Propuesta de Detección automática de la intención de búsqueda

Erick Ottoniel Jerónimo Toj* and Sharolin Guadalupe Lacunza González**

Resumen—**A continuación se detalla el análisis de un dataset con 22635 consultas reales de usuario, previamente etiquedas como infomacionales, navegacionales o transaccionales a partir de indicadores como URLs, citas textuales y patrones de busqueda. Tras descargar consultas ambiguas, limpiar y normalizar el texto, extraemos features lingüísticas simples y representamos el contenido como TF-IDF en n-gramas combinada con variables numéricas escaladas con estos datos entrenamo un SVM lineal balanceado para mitigar el desbalance de clases y evaluamos su desempeño mediante precision, recall y F1-score por clase F1 con 0.96 para transaccional y 0.7 para las otras, adeas de analizar la matriz de confusión y las curvas ROC. El objetivo es demostrar que, aun usando features sencillos y un modelo clásico, es posible detectar automáticamente la intención de búsqueda, lo cual podría mejorar la relevancia de los resultados y personalizar la interacción en motores de búsqueda y asistentes virtuales**

Palabras Claves

- **Ambigüedad de consultas:** Situación en la que una consulta de búsqueda no expresa con claridad la intención del usuario, dificultando su clasificación como informacional, navegacional o transaccional.
- **Clasificación de intención:** Proceso automático de asignar a una consulta de búsqueda una etiqueta de intención en función de sus características lingüisticas y estructurales.
- **Curva ROC:** Gráfica que muestra la relación entre la tasa de verdaderos positivos (TPR)y la tasa de falsos positivos (FPR) para distintos umbrales, empleada para evaluar la capacidad discriminativa de un modelo.
- **Extracción de características:** Conjunto de técnicas para derivar atributos numéricos o categóricos a partir del texto con el fin de alimentar un clasificador.
- **Matriz de confusión:** Tabla que compara las etiquetas reales frente a las predichas por el modelo, detallando aciertos y errores por cada clase.
- **Máquina de vectores de soporte (SVM):** Algoritmo de aprendizaje supervisado que busca un hiperplano óptimo en el espacio de características para maximizar el margen entre clases y clasificar nuevos ejemplos.
- **Vectorización TF-IDF:** Técnica de representación de texto que pondera la frecuencua de un término en un documento (TF) por su frecuencua inversa en el corpus (IDF), resaltando palabras discriminativas.

I. INTRODUCCIÓN

En la era digital, los motores de búsqueda se han convertido en el punto de encuentro entre los usuarios y la vasta información disponible en la web. Por tanto se ha convertido en algo fundamental identificar la intención de búsqueda con el fin de ofrecer un diferenciador y ademas mostrar resultados relevantes y personalizados al usuario. En este proyecto se propone un enfoque sencillo pero eficaz basado en SVM y características textuales elementales, como detección de dominios, contar palabras, longitud de texto. El SVM lineal se entrena con ponderacón de clases para compensar el desequilibrio natural que se encuentra en la data, dado que las consultas transaccionales predominan sobre las demás. Finalmente se encuentran los resultados y gráficas correspondientes.

II. PLANTEAMIENTO DEL PROBLEMA

A. Variable objetivo

Intent (la etiqueta final de intención)

B. Variables independientes

navigational, ambiguous, adult

C. Variables dependientes

text, quote question

* Carne: 16005196, e-mail: 16005196@galileo.edu.gt, Facultad de Ingeniería, Postgrado de IA, Universidad Galileo.
** Carne: 25005009, e-mail: 25005009@galileo.edu, Facultad de Ingeniería, Postgrado de IA, Universidad Galileo.

III. METODOLOGÍA

La clasificacion de la intencionalidad de busqueda se etiqueto de la siguiente forma:

- **Navigational:** es intencion navegacional
- **Informational:** es intencion informacional
- **Transactional:** es intencion transaccional

Este etiquetado manual sirvió para generar la columna **intent** y poder categorizar la consulta

Los pasos seguidos son los siguiente:

