Contribution Title

Miguel Alejandro Yáñez Martínez[1] and Darío Rodríguez LLosa[1]

Universidad de La Habana

Abstract. Este proyecto desarrolla un sistema de recomendación de hoteles basado en la satisfacción del cliente, evaluada a partir de las reseñas dejadas por los usuarios. Se utilizó un modelo de reconocimiento de emociones, sometido a un proceso de fine-tuning para optimizar su precisión en la tarea de analizar la positividad de las reseñas. Durante la fase de entrenamiento, se ajustaron los parámetros del modelo mediante un optimizador y un programador de tasa de aprendizaje, evaluando su rendimiento regularmente en un conjunto de validación. Para prevenir el sobreajuste, se seleccionó la versión del modelo que ofreció los mejores resultados sin pérdida significativa en los datos de entrenamiento. Se realizó un análisis de errores, calculando métricas de rendimiento y generando una matriz de confusión. Finalmente, se comparó el rendimiento del modelo ajustado con un modelo BERT clásico y otro preentrenado para el reconocimiento de emociones en reseñas. Además, se desarrolló un método para organizar un ranking de hoteles, combinando la salida del modelo con el rating de los usuarios

Keywords: BERT · Entrenamiento · Procesamiento del Lenguaje Natural· Sistema de Recomendaciones · Hoteles.

Repositorio:
`https://github.com/Matcom-Projects/Hotel-Recommendation-System`

1 Introducción

El análisis de comentarios en la industria hotelera ha cobrado relevancia debido a su impacto directo en la percepción y toma de decisiones de los clientes. En este contexto, el reconocimiento de emociones en el análisis de texto ha demostrado ser una herramienta eficaz para capturar la complejidad de las respuestas emocionales de los usuarios. A diferencia de los enfoques tradicionales de análisis de sentimientos, que se limitan a clasificar los comentarios en categorías básicas como "positivos" o "negativos", el reconocimiento de emociones permite identificar y cuantificar de manera más precisa las diversas emociones expresadas en las reseñas.

Este trabajo desarrolló un sistema de reconocimiento de emociones diseñado para analizar comentarios sobre hoteles y evaluar la positividad de los mismos. A partir de esta evaluación, se generó un ranking de hoteles basado en la intensidad y frecuencia de emociones positivas identificadas en las reseñas de los

usuarios. Esta metodología no solo ofreció una visión más detallada del grado de satisfacción de los clientes, sino que también proporcionó un mecanismo objetivo para clasificar los hoteles según la calidad percibida por sus huéspedes.

1.1 States of Arts

El análisis de sentimientos ha evolucionado significativamente desde sus inicios, donde se utilizaban técnicas tradicionales que dependían de la ingeniería manual de características para su funcionamiento. Con la llegada de modelos de aprendizaje profundo, como las redes neuronales recurrentes (RNN) y convolucionales (CNN), se mejoró la capacidad de manejar datos textuales complejos permitiendo análisis más precisos y contextuales. Estos avances reflejan una tendencia hacia el uso de modelos preentrenados, que ofrecen alta precisión y versatilidad en diversas aplicaciones.

Enfoques de Análisis de Sentimientos Los algoritmos de análisis de sentimientos se dividen en dos categorías:

- **Basados en reglas:** estos sistemas realizan análisis de sentimientos automáticamente basándose en un conjunto de reglas elaboradas manualmente.
- **Automáticos:** los sistemas se basan en técnicas de aprendizaje automático para aprender de los datos.

Enfoques Basados en Reglas Generalmente, un sistema basado en reglas intenta ayudar a determinar la subjetividad de una oración, la polaridad o el tema de una idea. La herramienta más utilizada aquí es "regex".

Estas reglas suelen incluir las siguientes dos técnicas de procesamiento del lenguaje natural (NLP):

- Stemmatización, tokenización, etiquetado de partes del discurso y análisis sintáctico.
- Léxicos (es decir, listas de palabras y expresiones).

El mecanismo de funcionamiento de estos sistemas es brevemente el siguiente:

1. Crear una lista de palabras polarizadas (por ejemplo, malo-bueno, peor-mejor, feo-hermoso, etc.). Estas listas se pueden encontrar como código abierto.
2. Calcular la proporción de palabras positivas y negativas en una oración.

