Construcción de corpus en español para la Identificación de sentimientos en Twitter

Construction of a corpus in Spanish for the identification of feelings on Twitter

Wendy Morales Castro[1], Gustavo Basurto Islas[2], Rafael Guzmán Cabrera[1]
[1] División de Ingenierías, Campus Irapuato-Salamanca, Universidad de Guanajuato.
[1] División de Ciencias e Ingenierías, Campus Leon, Universidad de Guanajuato.
w.moralescastro@ugto.mx
gustavo.basurto@ugto.mx
guzmanc@ugto.mx

Resumen

La prevalencia de personas con depresión ha aumentado en los últimos años, en parte debido a una mayor atención en la detección de enfermedades psiquiátricas. El propósito del presente trabajo es proponer una estrategia para detectar la enfermedad haciendo uso de la Inteligencia Artificial, tener mayores registros de su prevalencia y con ello exponer su impacto a nivel global. Dentro del presente trabajo se comenzará con la definición de la depresión, los principales factores, así como la sintomatología, a su vez también se hablará del uso de las redes sociales y como pueden ser utilizadas para la detección de depresión con la ayuda de la Inteligencia Artificial, específicamente el campo de Procesamiento de Lenguaje Natural.

Palabras clave: Depresión, Inteligencia Artificial, Procesamiento de Lenguaje Natural, Máquinas de Vectores de soporte, Regresión Logística, J48, Validación Cruzada, Conjuntos de Entrenamiento y Prueba.

Introducción

La depresión, según la Organización Mundial de la Salud (OMS), es un trastorno mental afectivo común a nivel mundial y es tratable. Está caracterizado por el cambio de ánimo con síntomas cognitivos y físicos. La depresión implica un estado de ánimo deprimido o la pérdida de placer o el interés por actividades durante largos periodos de tiempo y estos pueden ser de etiología primaria o secundaria al encontrarse enfermedades subyacentes, como la diabetes, cáncer, Parkinson, trastornos alimenticios o incluso el abuso de sustancias. Aunque la depresión y otros trastornos mentales son enfermedades muy frecuentes en todo el mundo, no se les da la importancia que debería, en concreto, la depresión puede convertirse en un problema de salud serio, si persiste por un largo periodo de tiempo, con una intensidad moderada o grave puede alterar las actividades laborales, escolares, sociales o familiares de la persona que la padece, y en casos muy extremos puede llevar al suicidio (Fernández-Berrocal et al., 2003; Retamal, 1998).

Al ser una enfermedad a la que no se le da la importancia necesaria, se vuelve difícil para las personas que la padecen pedir ayuda profesional, esto debido a cierto estigma que aún existe en la sociedad hacia este tipo de enfermedades. Algunos métodos utilizados por expertos para la detección de síntomas normalmente constan de encuestas o preguntas de carácter muy personal, y en consecuencia a esto los pacientes no acceden a someterse a dicha situación, o responden de forma errónea a estos métodos, lo cual afectaría en la veracidad de los síntomas (Molina & Martí, 2010). Por tal motivo, se ha recurrido al uso de la Inteligencia Artificial para apoyar a la identificación de depresión, mediante el uso de Procesamiento de Lenguaje Natural.

Resulta complicado definir lo que es la Inteligencia Artificial (IA), debido a que es un tema complejo, se pueden encontrar un sinfín de definiciones de ella, la definición comúnmente más usada es que la IA es la habilidad de los ordenadores para realizar actividades que normalmente realizamos los humanos, una definición más detalla podría indicar que la IA es la capacidad que tienen las máquinas para usar algoritmos, aprender de los datos y utilizar lo aprendido de esto para realizar una toma de decisiones como normalmente lo haría un humano. Dentro de la IA se tienen diversos campos, el campo en el que se centra el presente trabajo de investigación es el de Procesamiento de Lenguaje Natural por sus siglas (PLN), este campo lo podemos definir como la habilidad que tiene una máquina para procesar la información comunicada, entendemos que existen diversos tipos de comunicación; este estudio se centra en la comunicación escrita. El PLN actualmente ha adquirido mayor relevancia por la gran cantidad de información que se genera en internet, la