- **Selección de datos:**
  - Se tomo el data set de 2635 consultas extraídas de logs reales de la base de datos de Yahoo.
  - Cada consutla se etiquetó automáticamente como transaccional si contenía dominios o URLs, informacional si incluía citas textuales y navegacional en caso contrario.
  - Se eliminaron las consultas marcadas como ambiguas para garantizar claridad en las tres categorías.
- **Preprocesamiento del texto:**
  - Limpieza y estandarización de datos, conversión a minúsculas y normalización unicode.
- **Extracción de características:**
  - Lingüísticas básicas como conteo de palabras, texto con .com, longitud de caracteres. Esta extracción generó manualmente 4 columnas adicionales para guardar los valores si contiene palabra como "what", "how.º algun otra que sea de interrogante, la cantidad de caracteres del texto, número de palabras, y si contenía la cadena ".com"
  - Vectorización TF-IDF, se generaron matricexs dispersas de términos unigrama y Bigrama, con umbrales min de 2 y max de 0.9
- **Construcción del conjunto de entrenamiento:**
  - Escalado de las variables numéricas mediante StandarScaler
  - Combinación horizontal de la matriz TF-IDF y la matriz de características numéricas escaladas en una única representación de alta dimensionalidad
- **Modelado y evaluación:**
  - División estratificada en entrenamiento 70 % y prueba de 30 % para conservar la distribución de clases.
  - Entrenamiento con SVM con ponderación de clases.
  - Medición de desempeño mediante precisión, recall y F1-score por clase.
  - Análisis de matriz de confusión; y curvas ROC multiclasificación.
  - Interpretación cuantitativa y cualitativa para identificar puntos fuertes y oportunidades de mejora.

IV. RESULTADOS

A. Resultados intención de búsqueda

Precisión:

- **Informacional:** de todas las veces que el modelo evaluó acertó 74 %.
- **Navegacional:** 80 % de las predicciones fueron correctas.
- **Transaccional:** 95 %, lo que indica pocos falsos positivos para esta categoría.

Recall: porcentaje de verdaderos de cada clase que el modelo logra recuperar.

- **Informacional:** 67 % de todas las consultas realmente informacionales - pierde 33 %.
- **Navegacional:** 65 % de las verdaderas navegacionales.
- **Transaccional:** 97 %, de las búsquedas transaccionales, casi sin omisiones.

F1-score: promedio armónico entre precisión y recall, balancea ambos aspectos.

- **Informacional:** 0.70, refleja un equilibrio moderado pero mejorable.
- **Navegacional:** 0.72, refleja equilibrio moderado.
- **Transaccional:** 0.96, confirma un rendimiento excelente gracias a indicadores claros.

```
Classification Report:
               precision    recall  f1-score   support

informational       0.74      0.67      0.70        21
 navigational       0.80      0.65      0.72        80
transactional       0.95      0.97      0.96       656

     accuracy                           0.93       757
    macro avg       0.83      0.76      0.79       757
 weighted avg       0.93      0.93      0.93       757

Confusion Matrix:
 [[ 14   0   7]
 [  0  52  28]
 [  5  13 638]]
```

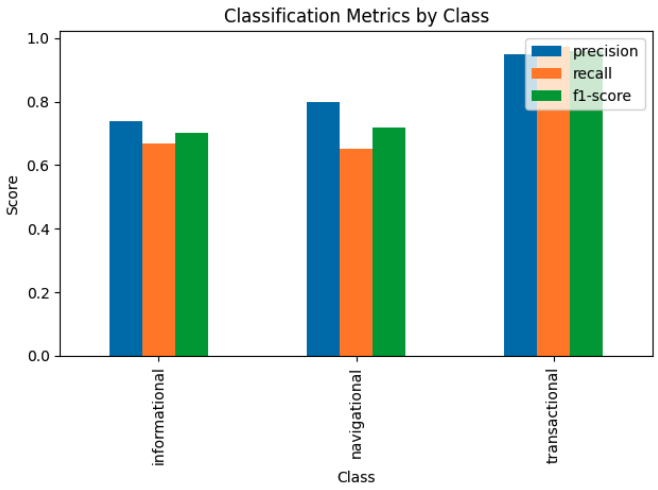
La exactitud es de 0.93 sobre un total de 757 consultas.

Los promedios macro avg tenemos 0.83 de precisión, 0.76 recall, 0.79 F1 refleja la menor representación y mayor dificultad de las clases informacional y navegacional, el weighted avg refleja un 0.93 en todas está dominado por el alto volumen y excelente desempeño transaccional.