Los enfoques basados en reglas están ahora obsoletos y no se utilizan tanto como antes. Estos enfoques fallan en la detección de ironías y no capturan con precisión cómo se sienten los usuarios. Por esta razón, los enfoques automatizados están ganando mayor importancia actualmente.

Enfoques Automáticos Estos sistemas no dependen de reglas elaboradas manualmente, sino de técnicas de aprendizaje automático, como la clasificación. La clasificación, que se utiliza para el análisis de sentimientos, es un sistema automático que necesita ser alimentado con texto de muestra antes de devolver una categoría, por ejemplo, positivo, negativo o neutral.

Así es como se puede implementar un clasificador de aprendizaje automático:

Algoritmos de Clasificación El paso de clasificación generalmente implica un modelo estadístico como Naïve Bayes, Regresión Logística, Máquinas de Vectores de Soporte (SVM) o Redes Neuronales:

- **Naïve Bayes:** es una familia de "clasificadores probabilísticos" simples, basados en la aplicación del teorema de Bayes con fuertes suposiciones (ingenuas) de independencia entre las características (ver clasificador de Bayes).
- **Regresión Lineal:** es un enfoque lineal para modelar la relación entre una respuesta escalar y una o más variables explicativas (también conocidas como variables dependientes e independientes).
- **Máquinas de Vectores de Soporte (SVM):** es un algoritmo de aprendizaje automático supervisado que puede ser utilizado para problemas de clasificación o regresión. Sin embargo, se usa principalmente en problemas de clasificación. La Máquina de Vectores de Soporte es un límite que mejor separa dos clases (hiperplano/línea).
- **Aprendizaje Profundo:** (también conocido como aprendizaje estructurado profundo) es parte de una familia más amplia de métodos de aprendizaje automático basados en redes neuronales artificiales con aprendizaje de representaciones. El aprendizaje puede ser supervisado, semisupervisado o no supervisado.

Información Breve sobre BERT: BERT utiliza Transformer, un mecanismo de atención que aprende las relaciones contextuales entre palabras (o subpalabras) en un texto. En su forma básica, Transformer incluye dos mecanismos separados: un codificador que lee el texto de entrada y un decodificador que produce una predicción para la tarea. Dado que el objetivo de BERT es generar un modelo de lenguaje, solo es necesario el mecanismo de codificación.

BERT es un transformer bidireccional para el preentrenamiento sobre una gran cantidad de datos textuales no etiquetados, con el fin de aprender una representación del lenguaje que se puede afinar para tareas específicas de aprendizaje automático. Aunque BERT superó el estado del arte en procesamiento de lenguaje natural (NLP) en varias tareas desafiantes, su mejora en el rendimiento se puede atribuir al transformer bidireccional, a nuevas tareas de preentrenamiento como el Modelo de Lenguaje enmascarado y la Predicción de la Estructura Siguiente, junto con una gran cantidad de datos y la potencia computacional de Google.

Información Breve sobre XLNet: XLNet es un transformer bidireccional que utiliza una metodología de entrenamiento mejorada, datos más amplios y mayor potencia computacional para lograr mejores métricas de predicción que BERT en 20 tareas lingüísticas.

Para mejorar el entrenamiento, XLNet introduce la modelización de lenguaje por permutación, donde se predicen todos los tokens pero en orden aleatorio. Esto contrasta con el modelo de lenguaje enmascarado de BERT, donde solo se predicen los tokens enmascarados (15%). También difiere de los modelos de lenguaje tradicionales, donde todos los tokens se predicen en orden secuencial en lugar de en orden aleatorio. Esto ayuda al modelo a aprender relaciones bidireccionales y, por lo tanto, maneja mejor las dependencias y relaciones entre palabras. Además, se utilizó Transformer XL como la arquitectura base, que mostró buen rendimiento incluso en ausencia de entrenamiento basado en permutaciones.

Información Breve sobre RoBERTa: RoBERTa es un enfoque robustamente optimizado de BERT, introducido por Facebook. RoBERTa es un reentrenamiento de BERT con una metodología de entrenamiento mejorada, 1000% más datos y mayor potencia computacional.

