.

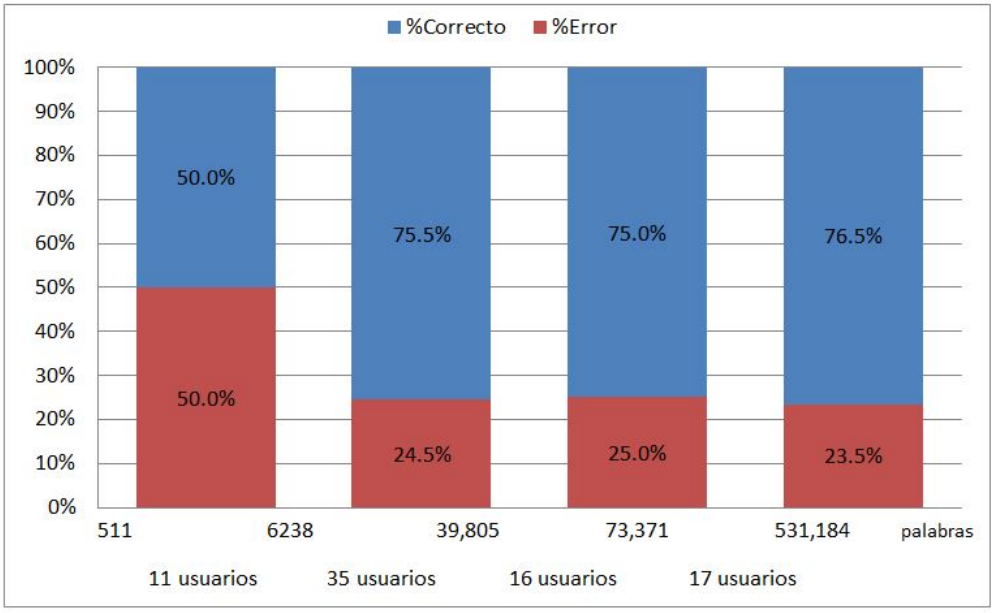


Figura 5.10: Error por tamaño de historial en el conjunto de datos de Reddit.

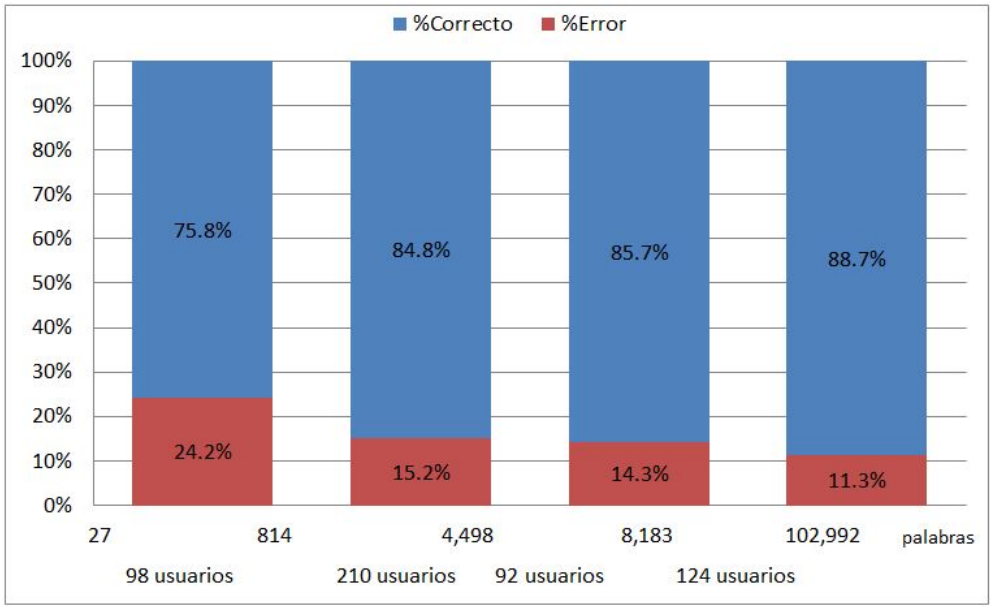


Figura 5.11: Error por tamaño de historial en el conjunto de datos de Twitter.

En las figuras 5.12 y 5.13 se presenta el análisis por **expansión temporal** para ambos conjuntos de datos. La expansión temporal es el tiempo que estuvo en observación cada usuario, el tiempo de observación se calcula con la fecha de la primera y la última publicación en el historial. El tiempo se obtiene en días.

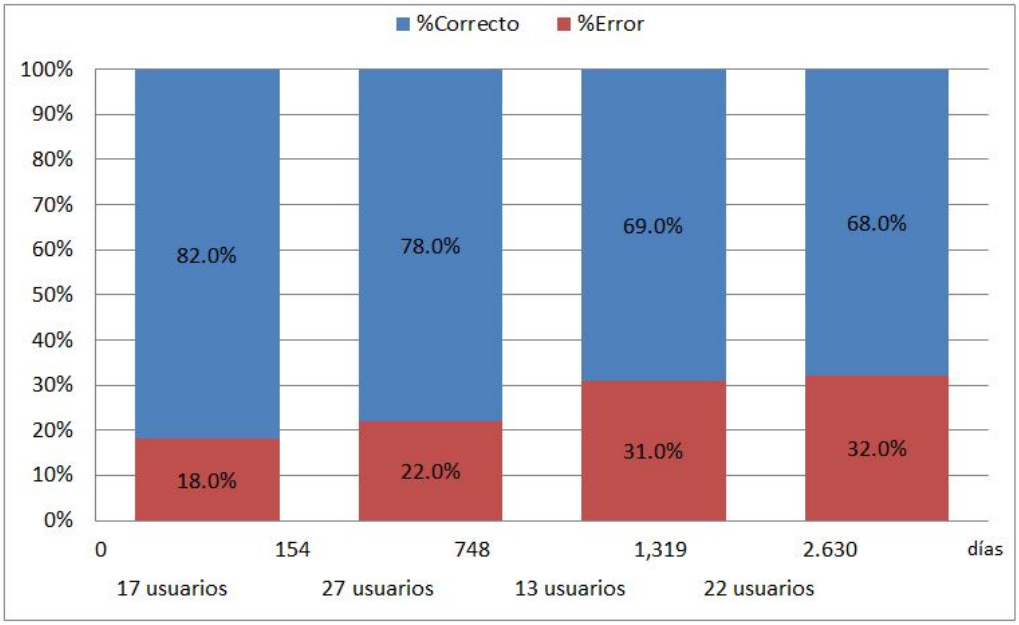


Figura 5.12: Error por expansión temporal en el conjunto de datos de Reddit.

