Tecnicatura Superior en Ciencia de Datos e Inteligencia Artificial

Materia: Técnicas de Procesamiento de Habla

Profesor: Moises Tinte

Integrantes:

- Nicolas González Da Silva
- Lilen Guzmán
- Silvia Carina Monzón
- Gabriela Cáceres

1. Comprensión del negocio

Objetivo del proyecto:

Desarrollar un sistema de análisis de sentimiento de tweets que clasifique de forma correcta las emociones expresadas en los mismos.

Beneficios e impacto en la solución propuesta:

- Permitir a las empresas comprender mejor las opiniones y sentimientos de sus clientes.
- Ayudar en la monitorización de las redes sociales para detectar cambios en el estado de ánimo del público.
- Proporcionar información útil para campañas de marketing y atención al cliente.

Criterios de éxito:

- Alta precisión en la clasificación de emociones.
- Capacidad de manejar grandes volúmenes de datos en tiempo real.
- Facilidad de uso e integración con otras herramientas de análisis de datos

2. Comprensión de los datos

El conjunto de datos:

El dataset tiene 40,000 tweets con las siguientes columnas: tweet_id, sentiment, y content.

Explorar los datos:

- Inspeccionar el dataset para comprender su estructura y contenido.
- Revisar el balance de clases en la columna sentiment para identificar posibles desbalances.

Identificar problemas de calidad:

- Revisar si hay datos faltantes o duplicados.
- Evaluar la distribución de las clases de sentimiento para ver si hay desbalance.

3. Preparación de los datos (técnicas de procesamiento de habla en el proyecto)

- ➔ Tokenización: Para este proyecto, se utiliza la biblioteca NLTK para tokenizar los tweets.
- ➔ Eliminación de stop words: Estas palabras se eliminan para reducir el ruido en los datos y mejorar la eficiencia del modelo.
- ➔ Lematización:Esto ayuda a normalizar el texto.
- ➔ Conversión a minúsculas:Para asegurar la uniformidad.
- ➔ Eliminación de caracteres especiales:No aportan información valiosa al análisis de sentimientos.
- ➔ Vectorización:Se convierte el texto preprocesado en vectores numéricos usando TF-IDF (Term Frequency-Inverse Document Frequency)
- ➔ Balanceo de datos: Se utiliza RandomOverSampler para balancear las clases de sentimientos en el conjunto de datos

4. Modelado

Selección de algoritmos:

- ★ Regresión logística (Ir al archivo)
- ★ Multinomial Naive Bayes (Ir al archivo)

Diseño de cada modelo:

- ○ Vectorización del texto
- ○ División del conjunto de datos
- ○ Entrenamiento del modelo
- ○ Predicción y evaluación

5. Evaluación y comparación

Evaluación del Modelo de Multinomial Naive Bayes

El modelo Multinomial Naive Bayes ha demostrado ser el mejor en este proyecto, este modelo es adecuado para tareas de clasificación de texto debido a su simplicidad y eficacia en problemas de clasificación con características de texto discretas como la frecuencia de palabras. Aunque la precisión y recall no son altos, el equilibrio entre estas métricas lo hace útil para una clasificación razonablemente buena y rápida.

Evaluación del Modelo de Regresión Logística

El modelo de Regresión Logística no ha tenido un rendimiento tan bueno como el de Naive Bayes, aunque la precisión es comparable a la de Naive Bayes, la menor exactitud y recall muestran que el modelo de Regresión Logística tiene más problemas para clasificar correctamente los sentimientos.

.

Tabla Comparativa de Valores Alcanzados		
Métrica	**Multinomial Naive Bayes**	**Regresión Logística**
Accuracy	0.5183	0.3761
Precision	0.4852	0.4847
Recall	0.5183	0.3761
F1 Score	0.4928	0.4212

Conclusión

En general, el modelo Multinomial Naive Bayes ha mostrado ser más util para esta tarea específica de clasificación de sentimientos en tweets. Sin embargo, ambos modelos tienen un margen significativo para la mejora. La aplicación de técnicas avanzadas de preprocesamiento y el uso de modelos más complejos como ensamblajes o redes neuronales podrían generar una mejora en la precisión y recall de la clasificación de sentimientos.

Mejoras a Tener en Cuenta

- Optimización de Hiperparámetros: Realizar una búsqueda de hiperparámetros más exhaustiva para ambos modelos podría mejorar su rendimiento.
- Enriquecimiento de Datos: Aumentar la cantidad de datos de entrenamiento mediante técnicas de aumento de datos (data augmentation) puede ayudar a los modelos a generalizar mejor.

- ❖ Mejora del Preprocesamiento: Refinar las técnicas de preprocesamiento, por ejemplo, utilizando técnicas más avanzadas de lematización y eliminación de stop words específicas para el dominio de los tweets.
- ❖ Análisis de la Matriz de Confusión: Analizar en detalle la matriz de confusión para identificar las clases que están siendo confundidas con mayor frecuencia y ajustar el modelo o el preprocesamiento para abordar esos casos específicos.

6.Bibliografía y fuentes

- *https://www.kaggle.com/*
- *The Process of Knowledge Discovery on Databases, Timarán-Pereira,*
- *Guía paso a paso de Minería de Datos, Pete Chapman (NCR),*
- *Introduction to Machine Learning with Python, Andreas C. Müller & Sarah Guido*
- *Hands-On Machine Learning with Scikit-Learn, Keras & TensorFlow - Aurelien Geron*
- *Scrum y xp desde las trincheras , Henrik Kniberg*