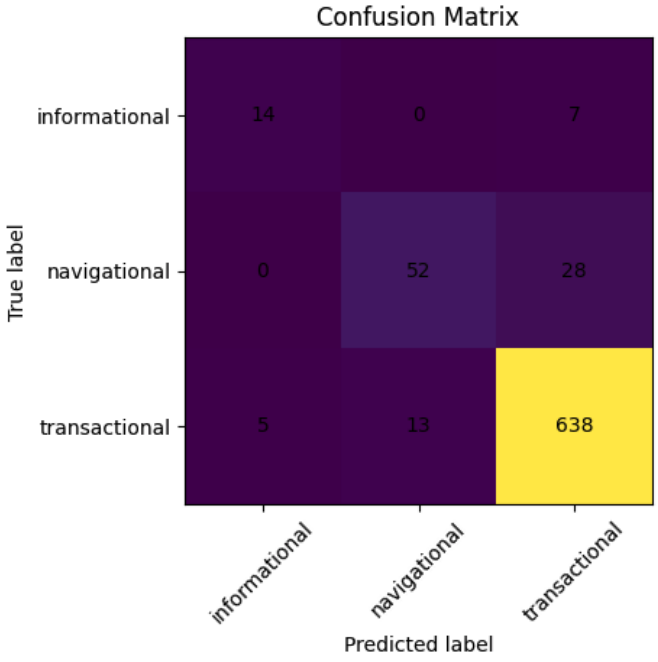
Como fuente de error encontramos que:

- de 21 consultas de informacionales, 14 clasifican correctamente pero 7 pasan a transaccional.
- de 80 consultas navegacionales, 52 son bien identificadas y 28 se confunde con transaccional.
- sólo 18 de 656 consultas transaccionales se asignan erróneamente a otro tipo

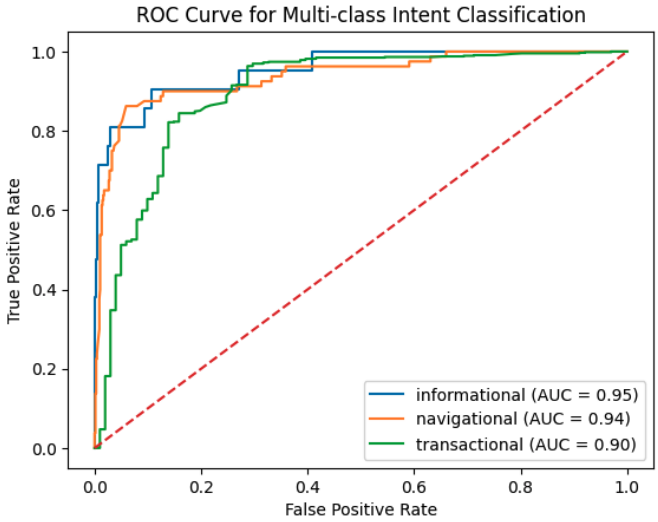
B. Resultados de métricas



C. Gráfica de matriz de confusión



D. Gráfica Curva de ROC



Muestra la capacidad del modelo para distinguir cada tipo de intención frente al resto, gráficando la tasa de verdaderos positivos (TPR) contra la tasa de falsos positivos (FPR) a distintos umbrales.

- **Informacional:** (Línea azul, AUC = 0.95) Al acercarse la curva muy pronto al vértice superior izquierdo, indica que el modelo logra alto TPR con bajo FPR para este tipo. Un AUC alto señala una excelente capacidad de discriminar consultas informacionales a pesar de su menor soporte.

- **Navegacional:** (Línea naranja, AUC = 0.94) Similar al anterior categoría el modelo separa bien las búsquedas navegacionales. La ligera caída respecto a informacional sugiere que en algunos umbrales confunde más ejemplos navegacionales con las otras clases.
- **Transaccional:** (Línea verde, AUC = 0.90) Aunque es la clase mayoritaria y con mejores métricas de F1, su AUC es el más bajo, lo que indica que en ciertos umbrales, el modelo sacrifica algo de especificidad o sensibilidad al discriminar transaccionales frente a las otras categorias. Aunque siempre mantiene un excelente rango.
- **Línea punteada roja:** (0.50) Representa el desempeño de un clasificador aleatorio. Al todas las curvas quedar por encima de esta confirma que el enfoque aporta valor real frente a la aleatoriedad.

V. CONCLUSIONES

El modelo se apoya con gran éxito en indicios claros de transacción pero muestra limitaciones para distinguir entre búsqueda informacionales y navegacionales cuando carecen de marcadores léxico evidentes. Aumentar datos de entrenamiento para estas clases o incorporar features semánticas adicionales podrían elevar su recall y equilibrar el desempeño global.

En conjunto, los AUC >0.90 en las tres clases validan que el modelo SVM con features simples destaca en la mayoría de lo spuntos de operación, aunque sugiere afinar aún mas la frontera de decisión especialmente para los casos transaccionales en escenarios muy conservadores o muy laxos.

VI. E-GRAFÍA

El data set utilizado fue extraído de Yahoo `https://webscope.sandbox.yahoo.com/catalog.php?datatype=l&did=66`