Descifrando el lenguaje emocional en Twitter: Un análisis predictivo basado en aprendizaje automático.

Autor: Neivys Luz González Gómez

1. Introducción

En la era actual de las redes sociales, Twitter se ha convertido en una de las plataformas más utilizadas en todo el mundo, donde millones de usuarios publican diariamente sus pensamientos, emociones y experiencias. Esta enorme cantidad de datos proporciona una fuente rica de información para analizar las tendencias y patrones de los usuarios.

La detección de emociones a partir de textos es uno de los desafíos más complejos en el procesamiento del lenguaje natural debido a la naturaleza multiclase del problema y la falta de conjuntos de datos etiquetados. Sin embargo, en este estudio se cuenta con un conjunto de datos ya etiquetado para la detección de emociones en tweets.

En este contexto, el objetivo de este estudio es desarrollar un modelo que permita detectar emociones en los tweets con el fin de ayudar en la detección temprana de trastornos emocionales, como la depresión y la ansiedad. Al lograr este objetivo, se puede proporcionar información valiosa para los profesionales de la salud mental y ayudar en la toma de decisiones clínicas tempranas.

Para alcanzar el objetivo, se utilizará la biblioteca Twint de Python para recolectar datos de Twitter y se recopilará información pública de data.world. Posteriormente, se aplicarán técnicas de procesamiento de lenguaje natural para limpiar y pre-procesar los datos. Después, se utilizarán distintos algoritmos de aprendizaje automático para entrenar un modelo capaz de detectar patrones en los datos y predecir el tipo de emoción en los tweets de un usuario de Twitter.

2. Información de la base dato.

Para la elaboración del proyecto se utilizarán dos bases de datos:

2.1. Tweet emotion: Se dispone de un conjunto de datos proveniente de data.world, una plataforma que proporciona acceso a conjuntos de datos públicos. Este dataset consiste en una colección de tweets etiquetados con la emoción que expresan. Contiene cuatro columnas que incluyen el identificador del tweet, el sentimiento expresado, el autor y el contenido del tweet. En total, se tienen 40,000 registros con

anotaciones para 13 emociones distintas. A continuación la descripción de las variables o atributos de la base de dato:

- **tweet_id:** corresponde al identificador numérico único asignado a cada registro de tweet en el conjunto de datos.
- **sentiment:** se trata de una variable de cadena que clasifica cada tweet en una de las 13 diferentes emociones.
- **author:** indica el nombre del usuario de Twitter que publicó el tweet. Es una variable de cadena.
- **content:** este campo contiene el texto completo del tweet y es una variable de cadena.

Estos campos son fundamentales para el estudio de detección de emociones en los tweets y serán utilizados para desarrollar nuestro modelo predictivo.

2.2. Tweet data 2023: Para llevar a la recolección de datos, se utilizará la librería Twint, una herramienta de web scrapping de Twitter que permite obtener tweets públicos sin necesidad de una API. La ventaja de utilizar Twint radica en su capacidad de obtener información de usuarios que no se encuentran disponibles mediante las APIs, así como de aquellos usuarios que han eliminado o bloqueado sus cuentas. Esta base de dato se utilizará para el test final del modelo.

A continuación la descripción de las variables o atributos de la base de dato:

- **user_id:** es una variable numérica (int64) que identifica de manera única a cada usuario de Twitter que publicó el tweet en cuestión. Este valor es asignado automáticamente por Twitter al crear una cuenta y se puede utilizar para distinguir entre diferentes usuarios.
- **username:** es una variable de tipo string (object) que representa el nombre de usuario de Twitter asociado con el user_id. Es un nombre elegido por el usuario y puede ser utilizado para hacer referencia a ellos en la plataforma.
- **tweet:** es una variable de tipo string (object) que contiene el contenido del tweet publicado por el usuario en cuestión. Este puede incluir texto, imágenes, enlaces y otros tipos de contenido multimedia. Es la información principal que se utiliza en este proyecto para detectar emociones y analizar patrones en el lenguaje utilizado en los tweets.

3. Objetivo:

3.1. Objetivo General:

Desarrollar un modelo que permita detectar emociones en los tweets y analizar patrones en el lenguaje utilizado en Twitter para ayudar en la detección temprana de trastornos emocionales como la depresión, la ansiedad, entre otros.

3.2. Objetivos especificos:

- Pre-procesar los datos obtenidos para limpiarlos y estandarizarlos, incluyendo la eliminación de caracteres especiales, hashtags y menciones, y la normalización de las palabras.
- Realizar un análisis exploratorio de los datos para identificar patrones y características relevantes para la detección de emociones, como las palabras más comunes asociadas con cada emoción y las tendencias de frecuencia de las emociones a lo largo del tiempo.
- Desarrollar y comparar dos algoritmos de procesamiento de lenguaje diferentes, como NTKL y Bert, para determinar cuál es el más efectivo para la detección de emociones en tweets.
- Evaluar rigurosamente los modelos utilizando técnicas de validación cruzada y otras métricas de evaluación, como la precisión, el recall y la F1-score, y seleccionar el mejor modelo.
- Realizar pruebas en datos nuevos para evaluar la eficacia del modelo en la detección de emociones en tweets que no se utilizaron durante el entrenamiento.
- Crear una pipeline de automatización del modelo, que incluya el pre-procesamiento de los datos de entrada, la ejecución del modelo y la generación de resultados.

3.3. Objetivos adicionales:

En caso de tener tiempo se realizarán los siguientes objetivos:

- Desarrollar y probar el modelo de Word2Vec.
- Desplegar el mejor modelo en una interfaz o API.

4. Metodología

En cuanto a la metodología de trabajo, se utilizaran dos algoritmos de procesamiento de lenguaje natural, los cuales que se evaluarán y compararán.

- **NTKL** es una biblioteca de procesamiento de lenguaje natural que se utiliza para el pre-procesamiento de texto, análisis de sentimientos, tokenización y clasificación de texto.
- **Bert** es un modelo de lenguaje pre-entrenado desarrollado por Google que utiliza redes neuronales para el procesamiento de texto y tiene una capacidad de comprensión semántica mucho más avanzada que otros modelos de lenguaje.

En caso de disponibilidad de tiempo se probará el siguiente modelo:

- **Word2Vec** es una técnica para el procesamiento de lenguaje natural publicada en 2013. El algoritmo Word2vec utiliza un modelo de red neuronal para aprender asociaciones de palabras a partir de un gran corpus de texto. Una vez entrenado, dicho modelo puede detectar palabras sinónimas o sugerir palabras adicionales para una frase sin terminar. Como su nombre indica, Word2vec representa cada palabra distinta con una lista particular de números llamada vector. Los vectores están escogidos cuidadosamente de forma que una función matemática sencilla (la similitud coseno entre los vectores) indica el nivel de la similitud semántica entre las palabras representada por dichos vectores.

5. Referencias

- https://github.com/twintproject/twint
- https://data.world/crowdflower/sentiment-analysis-in-text
- https://huggingface.co/docs/transformers/model_doc/bert
- Wong, Y. K., Lee, S. S., & Hung, E. (2017). Emotion Detection in Twitter Using Deep Learning Approaches. In Proceedings of the 2017 International Joint Conference on Neural Networks.