

HERRAMIENTA PARA DETECCIÓN DE ENFERMEDADES DE SALUD MENTAL

JOSÉ LUIS LEIVA, EDUARDO GARCÍA MALETA, JESÚS ALDAIR ALFONSO

CONTENTS

1	Estado del Arte	2
2	Introducción	2
3	Desarrollo	3
3.1	Datos y Procesamiento	3
3.2	Enfoque No Supervisado Utilizando KMeans	3
4	Aprendizaje No Supervisado	3

Abstract

Este artículo presenta un estudio comprensivo sobre el análisis de enfermedades de salud mental en redes sociales utilizando técnicas de aprendizaje automático. Exploramos diversas metodologías para la extracción de características, incluyendo el uso del léxico Empath para cuantificar los tonos emocionales en contenido generado por usuarios. En nuestro análisis se emplea herramientas de aprendizaje automático, tanto supervisado como no supervisado; evaluando el rendimiento de diferentes modelos a través de validación cruzada estratificada, proporcionando información sobre sus capacidades predictivas. Los modelos utilizados en este trabajo fueron K-Means y Redes Neuronales.

PALABRAS CLAVE :

Análisis de Sentimientos, Clasificación de Enfermedades de Salud Mental, Aprendizaje Automático, Redes Sociales, Extracción de Características, KMeans, Redes Neuronales.

CONTENTS

1 ESTADO DEL ARTE

2 INTRODUCCIÓN

En la última década, el uso de redes sociales ha crecido exponencialmente, convirtiéndose en una plataforma fundamental para la expresión personal y la interacción social. Este fenómeno ha generado un vasto volumen de datos que pueden ser analizados para obtener información valiosa sobre diversos aspectos de la vida cotidiana, incluida la salud mental. La clasificación de enfermedades de salud mental a partir de comentarios en redes sociales representa un desafío significativo, dado que los datos son inherentemente ruidosos y subjetivos.

El presente trabajo se centra en la aplicación de técnicas de aprendizaje automático para abordar este problema. A diferencia de estudios previos que han predominado en el uso de enfoques de aprendizaje supervisado, como Máquinas de Vectores de Soporte (SVM), Árboles de Decisión y Naive Bayes, nuestro enfoque inicial se basa en el aprendizaje no supervisado. Esta elección se fundamenta en la escasez de investigaciones que exploren esta metodología en el contexto específico del análisis de comentarios en redes sociales relacionados con la salud mental. Al emplear técnicas no supervisadas, buscamos evaluar su eficacia para identificar patrones y agrupaciones en los datos sin la necesidad de etiquetas predefinidas.

Posteriormente, complementamos nuestro análisis con un enfoque supervisado utilizando redes neuronales. A pesar de su creciente popularidad en el campo del aprendizaje automático, encontramos que su aplicación específica a la clasificación de enfermedades mentales a partir de datos extraídos de redes sociales es limitada en la literatura existente.

A través de este trabajo, esperamos proporcionar una visión integral sobre cómo algunas de las herramientas de aprendizaje automático pueden ser utilizadas para abordar problemas complejos en el ámbito de la salud mental, abriendo nuevas vías para futuras investigaciones y aplicaciones prácticas.

3 DESARROLLO

3.1 Datos y Procesamiento

Los datos utilizados en este estudio provienen de un conjunto de datos disponible en línea, el cual está catalogado en siete clases diferentes: Ansiedad, Depresión, Normal, Suicida, Estrés, Bipolar y Desorden de la Personalidad. Cada entrada del conjunto de datos está estructurada en dos columnas: una que contiene el texto y otra que proporciona la etiqueta correspondiente que describe la patología asociada al texto.

El primer paso en nuestro proceso de análisis fue la limpieza de los datos. Esto incluyó la eliminación de valores NaN (Not a Number) para asegurar la integridad del conjunto de datos.

Una vez limpiados los datos, utilizamos **Empath**, que es una herramienta diseñada para caracterizar y cuantificar aspectos psicológicos del texto. Esta nos permitió extraer características relevantes que reflejan el contenido emocional y psicológico de los comentarios, facilitando así un análisis más profundo.

Posteriormente, aplicamos una matriz de correlación para identificar y eliminar datos redundantes dentro de nuestro conjunto. Esta técnica es esencial para reducir la dimensionalidad del dataset y mejorar la eficiencia del modelo, al eliminar características que no aportan información adicional o que están altamente correlacionadas con otras. Al final de este proceso, obtuvimos un conjunto de datos más limpio y representativo, listo para ser utilizado en nuestros modelos de aprendizaje automático.

3.2 Enfoque No Supervisado Utilizando KMeans

Para abordar la clasificación de enfermedades de salud mental, como primer enfoque utilizando el algoritmo KMeans. Este algoritmo es ampliamente reconocido por su capacidad para agrupar datos en clústeres basados en similitudes, y en nuestro caso, se le indicó que clusterizara los datos en siete clases, alineándose con las categorías del conjunto de datos original.

Sin embargo, los resultados obtenidos no fueron satisfactorios. Al analizar la distribución de los clústeres generados, observamos que una de las clases contenía únicamente dos elementos, mientras que otra clase agrupaba 99 elementos. Dada la naturaleza del conjunto de datos, que cuenta con miles de entradas por clase, esta distribución indica un desempeño deficiente del modelo.

La presencia de un clúster con tan pocos elementos sugiere que el algoritmo no pudo identificar patrones significativos dentro de los datos para esa clase específica. Asimismo, la clase con 99 elementos podría estar representando una agrupación inadecuada, lo que implica que el modelo no logró capturar la complejidad del conjunto de datos.

En conclusión, estos resultados nos llevan a la reflexión sobre la idoneidad del enfoque no supervisado en este contexto particular. A pesar de las expectativas iniciales, los resultados obtenidos no cumplen con los estándares necesarios para una clasificación efectiva y nos motivan a explorar enfoques alternativos.