Proyecto de MMA:Clasificación de quejas financieras

María Fernanda Suárez González
Adrián Navarro Foya
Daniel Polanco Pérez

September 4, 2024

https://github.com/MaferGlez03/Complaints-Classification-MMA-Project.git

Introducción

En el ámbito financiero, las instituciones reciben miles de quejas de los consumidores cada semana, lo que puede ser un desafío para gestionarlas y clasificarlas manualmente. Clasificar estas quejas en categorías específicas de productos es crucial para poder ofrecer una respuesta rápida y eficiente. Este proyecto aborda la tarea de automatizar la clasificación de quejas utilizando técnicas de procesamiento de lenguaje natural (NLP) y modelos de aprendizaje automático.

El objetivo es desarrollar una solución que pueda analizar las descripciones de las quejas, identificar patrones en el texto, y asignar cada queja a su categoría de producto correspondiente, como tarjetas de crédito, préstamos hipotecarios, entre otros. Esto permite a las organizaciones responder de manera más proactiva a las necesidades de los consumidores. A continuación, se detallan las etapas clave de la implementación de esta solución.

Desarrollo

La solución se implementa en varias etapas clave:

Preprocesamiento de Datos

Comenzamos cargando un conjunto de datos de quejas de consumidores proporcionado en un archivo CSV. Este conjunto de datos fue preprocesado para seleccionar únicamente las columnas y filas relevantes (product y consumer_complaint_narrative.). Posteriormente, se aplicaron técnicas de limpieza de texto, como la eliminación de palabras comunes sin valor informativo (stopwords) y la reducción de las palabras a su raíz mediante stemming. Para convertir el texto en una forma que pudiera ser utilizada por los algoritmos de aprendizaje automático, aplicamos la técnica de TF-IDF (Term Frequency-Inverse Document Frequency), que transforma los textos en vectores numéricos que representan la importancia de cada palabra en un documento en relación con todo el conjunto de datos.

Entrenamiento del Modelo

Para mejorar la precisión, se utilizan dos modelos de aprendizaje automático combinados mediante un Voting Classifier. Regresión Logística: Este modelo estadístico estima la probabilidad de que una instancia pertenezca a una categoría particular. Se basa en la relación lineal entre las características de entrada y el logaritmo de las probabilidades de la variable objetivo. En el caso específico de nuestra solución, la aplicamos entrenando el modelo con los vectores TF-IDF de las quejas y las etiquetas de categorías correspondientes. Naive Bayes: Basado en el teorema de Bayes, este modelo asume que las características son independientes entre sí, lo que simplifica el cálculo de probabilidades. Naive Bayes es conocido por su eficacia en la clasificación de texto, especialmente con grandes volúmenes de datos. En mi implementación, entrené un clasificador Naive Bayes utilizando los datos preprocesados y los vectores TF-IDF, y lo evalué para comparar su rendimiento con otros modelos. Linear Support Vector Classifier (LinearSVC): es un clasificador basado en máquinas de soporte vectorial (SVM) que busca encontrar un hiperplano que divida los datos en diferentes clases con el margen máximo posible. A diferencia de otros tipos de SVM, LinearSVC utiliza una función de pérdida lineal y es particularmente eficiente para problemas de clasificación con un número alto de características, como es el caso de los datos de texto representados mediante TF-IDF

Para aprovechar al máximo las fortalezas de cada uno de los modelos mencionados, se utilizó un Voting Classifier con votación mayoritaria (hard voting). Este método toma las predicciones de cada modelo individual y selecciona la clase que recibe la mayoría de los votos como la predicción final.

Evaluación del Modelo

Para la evalución de la eficacia del modelo utilizamos dos métodos que nos brindan datos númericos La precisión (accuracy) mide el porcentaje de predicciones correctas que hace el modelo en relación con el total de predicciones realizadas. Es una métrica simple pero útil, especialmente cuando las clases están equilibradas. En nuestro caso los modelos de Regresión Logística, Naive Bayes y LinearSVC presentan una precisión de 84.4%, 79.1% y 84.6% respectivamente. El modelo resultante del Voting Classifier tiene una precisión de 84,6% lo que representa que la clasificaciones realizadas por nuestro modelo son confiables y bastante certeras.

El informe de clasificación (classification_report) proporciona un desglose detallado de otras métricas de rendimiento como: precisión (precision) que representa la proporción de verdaderos positivos entre los ejemplos que fueron clasificados como positivos, sensibilidad (recall) que muestra la proporción de verdaderos positivos entre los ejemplos que son realmente positivos, F1-Score contiene la media armónica entre la precisión y el recall, útil cuando se busca un equilibrio entre ambas y soporte (support) que es el número de instancias verdaderas de cada clase.

Implementación de la Interfaz Gráfica con PyQt5

La interfaz gráfica del usuario (GUI) se implementó utilizando PyQt5 para ofrecer una experiencia de usuario más moderna y funcional. La GUI permite al usuario ingresar texto de quejas financieras y recibir una clasificación instantánea de la categoría de la queja.

Conclusiones

La precisión del Voting Classifier combinado se evalúa en un conjunto de datos de validación, mostrando resultados positivos al combinar las fortalezas de múltiples modelos de aprendizaje automático. La interfaz gráfica basada en PyQt5 facilita la interacción con el modelo, permitiendo que usuarios no técnicos clasifiquen las quejas de manera rápida y precisa. La implementación de esta solución proporciona una herramienta robusta y accesible para la clasificación automatizada de quejas financieras, mejorando la eficiencia y capacidad de respuesta de las organizaciones en la gestión de quejas de clientes.