cual resulta relevante en diversos ámbitos como su uso dentro de la organización para evaluar servicios o productos (Joyanes Aguilar, 2013), o su uso dentro del área social (Garcés Chaparro, 2019), para conocer opiniones o comentarios de las personas acerca de algún tema (Caicedo Ortiz, 2016) en el área médica, como la identificación de depresión, en la cual se tiene poca información relevante en el idioma español, algunos trabajos encontrados en el área enfocados en el idioma español son basados en cuestionarios como la implementación de chatboots (Arrabales, 2008.) u otro tipo de herramientas como los trabajos de (Del Pilar et al., 2018; Coll-Florit et al., 2020; Leis et al., 2019; Ramirez-Esparza et al., 2008), los cuales presentan algunas problemáticas como la mencionada con anterioridad, lo que conlleva a la reducción de la veracidad de los resultados, o el sesgo producido por la traducción de los datos. Es importante mencionar que esta tarea ha adquirido mayor auge en países como Estados Unidos, Reino Unido y China (Colombo-Mendoza et al., 2020), pero debido a que cada idioma tiene sus especificidades es importante darles a las oraciones el tratamiento correcto para entender el contexto de estas.

Para extraer este tipo de información, diversos investigadores se han enfocado en el uso de las redes sociales como Twitter, Facebook, Instagram, blogs, foros, etc., debido a que son plataformas que contienen grandes cantidades de usuarios y dentro de las cuales los usuarios realizan post, comentarios u opiniones de su vida cotidiana. Es por ello por lo que en el presente trabajo se usará un conjunto de corpus extraídos previamente de Twitter, los cuales han sido etiquetados por expertos con la finalidad de entrenar y probar la metodología propuesta.

Metodología Propuesta

En el siguiente diagrama se presenta la metodología llevaba a cabo con este corpus.

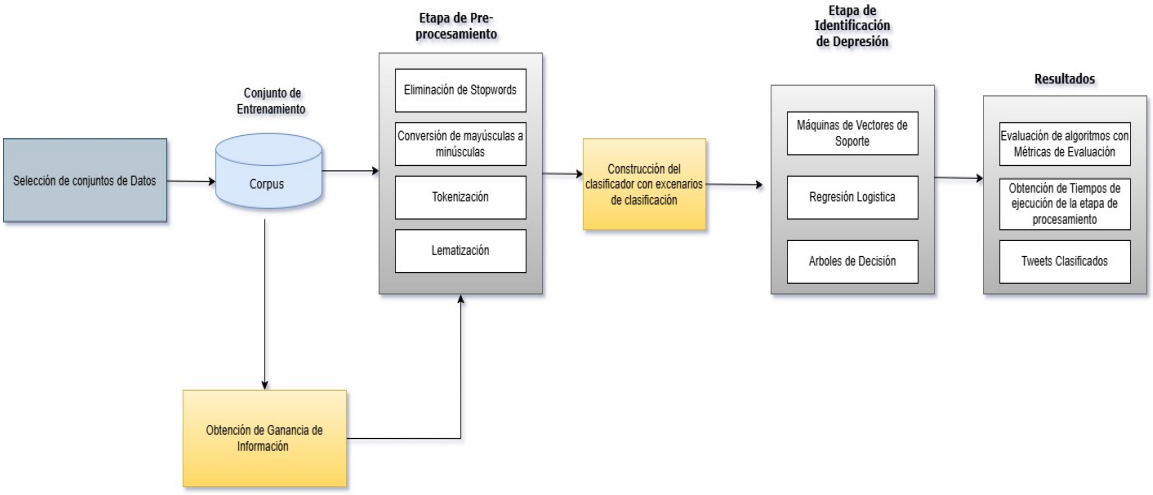

Figura 1. Metodología Propuesta. Fuente Autoría propia.

Para llevar a cabo la metodología propuesta se creó un nuevo corpus, con la unión de dos diferentes corpus etiquetados con anterioridad por expertos en el área, el primero de ellos llamado "conjunto de datos de usuarios depresivos[1] ", los tweets fueron creados por 90 usuarios, los cuales fueron analizados por expertos y etiquetados con la leyenda "depresión", por el contenido del tweet, contiene un total de 1000 tweets, el otro conjunto de datos llamado "Análisis del sentimiento de Twitter en los Tweets en español [2]" fue incluido dentro del presente trabajo por el enfoque y el contenido distinto a los anteriores tweets con la finalidad de tener una clase distinta y poder colocar la etiqueta "no depresivos", este corpus contenía 2590 tweets, de los cuales se usaron solo 1000 para tener un corpus balanceado.

[1]Spanish tweets suggesting depression
[2]Twitter Sentiment Analysis in Spanish Tweets

Con la finalidad de observar el comportamiento de los datos, se calculará la ganancia de información (IG) para evaluar también la efectividad de la metodología empleada. La obtención de IG permitirá reducir la dimensionalidad de la matriz de confusión generada durante el preprocesamiento, considerando únicamente los atributos más relevantes de los tweets. Esto facilitará la obtención de resultados distintos a los generados en el corpus anterior.