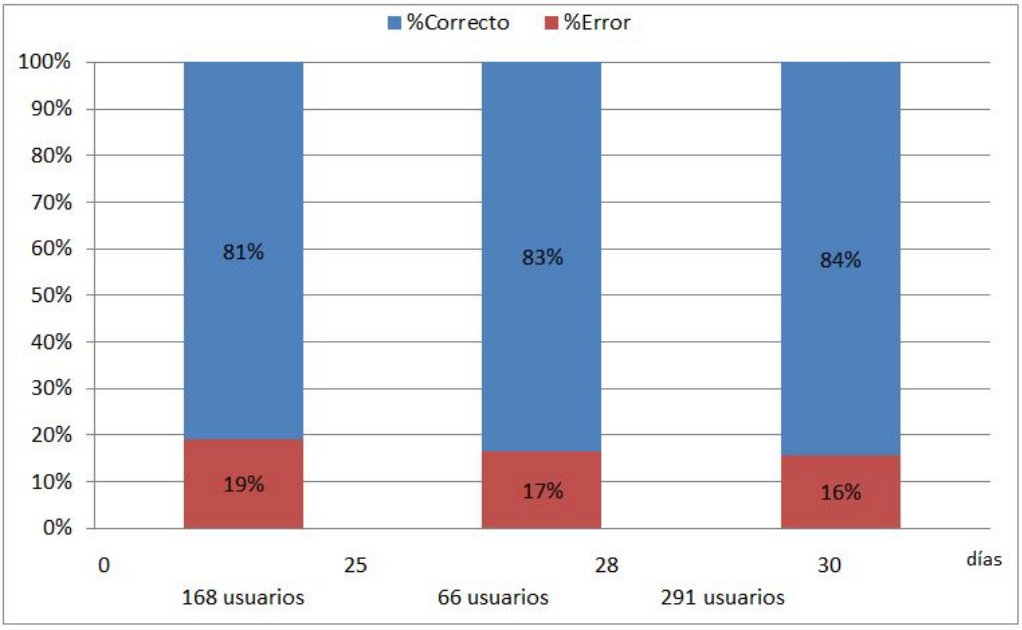


Figura 5.13: Error por expansión temporal en el conjunto de datos de Twitter.

Comparando los resultados obtenidos se observa un comportamiento diferente para cada conjunto. Esto se debe principalmente al criterio de construcción para el conjunto de datos de Twitter donde los usuarios fueron obtenidos en un lapso de

tiempo fijo (30 días). En el conjunto de datos de Twitter mientras más tiempo se observa a un usuario el porcentaje de error en la predicción es menor; algo que pasa de forma contraria en Reddit donde si un usuario se observa demasiado tiempo el porcentaje de error en la predicción sube. Este comportamiento se puede explicar debido a que los usuarios que principalmente están deprimidos logran superar su estado depresivo pero pasado un tiempo tienden a tener recaídas afectando así su clasificación.

5.5.2. Diferencia de vocabulario por Género

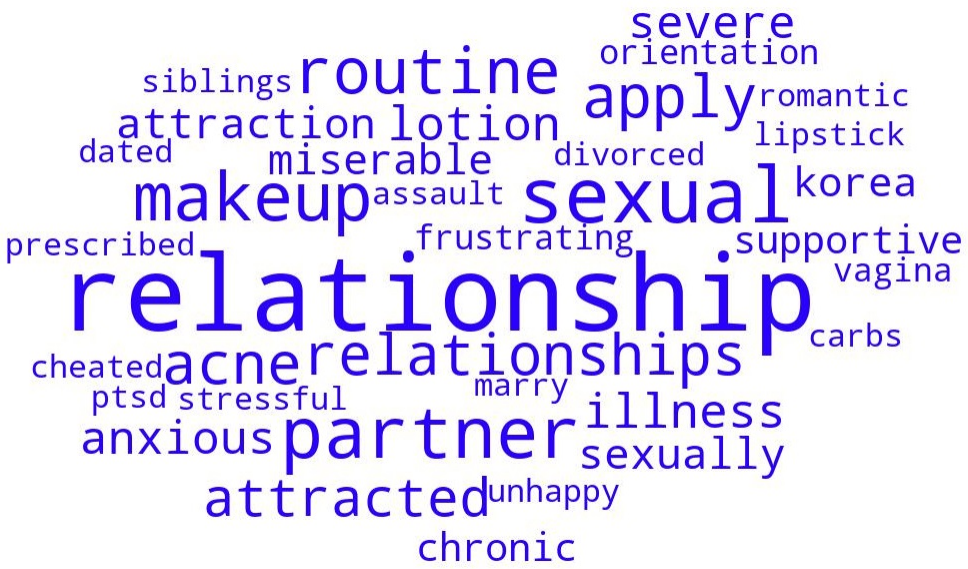
Respecto a la diferencia entre hombres y mujeres que sufren depresión, en cada conjunto de datos se obtiene el vocabulario que diferencia a cada género. La figura 5.14 muestra palabras comunes en ambas redes sociales[4]. Las figuras 5.15 y 5.16 muestra las nubes de sus palabras más frecuentes.

En Reddit es posible observar diferencias claras entre ambos géneros, mientras los hombres mencionan más aspectos de enfermedades o medicamentos (therapy, medication, psychiatrist, xanax, vitamin), las mujeres usualmente mencionan más términos relacionados a aspectos afectivos o de apariencia física. Similarmente, en el conjunto de datos de Twitter los hombres siguen mencionando aspectos de sus enfermedades mientras que las mujeres, además de mencionar términos relacionados a su apariencia física, también mencionan términos relacionados a su peso (calories, weightloss, anorexic).

Figura 5.14: Palabras en común para el género en ambas redes sociales (Reddit y Twitter). Por simplicidad las palabras de hombres y mujeres se presentan en una sola nube.

[4]Se obtuvieron las palabras comunes de ambos géneros, pero se daba el caso de que en un género solo se obtenían 4 o 5 palabras y por eso se optó por presentarlas en una sola nube.

