Bootcamp Data Science - Ejercicio 13 - Gerardo Rodríguez

Diseño de un Sistema de Análisis de Sentimiento

Imagina que eres un científico de datos en una startup que ha desarrollado una plataforma de reseñas de productos tecnológicos. Tu objetivo es diseñar un sistema que automáticamente clasifique y etiquete las reseñas de los usuarios según su sentimiento: positivo, negativo o neutro.

1. Recopilación y Preprocesamiento de Datos:

- ¿Qué tipo de datos necesitarías recopilar para entrenar tu modelo?
    - Datos de entrenamiento y prueba que incluirán:
        - Las recomendaciones realizadas por los clientes
        - Etiquetas clasificando el sentimiento de cada comentario definidas por un humano.
        - Otros datos, como fechas de compra, puntuación del producto, fabricante, Modelo del producto.

- Describe tres técnicas de preprocesamiento de texto que aplicarías a las reseñas antes de alimentarlas al modelo.
    - Tokenización: indexa las palabras contenidas en una oración para poderlas identificar en una matriz. Esto permitirá analizar cada palabra posteriormente.
    - Eliminar Stop Words: Mediante un diccionario elimina aquellas palabras que usualmente no son relevantes para el análisis de sentimientos. El diseñador del modelo debe validar si se requiere editar la lista agregando o eliminando palabras del diccionario.
    - Lematización y Stemming: ambas técnicas reducen las palabras a su forma base, aunque el Steeming utiliza el sentido común del sistema para rastrear la raíz y no es tan preciso. Por otro lado la Lematización es más precisa porque sigue reglas gramaticales, pero puede ocupar más recursos del sistema. Por tanto, habría que validar si nuestro entorno nos da el lujo de Lematizar en lugar de usar Stemming.

2. Modelado:

- ¿Qué modelos de lenguaje considerarías para esta tarea y por qué?
    - Bert, ELMo, RoBERTa, DistilBERT o LSTM son modelos que han demostrado eficiencia identificando sentimientos en textos.
    - Sin embargo, cada uno tiene pros y contras, principalmente basados en la capacidad de performance para procesar grandes o pocas cantidades de datos.
    - Por lo que la elección del modelo dependerá en buena parte de la cantidad de datos a analizar.
    - En este caso pensando que tendremos muchos muchos datos, recomendaríamos usar Bert o RoBERTa.

- Suponiendo que decides utilizar un modelo basado en embeddings como Word2Vec o BERT, ¿cómo podrías usar esos embeddings para la clasificación de sentimientos?
    - Bert puede usar embeddings para identificar la correlación de otras palabras y dar cierta orientación más objetiva a como puede asignarse un sentimiento a un enunciado.
    - Esto lo realizará utilizando diccionarios de palabras relacionadas, donde se buscarán palabras que aparezcan en una misma oración para sumar puntos en caso de que tengan relevancia.

3. Evaluación:

- Describe al menos dos métricas que usarías para evaluar el rendimiento de tu sistema de análisis de sentimiento.
    - Precisión (Accuracy) es un indicador que nos da como resultado el porcentaje de eficiencia en la evaluación de los datos de prueba, calculando el total de predicciones correctas entre el total de predicciones.
    - Por otro lado el índice de Sensibilidad o Recall, es de importancia para identificar los verdaderos positivos entre la suma de verdaderos positivos y falsos negativos.

4. Aplicación Práctica:

- Una vez implementado tu modelo, un cliente quiere usarlo para identificar reseñas que podrían ser útiles para mejorar un producto.
  ¿Cómo podrías adaptar o utilizar tu sistema de análisis de sentimientos para ayudar en esta solicitud?
    - Sería posible identificar las reseñas positivas y rastrear en un nuevo modelo las características positivas del producto con el fin fortalecer aquellas características que son del agrado del cliente.
    - Identificar palabras clave como recomendación, mejora, cambio (y sus variantes) para rastrear posibles recomendaciones.
    - Obviamente las clasificaciones negativas permitirán tomar acciones correctivas, pero deben catalogarse por clusters o de alguna forma que permita simplificar el análisis, por ejemplo buscando las palabras más repetidas, o las que tengan más puntos de negatividad.
    - Esto último permitirá identificar:
        - problemas y deficiencias del producto,
        - problemas no relativos directamente al producto como puede ser: la entrega, el soporte, pago, garantía, entre otros.

5. Reflexión:

- Considera las implicaciones éticas del análisis automático de sentimientos en reseñas de usuarios. ¿Qué preocupaciones éticas podrían surgir y cómo podrías abordarlas?

    - Validar que tenemos consentimiento del cliente para analizar y aplicar acciones basados en sus comentarios. No hacerlo podría representar un riesgo legal e incluso hacer que perdamos la confianza de nuestros clientes.

- Que los datos de los clientes estén anonimizados, esto evitará que los analistas puedan acceder a ellos de manera indebida y que se discrimine mediante supuestos como raza, religión, o afiliación política por mencionar algunas variables.
- Que la información que se recopilada se conserve de manera interna y esta no sea publicada de otras formas al publico en general.