Para seguir con la metodología propuesta una vez que el corpus fue creado y obtenida también la IG el siguiente paso será preprocesarlo para ello se usarán técnicas de PLN (S. González et al., 2023) que nos permitan realizar una limpieza al corpus, entre las técnicas realizadas encontramos:

- Eliminación de StopWords.
- Conversión de mayúsculas a minúsculas.
- Lematización.
- Tokenización.

Una vez que los datos fueron preprocesados son divididos en dos escenarios de clasificación los cuales serán Validación Cruzada (Pérez-Planells et al., 2015) con 10 folds y Conjuntos de Entrenamiento y Prueba (Águila et al., 2017) con un 80% a entrenamiento y un 20% de prueba, el siguiente paso es llevar a cabo el procesamiento de los datos, una vez realizados los escenarios serán uno a uno evaluados por los siguientes clasificadores:

- Máquinas de Vectores de Soporte (SVM) (C. González et al., 2011.).
- Regresión Logística (Kleinbaum et al., 2002).
- J48 (Bhargava et al., 2013).

Terminando la etapa de procesamiento, se continua con la etapa de resultados en la cual se obtendrá la métrica de evaluación llamada precisión y a su vez se estarán obteniendo cada uno de los tweets con su respectiva etiqueta de clasificación.

Resultados.

Se analizaron diversas metodologías con la finalidad de determinar la variabilidad en cuanto a la precisión obtenida en los resultados, para ello la metodología se dividió en 3 experimentos los cuales se describen a continuación y de los cuales también se presentan los resultados obtenidos.

Experimento 1. Baseline.

Para este experimento no se eliminaron las StopWords, solamente se prepararon los datos, para no realizar otro tipo de modificación al contenido del corpus, y observar cómo se comportaba el corpus de forma “cruda”. Con el corpus preparado se crearon los dos escenarios de clasificación Validación Cruzada y Conjuntos de Entrenamiento y Prueba y se usaron los 3 algoritmos de clasificación mencionados en la metodología, para finalmente obtener el tiempo de ejecución del procesamiento y la precisión de esta evaluación, estos resultados se muestran en la tabla 1.

Como datos estadísticos del corpus se obtuvieron las palabras distintas del corpus es decir las palabras distintas encontradas en todos los tweets del corpus y el vocabulario del corpus es decir todas las palabras existentes y sus respectivas frecuencias dentro del corpus, con la finalidad de observar como va variando esto a lo largo de los experimentos.

Tabla 1. Baseline

Métodos de Aprendizaje.	Validación Cruzada	Conjuntos de Entrenamiento y Prueba	Vocabulario	Palabras Distintas	Tiempo de ejecución VC	Tiempo de ejecución CEYP
SVM	93.65%	93%	35675	5795	33.78 s	0.06 s
J48	77.95%	75.75%	35675	5795	30.58 s	0.02 s
KNN	70.15%	69.5%	35675	5795	4.95 s	0.06 s

En este caso los mejores resultados fueron obtenido por el clasificador SVM.

Experimento 2. Eliminación de StopWords.

Para este experimento, si se eliminaron las StopWords, con la finalidad de observar el cambio en cuanto a la precisión obtenida con el experimento anterior, y a su vez observar el cambio que presenta dentro del tamaño del vocabulario, de igual forma, se obtuvieron las mismas características como las palabras distintas del corpus, el vocabulario del corpus y tiempos de ejecución. Se usaron los mismos dos escenarios de clasificación y los mismos 3 algoritmos de clasificación mencionados dentro de la metodología. Finalmente se obtendrá la precisión de esta evaluación, estos resultados se muestran en la tabla 2.

Tabla 2. Eliminación de StopWords.

Métodos de Aprendizaje.	Validación Cruzada	Conjuntos de Entrenamiento y Prueba	Vocabulario	Palabras Distintas	Tiempo de ejecución VC	Tiempo de ejecución CEYP
SVM	93.75	92.75	17579	5631	25.8	0.09
J48	73.8	72	17579	5631	20.3	0.02
KNN	71.65	70	17579	5631	15.07	0.27

Para este caso los mejores resultados también fueron obtenidos por el clasificador SVM.

Experimento 3. Con eliminación de StopWords y IG.

En este experimento se empleó la Ganancia de información como entrada a la etapa de preprocesamiento, y se estuvieron eliminando las StopWords para hacer uso de una información más limpia para llevar a cabo el procesamiento de los datos y observar cómo esto afectaba en tiempos de ejecución, tamaño de vocabulario y tamaño de palabras distintas, los resultados a este experimento se presentan a continuación.