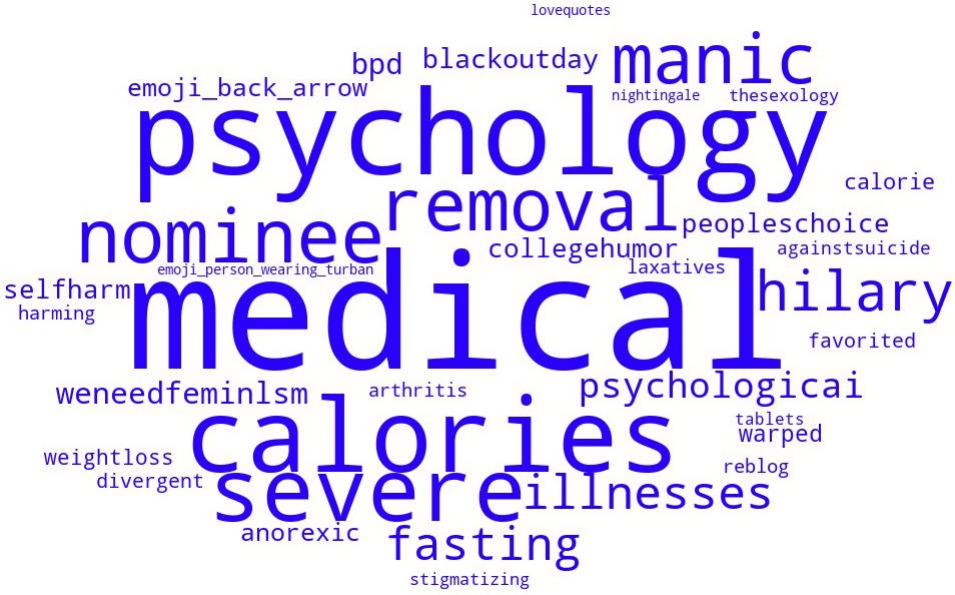
(a) Hombres



(b) Mujeres

Figura 5.15: Top 50 de las palabras usadas por hombres y mujeres deprimidos en el conjunto de datos de Reddit.

(a) Hombres



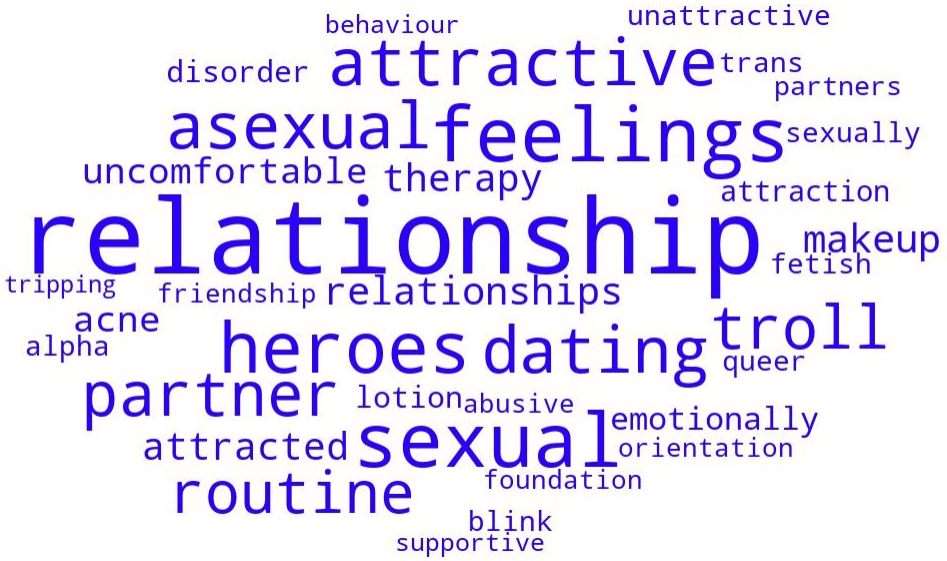
(b) Mujeres

Figura 5.16: Top 50 de las palabras usadas por hombres y mujeres deprimidos en el conjunto de datos de Twitter.

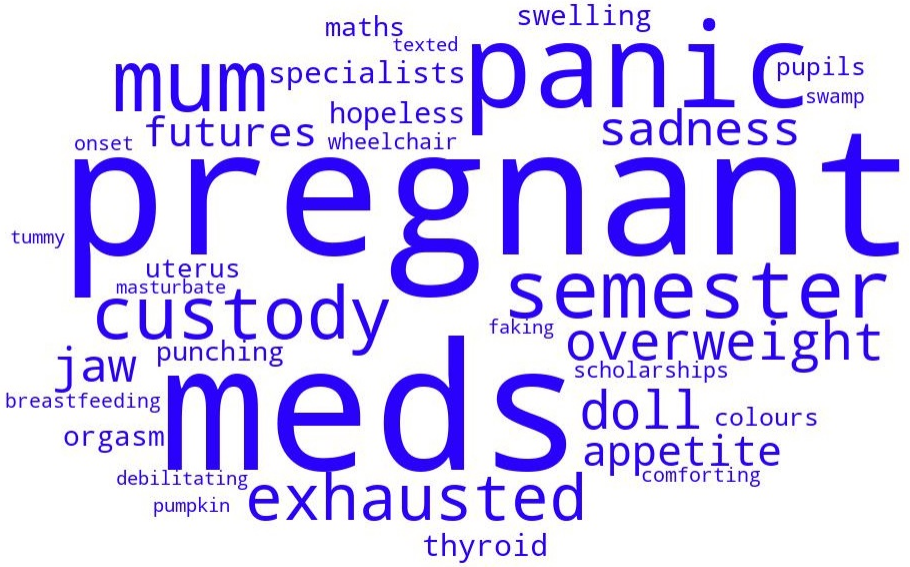
5.5.3. Diferencia de vocabulario por Edad

Cuando se hace la diferenciación entre jóvenes y adultos (figura 5.17) es normal que los temas cambien al igual que las palabras más comunes. En Reddit, los usuarios

jóvenes hablan sobre relaciones afectivas y apariencia física mientras que los usuarios adultos hablan más sobre sexualidad, medicamentos y cansancio.