Para mejorar el procedimiento de entrenamiento, RoBERTa elimina la tarea de Predicción de la Siguiente Oración (NSP) del preentrenamiento de BERT e introduce un enmascaramiento dinámico, de manera que el token enmascarado cambia durante las épocas de entrenamiento. También se encontró que el uso de tamaños de lotes de entrenamiento más grandes era más útil en el procedimiento de entrenamiento.

2 Entrenamiento

2.1 Preprocesamiento de datos

Debido a que para el entrenamiento los datos de entrada debían presentar un formato específico, se procedió a modificar los datos. En primer lugar se limpió el texto eliminando caracteres no deseados, como emojis, signos de puntuación y números, y se tokenizó el texto para dividirlo en palabras y caracteres significativos.

Se codificaron las etiquetas de las reseñas según sus calificaciones, categorizándolas como "Negativas", "Neutras" o "Positivas", ya que estas son las clasificaciones utilizadas por el modelo. Posteriormente, se tokenizaron las reseñas y se calculó la longitud de los tokens para cada una con el fin de verificar si la cantidad de tokens excedía el tamaño máximo permitido por BERT (512 tokens).

Finalmente, se realizó un estudio de estadísticas relevantes, como la distribución de las etiquetas, la longitud de los tokens y la cantidad de caracteres por reseña, entre otros, para comprender mejor la naturaleza de los datos.

2.2 Entrenamiento

Debido al alto costo asociado con el entrenamiento de un modelo, se optó por realizar este proceso en Google Colab, ya que proporciona acceso a recursos de computación en la nube, especialmente GPUs, que facilitan el entrenamiento del modelo. Durante la fase de entrenamiento, se llevó a cabo el ajuste de los parámetros del modelo a lo largo de varias épocas utilizando un optimizador y un programador de tasa de aprendizaje. Al final de cada época, se realizó la retropropagación y se actualizaron los parámetros del modelo, los cuales fueron guardados posteriormente. Además, se evaluó el rendimiento del modelo en un conjunto de validación.

Para evitar el sobreajuste (overfitting), se seleccionó la versión del modelo que ofreciera los mejores resultados sin mostrar una disminución en el f1-score ya que indicaría la presencia de sobreajuste.

Finalmente, se realizó un análisis de errores en el conjunto de validación, calculando diversas métricas de rendimiento y generando una matriz de confusión. El rendimiento del modelo entrenado se comparó con el modelo BERT clásico utilizado para clasificación, así como con otro modelo previamente entrenado para el reconocimiento de emociones en reseñas, encontrado en internet, evaluando las métricas y las matrices de confusión de ambos modelos.

3 Sistema de recomendación

Para organizar el ranking, se utilizó la salida del modelo junto con el rating. La salida del modelo consistió en una etiqueta (positivo, negativo o neutro) y la confianza del modelo en la precisión de esa etiqueta. Dado que el valor de confianza refleja el grado de certeza del modelo respecto a la correctitud de la etiqueta devuelta, se interpretó como una medida de cuán cerca estaba la emoción expresada en una reseña con la etiqueta asignada.

Para calcular el valor correspondiente a cada reseña, se decidió sumar el valor de confianza proporcionado por el modelo y el rating asociado, ambos multiplicados por un factor que refleja la importancia otorgada a cada componente. En el caso del rating, se multiplicó por 1.5, mientras que la confianza se multiplicó por -0.5 si el sentimiento era negativo, por 1.0 si era neutro, y por 2.5 si era positivo. Posteriormente, para cada hotel, se realizó una sumatoria de todas las reseñas correspondientes y, finalmente, se ordenaron los resultados de menor a mayor.

4 Aspectos a mejorar

- Utilizar un modelo base mejor,talvez un modelo Bert de mayor tamaño u otro modelos de lenguaje como por ejemplo RoBERTa
- Realizar algunas operaciones extra en el preprocesamiento de los datos que mejoren el entrenamiento
- Aplicar mecanismos para evitar el overfitting
- Mejorar el sistema de recomendaciones

References

1. Base de datos usada, https://data.world/datafiniti/hotel-reviews
2. LNCS Homepage, http://www.springer.com/lncs
3. https://youtu.be/mvh7DV84mr4