Tabla 3. Con eliminación de StopWords y IG.

Métodos de Aprendizaje.	Validación Cruzada	Conjuntos de Entrenamiento y Prueba	Vocabulario	Palabras Distintas	Tiempo de ejecución VC	Tiempo de ejecución CEYP
SVM	97.1	96.75	6586	765	5.75	0.01
J48	73.5	72.75	6586	765	6.84	0.01
KNN	83.25	80	6586	765	0.06	1.28

Como se observa también los mejores resultados se obtuvieron con SVM.

Conclusiones.

Dentro de los resultados de los experimentos se pudo observar lo siguiente:

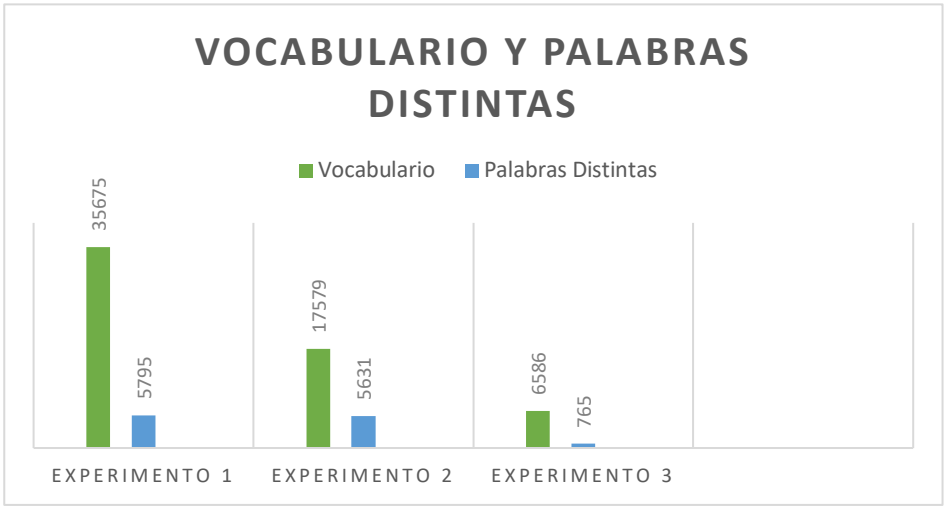

Figura 2. Gráfica de Vocabulario y Palabras distintas. Fuente Autoría propia.

Como se observa en la gráfica, en cada experimento el tamaño del vocabulario y de palabras distintas disminuyendo progresivamente debido a los tipos de preprocesamiento aplicados. Estos procedimientos de preprocesamiento impactaron los resultados obtenidos, tanto en términos de precisión como en el tiempo de ejecución de cada experimento. Tomando en cuenta la precisión mostrada en las tablas de resultados se observó el aumento gradual a lo largo de los experimentos, teniendo un aumento del 3.45% entre el primer experimento y el último experimento, con base al mejor resultado obtenido de cada experimento, y una disminución de 28 segundos con 3 milisegundos, en cuanto al tiempo de ejecución del resultado con mejor precisión del experimento 1, contra el mejor resultado de precisión obtenido del experimento 3. Con estos datos se puede deducir que el realizar y variar el tipo de tratado de datos dentro del preprocesamiento puede ayudar a la optimización de los resultados de clasificación y el tiempo de ejecución. A su vez el experimentar con diversos algoritmos de clasificación puede ayudar a discriminar cuales funcionan mejor que otros en cuanto a la identificación de depresión en tweets. Lo cual resulta relevante para que expertos en el área como psicólogos o psiquiatras puedan tener herramientas que ayuden con la identificación de la enfermedad y asegurarse de realizar un diagnóstico médico correcto y seguir con su respectivo tratamiento médico.

Referencias

Águila, J., Trabajo, M., & Solanas Gómez, A. (2017). *Aprendizaje supervisado en conjuntos de datos no balanceados con redes neuronales artificiales: métodos de mejora de rendimiento para modelos de clasificación*. https://openaccess.uoc.edu/handle/10609/64768

Arrabales, R. (2008). *Perla: un Agente Conversacional para la Detección de Depresión en Ecosistemas Digitales. Diseño, Implementación y Validación*. Retrieved November 2, 2022, from https://arxiv.org/abs/2008.12875]

Bhargava, N., Sharma, G., Bhargava, R., & Mathuria, M. (2013). Decision Tree Analysis on J48 Algorithm for Data Mining. *International Journal of Advanced Research in Computer Science and Software Engineering*, *3*(6), 2277. http://www.cs.waikato.ac.nz/ml/weka.