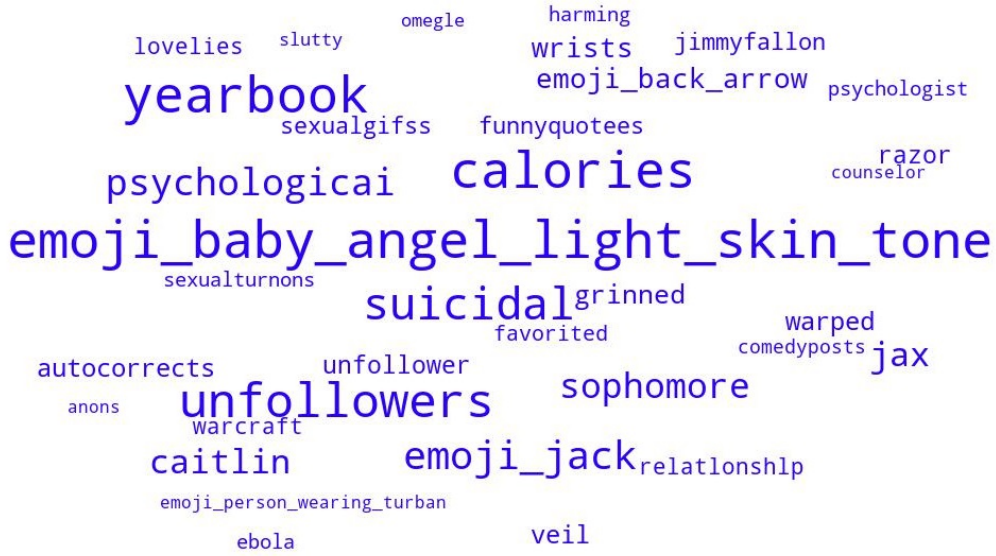


(a) Jóvenes

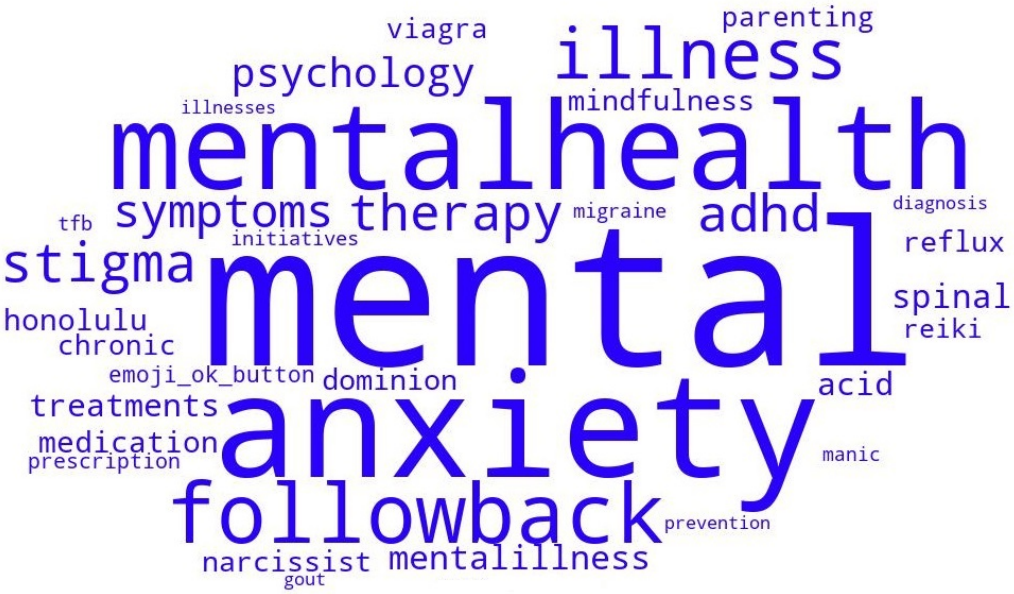


(b) Adultos

Figura 5.17: Top 50 de las palabras usadas por jóvenes y adultos deprimidos en el conjunto de datos de Reddit.



(a) Jóvenes



(b) Adultos

Figura 5.18: Top 50 de las palabras usadas por jóvenes y adultos deprimidos en el conjunto de datos de Twitter.

Hay que recordar que en Reddit algunos usuarios fueron seleccionados por la pertenecía a un subreddit aparte de haber auto-declarado el diagnóstico de la depresión. Esto se ve reflejado por la diferencia de vocabulario que se presenta en Twitter (fi-

gura 5.18) donde los jóvenes hacen uso de *emojis* y hablan sobre temas más diversos (seguidores, videojuegos, relaciones afectivas); por otra parte, los adultos son los que presentan de forma más explicita temas sobre salud mental y enfermedades. Por último, la figura 5.19 muestra palabras comunes en ambas redes sociales[5].

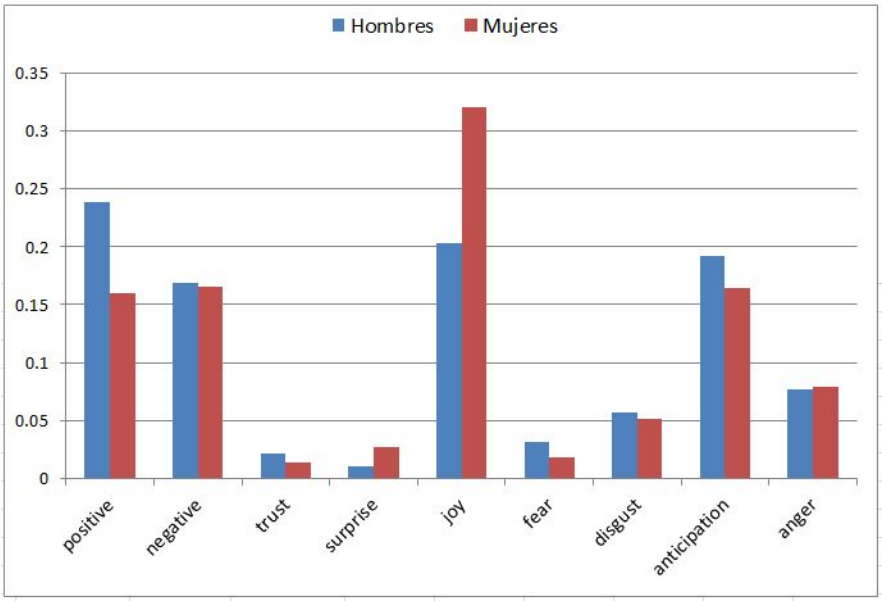
Figura 5.19: Palabras en común para la edad en ambas redes sociales (Reddit y Twitter). Por simplicidad las palabras de jóvenes y adultos se presentan en una sola nube.

