Tarea 3: Comparación de distintas vectorizaciones de texto de algunos comentarios recibidos por el cliente final

Aldo Daniel Ojeda Rodriguez

17 Julio 2024

Resumen

Este estudio evaluó técnicas de vectorización de texto y modelos de clasificación para analizar comentarios de clientes. Se utilizaron Bosques Aleatorios, Regresión Logística y Máquinas de Soporte Vectorial, aplicando transformaciones como conteo de palabras y TF-IDF. Los hallazgos resaltan la importancia de elegir adecuadamente las técnicas de vectorización y modelos de clasificación para lograr un análisis efectivo de textos.

1. Introducción

Para el presente proyecto se realizó un diseño de Experimentos para comparar el efecto de distintas técnicas de vectorización en datos de quejas recibidas por clientes de una empresa cervecera.

El conjunto de datos contiene comentarios de cuatro clasificaciones de quejas capturadas a lo largo del tiempo relacionadas con :

- Disponibilidad del producto :10226
- Malas experiencias del cliente por parte del Staff : 2139
- Problemas con entrega : 2169
- Tratos con el cliente : 2468

se busca entrenar distintos modelos multiclase para entre ellos comparar cuales resultan ser los que generan mejores resultados, entre ellos se tienen los siguientes modelos.

Random Forest es un algoritmo de aprendizaje automático basado en árboles de decisión. Es un método de ensamble que construye múltiples árboles y los combina para mejorar la precisión de las predicciones y reducir el riesgo de sobre ajuste. Cada árbol en el conjunto se entrena con una muestra aleatoria del conjunto de datos de entrenamiento y realiza predicciones de manera independiente para al final generar una agregación de los resultados[1].

Support Vector Machine (SVM) es un algoritmo de aprendizaje supervisado que se utiliza tanto para clasificación como para regresión. En el caso de clasificación, el objetivo

de SVM es encontrar el hiperplano que mejor separe las clases en el espacio de características. Para problemas de clasificación binaria, SVM busca el hiperplano con el mayor margen, es decir, la distancia máxima entre las instancias más cercanas de cada clase, conocidas como vectores de soporte [2].

La Regresión Logística Multinomial es una extensión de la regresión logística utilizada para problemas de clasificación multiclase, donde la variable de respuesta tiene tres o más categorías. A diferencia de la regresión logística binaria, que predice la probabilidad de un evento en una situación de dos clases, la regresión logística multinomial predice la probabilidad de pertenecer a cada una de las posibles clases [3].

por otro lado se utilizan distintas técnicas de vectorización para texto, esto con el fin de transformar las palabras a números interpretables por una computadora.

El TfidfVectorizer es una técnica de extracción de características del texto que convierte una colección de documentos en una matriz TF-IDF. El término TF-IDF (Term Frequency-Inverse Document Frequency) es una medida que refleja la importancia de un término en un documento, en relación con su frecuencia en una colección de documentos (corpus). El objetivo principal de TF-IDF es dar mayor peso a la palabra o palabras que son más importantes[4], siendo utilizadas en el presente proyecto las técnicas de palabra a palabra, bigramas y trigramas para la vectorización.

por otro lado el CountVectorizer es una técnica de extracción de características del texto utilizada para convertir una colección de documentos en una matriz de conteo de términos.

2. Metodología

En el siguiente proyecto se usan tres modelos: Regresión Logística, Bosques Aleatorios y Máquinas de Soporte Vectorial, aplicados a datos transformados de cuatro maneras diferentes, incluyendo conteos de palabras y análisis por caracteres.

Preparación: Se cargan los datos, preparan y se definen los modelos. Transformaciones: Los datos se representan de diferentes maneras, como conteo de palabras o TF-IDF. Evaluación: Cada modelo se entrena y se evalúa con cada tipo de datos para determinar cuál combinación es más precisa.

3. Resultados y Discusión

Transformación	**Modelo**	**Accuracy**	**Precision**	**Recall**
Conteo	LR	0.9262	0.93	0.93
Conteo	RF	0.9271	0.93	0.93
Conteo	SVC	0.9233	0.92	0.92
Palabras	LR	0.9233	0.92	0.92
Palabras	RF	0.9203	0.92	0.92
Palabras	SVC	0.9277	0.93	0.93
N-gramas	LR	0.7330	0.78	0.73
N-gramas	RF	0.7548	0.77	0.75
N-gramas	SVC	0.7639	0.79	0.76
Caracteres	LR	0.7865	0.78	0.79
Caracteres	RF	0.8074	0.81	0.81
Caracteres	SVC	0.7865	0.78	0.79

El modelo Random Forest (RF) resulto ser el mejor modelo utilizando las distintas vectorizaciónes, alcanzando una Exactitud del 92.71 % y buenos valores de precisión y recall ponderados. por otro lado para las tecnicas de vectorización que más destacaron fueron las de conteo y frecuencia de palabras, por otro lado los bigramas y trigramas (N-gramas) no resultaron ser los mejores así mismo como el tema de caracteres. esto se debe a que es posible que al haber eliminado Stopwords haya hecho que los bigramas pierdan intención.

4. Conclusión

Existen diversas tecnicas para el análisis de texto y para la clasificación del mismo, modelos como Random Forest o Suport Vector Machine son muy utilizados al tratarse de un problema de clasificación, en conclusión las tecnicas de clasificación también sirven para analizar texto encontrando patrones y clasificando con mucha exactitud.

Referencias

[1] Leo Breiman *Random Forests*, www.stat.berkeley.edu, 2001

[2] Corinna Cortes y Vladimir Vapnik *Support-vector networks*, Springer, 1995

[3] Trevor Hastie and Robert Tibshirani y Jerome Friedman, *he Elements of Statistical Learning*, Stanford,edu, 2009

[4] Pedregosa, F. and Varoquaux et. all, *scikit-learn: Machine Learning in Python*, Journal of Machine Learning Research, 12, pp. 2825-2830, 2011