Caicedo Ortiz. (2016). *Análisis del sentimiento político mediante la aplicación de herramientas de minería de datos a través del uso de redes sociales*. https://repository.javeriana.edu.co/handle/10554/20516

Coll-Florit, M., Climent, S., Sanfilippo, M., & Hernández-Encuentra, E. (2020). Metáforas de la depresión. Estudiar relatos en primera persona de la vida con depresión publicados en blogs. *Taylor & FrancisM Coll-Florit, S Climent, M Sanfilippo, E Hernández-EncuentraMetáfora y Símbolo, 2021•Taylor & Francis*, *36*(1), 1–19. https://doi.org/10.1080/10926488.2020.1845096

Colombo-Mendoza, A. E., Luis Sánchez-Cervantes, J., Alor-Hernández, G., Rodríguez-Mazahua, L., Del, M., & Salas-Zarate, P. (2020). Análisis comparativo de herramientas y APIs para la identificación y detección de depresión. *Congreso Estudiantil de Inteligencia Artificial Aplicada a La Ingeniería y Tecnología, UNAM*.

Del Pilar, M., Zárate, S., & Rodríguez González, D. A. (2018). Detección de Patrones Psicolingüísticos para el Análisis de Lenguaje Subjetivo en Español. *Procesamiento Del Lenguaje Natural*, *60*(0), 79–82. https://doi.org/10.26342/2018-60-10

Fernández-Berrocal, P., … N. E.-E. en psicología, & undefined. (2004). Inteligencia emocional y depresión. *Academia.Edu*. Retrieved October 25, 2024, from https://www.academia.edu/download/32027296/PDF12depresion.pdf

Garcés Chaparro, T. I. (2019). *Análisis de sentimientos en redes sociales orientado a la percepción de la calidad de servicios de internet, redes móviles, tv cable y electricidad*. https://repositorio.unab.cl/xmlui/handle/ria/17963

González, C., Sistemas, G. de S. I.-G. de, & undefined. (2021). SVM: Máquinas de vectores soporte. *Infor.Uva.Es*. Retrieved February 13, 2024, from http://www.infor.uva.es/~calonso/MUI-TIC/MineriaDatos/SVM.pdf

González, S., Reyes, J., Cedeño, D., & Azcuy, R. (2023). *Herramienta de PNL para la detección de ambigüedades en requisitos de software escritos en español NLP tool for the detection of ambiguities in software*. https://easychair.org/publications/preprint_download/RgRDR

Joyanes Aguilar, L. (2013). Big Data: Análisis de grandes volúmenes de datos en organizaciones. *Journal of Chemical Information and Modeling*, *1*, 428.

Kleinbaum, D., Dietz, K., Gail, M., Klein, M., & Klein, M. (2002). *Logistic regression*. https://link.springer.com/content/pdf/10.1007/978-1-4419-1742-3.pdf

Leis, A., Ronzano, F., Mayer, M. A., Furlong, L. I., & Sanz, F. (2019). Detecting Signs of Depression in Tweets in Spanish: Behavioral and Linguistic Analysis. *Journal of Medical Internet Research*, *21*(6). https://doi.org/10.2196/14199

Molina, L. S., & Martí, B. (2010). *Comprender la depresión*. https://books.google.com.mx/books?hl=es&lr=&id=yC_1xY4jzNUC&oi=fnd&pg=PA5&dq=depresion+&ots=nLhEHeUPHV&sig=VxZbXhyiT1ljZ9djZk3KWcZtZ1A

Pérez-Planells, L., Delegido, J., … J. R.-C.-R., & undefined. (2015). Análisis de métodos de validación cruzada para la obtención robusta de parámetros biofísicos. *Ojs.Upv.Es*. Retrieved February 13, 2024, from http://ojs.upv.es/index.php/raet/article/view/4153

Ramirez-Esparza, N., Chung, C., … E. K.-P. of the, & undefined. (2008). La psicología del uso de palabras en foros de depresión en inglés y en español: Probando dos enfoques analíticos de texto. *Ojs.Aaai.OrgN Ramirez-Esparza, C Chung, E Kacewic, J PennebakerActas de La Conferencia Internacional AAAI Sobre Web y Redes Sociales, 2008•ojs.Aaai.Org*. https://ojs.aaai.org/index.php/ICWSM/article/view/18623

Retamal, P. (1998). *Depresión*. https://books.google.com.mx/books?hl=es&lr=&id=1kwVmA7st_cC&oi=fnd&pg=PA3&dq=depresion+&ots=7RTaQyxX5-&sig=f7pI54y3ddHyfMuOQ94icQLCIng