5.5.4. Diferencia en el uso de emociones

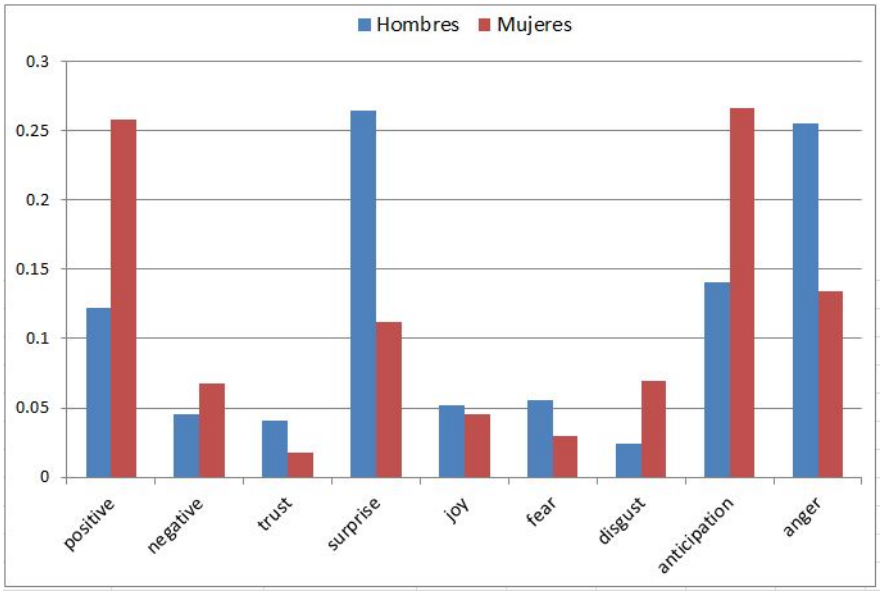
Al igual que las palabras, el uso de emociones para ambos grupos es claramente diferente. Los hombres deprimidos en el conjunto de datos de Reddit expresan más las emociones *positivas* y de *anticipación*, mientras que las mujeres deprimidas parecen usar más palabras de *alegría*. Por otro lado, los hombres deprimidos en el conjunto de datos de Twitter expresan más emociones de *sorpresa* y *enfado* que las mujeres, y las mujeres deprimidas usan palabras más *positivas* cuando se expresan.

En el análisis de emociones para la edad en Reddit la diferencia que existe entre jóvenes y adultos deprimidos no es muy visible. La diferencia sutil que se puede observar en los jóvenes es que se expresan con más *enfado* y los adultos con más palabras *positivas*; sin observar estas diferencias, las emociones que más abundan son las de expresiones *positivas, negativas*, de *alegría* y *anticipación*. De forma contraria, en Twitter los jóvenes deprimidos se expresan con palabras más *positivas* en sus publicaciones, mientras que los adultos deprimidos expresan más las emociones de *sorpresa* y *enfado*.

[5]Se obtuvieron las palabras comunes de ambas edades, pero se daba el caso de que en una edad solo se obtenían 4 o 5 palabras y por eso se optó por presentarlas en una sola nube.

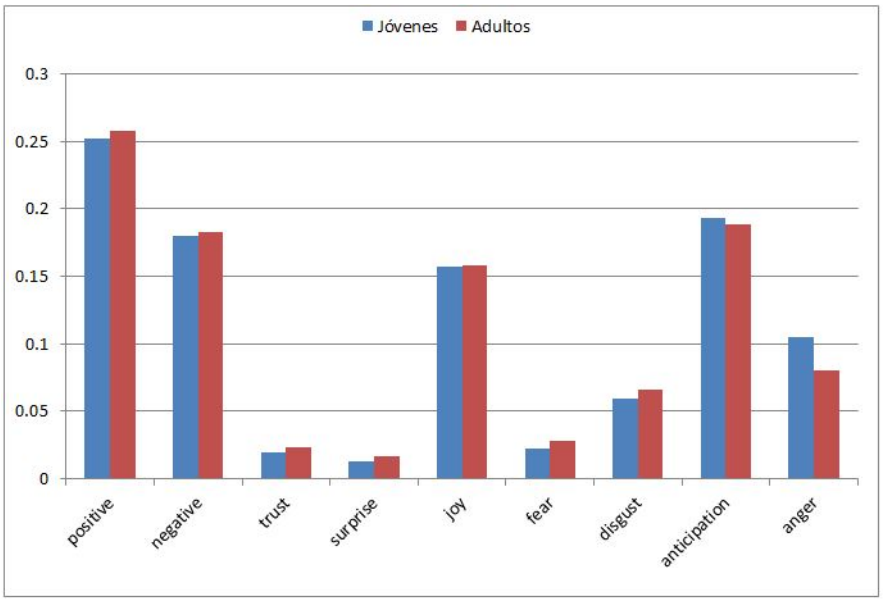


(a) Reddit

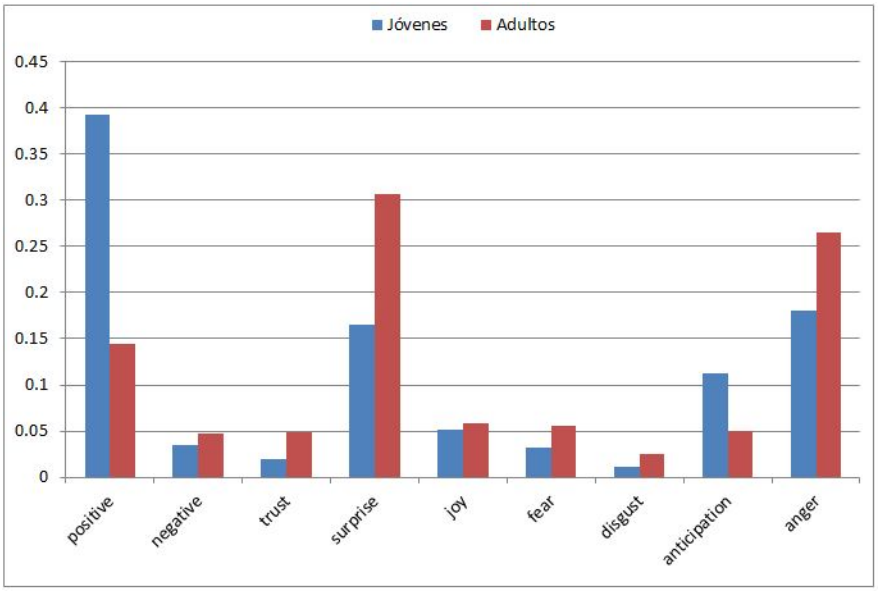


(b) Twitter

Figura 5.20: Histogramas de emociones en hombres y mujeres deprimidos para ambos conjuntos de datos, obtenido de las 100 sub-emociones más frecuentes de la representación BoSE.



(a) Reddit



(b) Twitter

Figura 5.21: Histogramas de emociones en jóvenes y adultos deprimidos para ambos conjuntos de datos, obtenido de las 100 sub-emociones más frecuentes de la representación BoSE.

5.5.5. Utilidad de palabras según su polaridad

Una de las ideas principales del presente trabajo es que las palabras por sí mismas no son suficiente para detectar a usuarios depresivos. Un elemento clave para una

adecuada detección es su contexto ocurrencia, ya sea positivo o negativo.

En las tablas 5.8 y 5.9 se presentan ejemplos de palabras con una diferencia notable de ganancia de información en contexto positivo y negativo para cada tipo de usuario. Estas palabras se obtienen del vocabulario de palabras más frecuentes en hombres y mujeres, jóvenes y adultos; después se escogen las palabras que muestren una diferencia visible de ganancia de información entre ambos contextos.

Reddit					
Hombres			**Mujeres**		
Palabra	**GI - P**	**GI - N**	**Palabra**	**GI - P**	**GI - N**
therapy	**0.0221**	0.0124	relationship	0.0437	**0.0727**
diagnosis	0.0048	**0.0190**	sexual	**0.0162**	0.0072
friendship	**0.0146**	0.0037	attracted	**0.0346**	0.0144
unattractive	0.0037	**0.0068**	makeup	0.0153	**0.0259**
depressive	**0.0301**	0.0153	acne	0.0030	**0.0105**
suicidal	0.0114	**0.0307**	therapist	**0.0746**	0.0259
Twitter					
Hombres			**Mujeres**		
Palabra	**GI - P**	**GI - N**	**Palabra**	**GI - P**	**GI - N**
mental	0.0001	**0.0157**	psychology	**0.0026**	0.0008
anxiety	0.0108	**0.0308**	calories	0.0002	**0.0032**
therapy	**0.0088**	0.0023	illnesses	0.0023	**0.0049**
treatment	**0.0015**	0.0005	manic	0.0017	**0.0057**
pregnancy	**0.0020**	0.0004	hugging	**0.0342**	0.0177
addiction	0.0003	**0.0011**	crying	0.0192	**0.0393**

Tabla 5.8: Ganancia de información (GI) en hombres y mujeres usando uni-gramas de palabras en ambos conjuntos de datos.

Reddit					
Jóvenes			**Adultos**		
Palabra	**GI - P**	**GI - N**	**Palabra**	**GI - P**	**GI - N**
relationship	0.0151	**0.0377**	meds	**0.0470**	0.0316
sexual	**0.0119**	0.0068	mum	**0.0078**	0.0028
feelings	**0.0241**	0.0107	exhausted	0.0097	**0.0130**
dating	**0.0283**	0.0076	overweight	0.0039	**0.0055**
family	0.0078	**0.0143**	therapist	**0.0490**	0.0136
suicidal	0.0312	**0.0538**	cope	0.0085	**0.0280**
Twitter					
Jóvenes			**Adultos**		
Palabra	**GI - P**	**GI - N**	**Palabra**	**GI - P**	**GI - N**
emoji: baby angel light skin tone	**0.0100**	0.0064	anxiety	0.0173	**0.0372**
calories	0.0024	**0.0062**	illness	0.0083	**0.0145**
suicidal	0.0034	**0.0071**	symptoms	0.0045	**0.0104**
psychological	**0.0030**	0.0018	psychology	**0.0021**	0.0003
blessings	**0.0185**	0.0037	fucked	0.0048	**0.0205**
sexuality	**0.0040**	0.0002	mad	0.0095	**0.0251**

Tabla 5.9: Ganancia de información (GI) en jóvenes y adultos usando uni-gramas de palabras en ambos conjunto de datos.

Capítulo 6

Conclusiones y trabajo futuro

En este capítulo se presenta un breve resumen de las ideas presentadas, las conclusiones que se obtuvieron al realizar los experimentos, y algunas ideas del trabajo que se puede desarrollar a futuro.

En la detección de depresión, así como en cualquier tarea de clasificación, la representación de los documentos y los atributos que se utilizan para el proceso de clasificación son de gran importancia para el desempeño de los métodos y los clasificadores. Dicho esto, en el presente trabajo presentamos alternativas para abordar la tarea de detección de depresión, de forma general los métodos propuestos se componen de dos enfoques:

- Exploración de atributos demográficos: se propusieron dos alternativas para el uso de la información de perfil en la detección de usuarios en estado depresivo; la primera consiste simplemente en incluir los atributos de género y edad en la representación del usuario, y la segunda consiste en entrenar clasificadores especializados en género o edad.

- Integración de la polaridad de las publicaciones: se hace una clasificación de las publicaciones de los usuarios según su polaridad (positiva o negativa) para construir una nueva representación que integre esa información. Esta representación se combina con la información de perfil a nivel atributos y usando clasificadores especializados por tipo de usuario.

6.1. Conclusiones

En este trabajo se presentó un estudio sobre la información del perfil de los usuarios que sufren de depresión bajo la idea de que, especializar la clasificación por tipo de usuario, y considerarla junto a la polaridad de sus publicaciones ayuda a mejorar su detección en redes sociales. Respecto a las representaciones usadas, se probaron los métodos antes descritos considerando dos tipos, usamos bolsa de palabras (BoW) para hacer una clasificación orientada a los temas de interés y bolsa de sub emociones (BoSE) para hacer una clasificación basada en las emociones expuestas por los usuarios.

Los experimentos realizados para la exploración de atributos demográficos mostraron que el uso de clasificadores específicos ya sea por género o edad ofrecen mejores resultados que solo incluir esta información como atributos extra a la representación. Esto demostró que las publicaciones hechas por hombres y mujeres o jóvenes y adultos se distinguen por intereses o preocupaciones de su propio grupo (género o edad). Se observó además, que considerar la información de perfil de un usuario provee información relevante y complementaria sobre aquellos usuarios que sufren de depresión. Respecto a las dos representaciones usadas, ambas se ven beneficiadas en el uso de clasificadores específicos, aunque este beneficio es más claro en el género usando BoSE como representación base, donde la diferencia emocional de hombres y mujeres con depresión muestra una diferencia clara.

En cuanto a la construcción de una representación que, incluyera de forma conjunta la información del perfil y la polaridad de las publicaciones, se observó que esta representación es de utilidad y contribuye a mejorar la detección de usuarios que sufren de depresión cuando se usan clasificadores específicos por tipo de usuario. Reflejan además que cuando se combina la idea de usar la información del perfil junto con las polaridades se obtienen mejores resultados. Pensamos que esto se debe principalmente a que los diferentes clasificadores especializados logran capturar con más detalles la información de los usuarios deprimidos y no deprimidos.

Cuando se analizan las diferencias de vocabularios, tanto con género como con edad se obtienen diferencias claras en los temas e intereses que presentan. En las

nubes de palabras se pudieron observar las palabras más frecuentes que los usuarios con depresión usan para compartir experiencias y/o expresarse a través de sus publicaciones. Cabe mencionar que incluso se pueden notar diferencias entre ambos medios sociales (Reddit y Twitter). Reddit, al estar compuesto de subreddits dedicados a un tema en particular, ayuda a que los métodos propuestos sean más efectivos y que las diferencias entre los usuarios deprimidos y no deprimidos puedan estar mejor definidas. Twitter al contrario, al no estar enfocado en un tema en particular, dificulta la detección de depresión; sumado a esto también encontramos la limitación de caracteres y el ruido que se produce en las publicaciones.

6.2. Trabajo futuro

A continuación se presentan algunas ideas que se proponen como trabajo futuro con el fin de aprovechar los métodos propuestos en esta investigación:

- Evaluar los métodos propuestos en conjuntos con mayor cantidad de datos o de distintas redes sociales, principalmente para observar la relación que pueda existir entre género y edad.
- Aprovechar la relación que existe entre el perfil de las personas y la manifestación de las enfermedades mentales, a través de un enfoque de multi-task learning, basados en redes neuronales. El propósito es probar que, dada la relación que existe entre ambos tipos de atributos, es conveniente su aprendizaje conjunto.
- Aprender modelos de atención específicos para cada tipo de usuario, es decir, aprender la representación considerando aspectos del perfil.
- Aplicar un estudio similar en tareas relacionadas con otros desórdenes mentales como la detección de usuarios con anorexia, o incluso en la detección de comentarios agresivos en redes sociales. Esto bajo la premisa de que la manera en que se expresan comentarios agresivos, irónicos, vulgares, es dependiente del contexto, entre ellos el tipo de persona que los produce.

Apéndice

Revisión de la predicción de género

Se hace una revisión manual sobre la predicción de género en ambos conjuntos de datos. Hacer esta revisión nos permite tener certeza sobre qué tan acertada fue la predicción de los datos. Dado que no hay una forma muy clara de cómo evaluar la predicción en la edad solo se hace la revisión sobre el género. En la revisión se buscan palabras o frases en las que se pueda diferenciar ambos géneros y así verificar que la predicción fue correcta. Para seguir manteniendo la anonimidad no se buscan indicios del perfil en la red social de forma deliberada a menos que el usuario haya dejado una referencia o información sobre si mismo a propósito.

La tabla A.1 presenta detalles y resultados sobre la revisión manual del historial de algunos usuarios, esta revisión se hizo sobre historiales del conjunto de datos de Reddit. El tamaño de la población es el número total de usuarios combinando el conjunto de entrenamiento y prueba, el nivel de confianza deseado está en forma de porcentaje (90 %) y el tamaño de la muestra es el número de historiales revisados como correctos o incorrectos al que se desea llegar. La predicción de género fue muy acertada , pues como se puede observar la exactitud fue de 118/129.

Tamaño de la población:	1707
Nivel de confianza (%)	90
Tamaño de la muestra	**129**
Historiales correctos:	**118**
Historiales incorrectos	11

Tabla A.1: Nivel de confianza obtenido en la revisión manual de la predicción de género.

Bibliografía

Angst, J.; Gamma, A.; Gastpar, M.; p Lepine, J.; Mendlewicz, J.; y Tylee, A. T. 2002. Gender differences in depression. epidemiological findings from the european depres i and ii studies. *European archives of psychiatry and clinical neuroscience* 252 5:201–9.

Aragón, M. E.; López-Monroy, A. P.; González-Gurrola, L. C.; y Montes-y Gómez, M. 2019. Detecting depression in social media using fine-grained emotions. In *NAACL-HLT*, 1481–1486.

Baccianella, S.; Esuli, A.; y Sebastiani, F. 2010. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *LREC*.

Blei, D. M.; Ng, A. Y.; y Jordan, M. I. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.* 3:993–1022.

Call, J. B., y Shafer, K. M. 2018. Gendered manifestations of depression and help seeking among men. *American Journal of Men's Health* 12:41 – 51.

Chen, X.; Sykora, M. D.; Jackson, T. W.; y Elayan, S. 2018. What about mood swings: Identifying depression on twitter with temporal measures of emotions. *Companion Proceedings of the The Web Conference 2018.* 1653–1660.

Choudhury, M. D.; Gamon, M.; Counts, S.; y Horvitz, E. 2013. Predicting depression via social media. In *ICWSM*, 128–137.

Coppersmith, G.; Dredze, M.; Harman, C.; Hollingshead, K.; y Mitchell, M. 2015. Clpsych 2015 shared task: Depression and ptsd on twitter. In *CLPsych@HLT-NAACL*, 31–39.

Cortes, C., y Vapnik, V. 1995. Support-vector machine. *Machine Learning* 20:273–297.

Ekman, P. E., y Davidson, R. J. 1994. *The nature of emotion: Fundamental questions*. New York, NY, US: Oxford University Press.

Guntuku, S. C.; Yaden, D. B.; Kern, M. L.; Ungar, L. H.; y Eichstaedt, J. C. 2017. Detecting depression and mental illness on social media: an integrative review. *Current Opinion in Behavioral Sciences* 18:43–49.

Hassan, A. U.; Hussain, J.; Hussain, M.; Sadiq, M.; y Lee, S. 2017. Sentiment analysis of social networking sites (sns) data using machine learning approach for the measurement of depression. *2017 International Conference on Information and Communication Technology Convergence (ICTC)* 138–140.

Hiraga, M. 2017. Predicting depression for japanese blog text. In *ACL*, 107–113.

Jamil, Z.; Inkpen, D.; Buddhitha, P.; y White, K. 2017. Monitoring tweets for depression to detect at-risk users. In *CLPsych@ACL*, 32–40.

Losada, D. E., y Crestani, F. 2016. A test collection for research on depression and language use. In *CLEF*.

Losada, D. E.; Crestani, F.; y Parapar, J. 2018. Overview of erisk: Early risk prediction on the internet (extended lab overview). In *CLEF*.

Loveys, K.; Torrez, J.; Fine, A.; Moriarty, G.; y Coppersmith, G. 2018. Cross-cultural differences in language markers of depression online. In *CLPsych@NAACL-HTL*, 78–87.

McCrae, N.; Gettings, S.; y Purssell, E. 2017. Social media and depressive symptoms in childhood and adolescence: A systematic review. *Adolescent Research Review* 2:315–330.

Mowery, D. L.; Park, A.; Bryan, C. J.; y Conway, M. 2016. Towards automatically classifying depressive symptoms from twitter data for population health. In *PEOPLES@COLING*, 118–125.

Nolen-Hoeksema, S. 2001. Gender differences in depression. *Current Directions in Psychological Science* 10:173–176.

Olson, D., y Delen, D. 2008. Advanced data mining techniques. *Springer*. 138.

O.M.S. 2013. Plan de acción integral sobre salud mental 2013-2020. *Technical report*.

OPS, y OMS. 2014. Plan de acción sobre salud mental 2015-2020. *Technical report*.

Ortega-Mendoza, R. M.; López-Monroy, A. P.; Franco-Arcega, A.; y y Gómez, M. M. 2018. Emphasizing personal information for author profiling: New approaches for term selection and weighting. *Knowl. Based Syst.* 145:169–181.

Pennebaker, J. W.; Mehl, M. R.; y Niederhoffer, K. 2003. Psychological aspects of natural language. use: our words, our selves. *Annual review of psychology* 54:547–77.

Preotiuc-Pietro, D.; Eichstaedt, J. C.; Park, G. J.; Sap, M.; Smith, L.; Tobolsky, V.; Schwartz, H. A.; y Ungar, L. H. 2015. The role of personality, age, and gender in tweeting about mental illness. In *CLPsych@HLT-NAACL*, 21–30.

Prince, M. U.; Patel, V.; Saxena, S.; Maj, M.; y Rahman, A. 2007. No health without mental health. *The Lancet* 370:859–877.

Reece, A. G.; Reagan, A. J.; Lix, K. L. M.; Dodds, P. S.; Danforth, C. M.; y Langer, E. J. 2016. Forecasting the onset and course of mental illness with twitter data. *Scientific Reports* 7:e12948 – a33.

Resnik, P.; Armstrong, W.; Claudino, L.; y Nguyen, T. 2015. The university of maryland clpsych 2015 shared task system. In *CLPsych@HLT-NAACL*, 54–60.

Resnik, P.; Garron, A.; y Resnik, R. 2013. Using topic modeling to improve prediction of neuroticism and depression in college students. In *EMNLP*, 1348–1353.

Sadeque, F.; Pedersen, T.; Solorio, T.; Shrestha, P.; Rey-Villamizar, N.; y Bethard, S. 2016. Why do they leave: Modeling participation in online depression forums. In *SocialNLP@EMNLP*, 14–19.

Sap, M.; Park, G. J.; Eichstaedt, J. C.; Kern, M. L.; Stillwell, D.; Kosinski, M.; Ungar, L. H.; y Schwartz, H. A. 2014. Developing age and gender predictive lexica over social media. In *EMNLP*, 1146–1151.

Schwartz, H. A.; Eichstaedt, J. C.; Kern, M. L.; Park, G. J.; Sap, M.; Stillwell, D.; Kosinski, M.; y Ungar, L. H. 2014. Towards assessing changes in degree of depression through facebook. In *CLPsych@ACL*, 118–125.

Shen, G.; Jia, J.; Nie, L.; Feng, F.; Zhang, C.; Hu, T.; Chua, T.-S.; y Zhu, W. 2017. Depression detection via harvesting social media: A multimodal dictionary learning solution. In *IJCAI*, 3838–3844.

Tian, L.; Lai, C.; y Moore, J. D. 2018. Polarity and intensity: the two aspects of sentiment analysis. *ArXiv* abs/1807.01466.

Trotzek, M.; Koitka, S.; y Friedrich, C. M. 2018. Word embeddings and linguistic metadata at the clef 2018 tasks for early detection of depression and anorexia. In *CLEF*.

Trotzek, M.; Koitka, S.; y Friedrich, C. 2020. Utilizing neural networks and linguistic metadata for early detection of depression indications in text sequences. *IEEE Transactions on Knowledge and Data Engineering* 32:588–601.

Wang, S. I., y Manning, C. D. 2012. Baselines and bigrams: Simple, good sentiment and topic classification. In *ACL*, 90–94.

Wolohan, J.; Hiraga, M.; Mukherjee, A.; Sayyed, Z. A.; y Millard, M. 2018. Detecting linguistic traces of depression in topic-restricted text: Attending to self-stigmatized depression with nlp. In *Proceedings of the First International Workshop on Language Cognition and Computational Models*, 11–21.

Yan, J. 2009. *Text Representation*. 3069–3072.

Yates, A.; Cohan, A.; y Goharian, N. 2017. Depression and self-harm risk assessment in online forums. In *EMNLP*, 2968–2978.

Zhang, Y., y Wallace, B. 2017. A sensitivity analysis of (and practitionersǵuide to) convolutional neural networks for sentence classification. In *IJCNLP*, 253–263.