

# PAPERS BASE



Cecilia del Pilar Vilca Alvites

# TEXTLENS:ANALISIS VISUAL POTENCIADO POR MODELOS DE LENGUAJE DE GRAN ESCALA PARA MEJORAR EL AGRUPAMIENTO DE TEXTO



## Introducción

El agrupamiento de texto organiza documentos según similitudes, normalmente utilizando algoritmos de clustering sobre representaciones de texto generadas por modelos preentrenados. La integración de LLMs ha mejorado los resultados, pero persisten tres desafíos principales:

- Rendimiento limitado en dominios especializados, y altos costos al ajustarlos para distintos conjuntos de datos.
- Falta de mecanismos sistemáticos para identificar datos anómalos en el proceso de clustering, confiando en métricas personalizadas sin intervención humana.
- Evaluación deficiente en escenarios sin etiquetas, donde es difícil determinar el número óptimo de clusters, y donde métricas tradicionales como el "silhouette score" no son efectivas para datos textuales.





# OBJETIVOS

## • OBJETIVO 1


Un marco de clustering basado en LLMs que mejora resultados mediante extracción de temas, filtrado de anomalías y evaluación de cambios, incluyendo una nueva métrica interna de clusters.

## • OBJETIVO 2

Un sistema de análisis visual interactiva que facilita la exploración, inspección y modificación de clusters.

## • OBJETIVO 3

Evaluación de la efectividad y escalabilidad de TextLens mediante cuatro estudios de caso y un estudio de usuarios en dos conjuntos de datos distintos.





# REQUERIMIENTOS

Para mejorar el análisis de grandes volúmenes de texto, se busca perfeccionar el proceso de clustering de texto. Han identificado cuatro áreas clave para lograrlo:

01

## Mejorar el preprocesamiento de texto

Necesitan métodos más avanzados para destacar las características temáticas de cada texto. Esto permitirá que los modelos de lenguaje (LLMs) codifiquen la información de manera más efectiva, lo que se traducirá en clusters más precisos y significativos.

02

## Entender la semántica de los clusters

No solo quieren saber qué textos están agrupados, sino por qué lo están. Buscan una comprensión profunda de los temas semánticos que unen a los elementos dentro de cada cluster.

03

## Evaluar los resultados del clustering

Requieren métricas cuantificables para evaluar la efectividad de sus agrupaciones. Estas métricas deben funcionar incluso sin etiquetas preexistentes y ayudarles a determinar el número óptimo de clusters.

04

## Desarrollar un sistema interactivo

Implementar todo el proceso de clustering en un sistema inteligente que combine los algoritmos con análisis visual. Esto les permitirá operar y explorar los clusters de texto de forma interactiva e iterativa.

# PIPELINE DE TEXTLENS

## 01. Procesamiento de Datos

el sistema usa un LLM para extraer palabras clave (intenciones) del texto original. Luego, estas palabras clave se concatenan con el texto original para enriquecer la información. Finalmente, todo este texto enriquecido se convierte en vectores numéricos usando modelos de codificación existentes, preparándolo para el clustering.

## 02. Clustering de Texto

Una vez que los datos están en formato vectorial, se aplican algoritmos de clustering para agrupar inicialmente los textos basándose en sus representaciones numéricas.

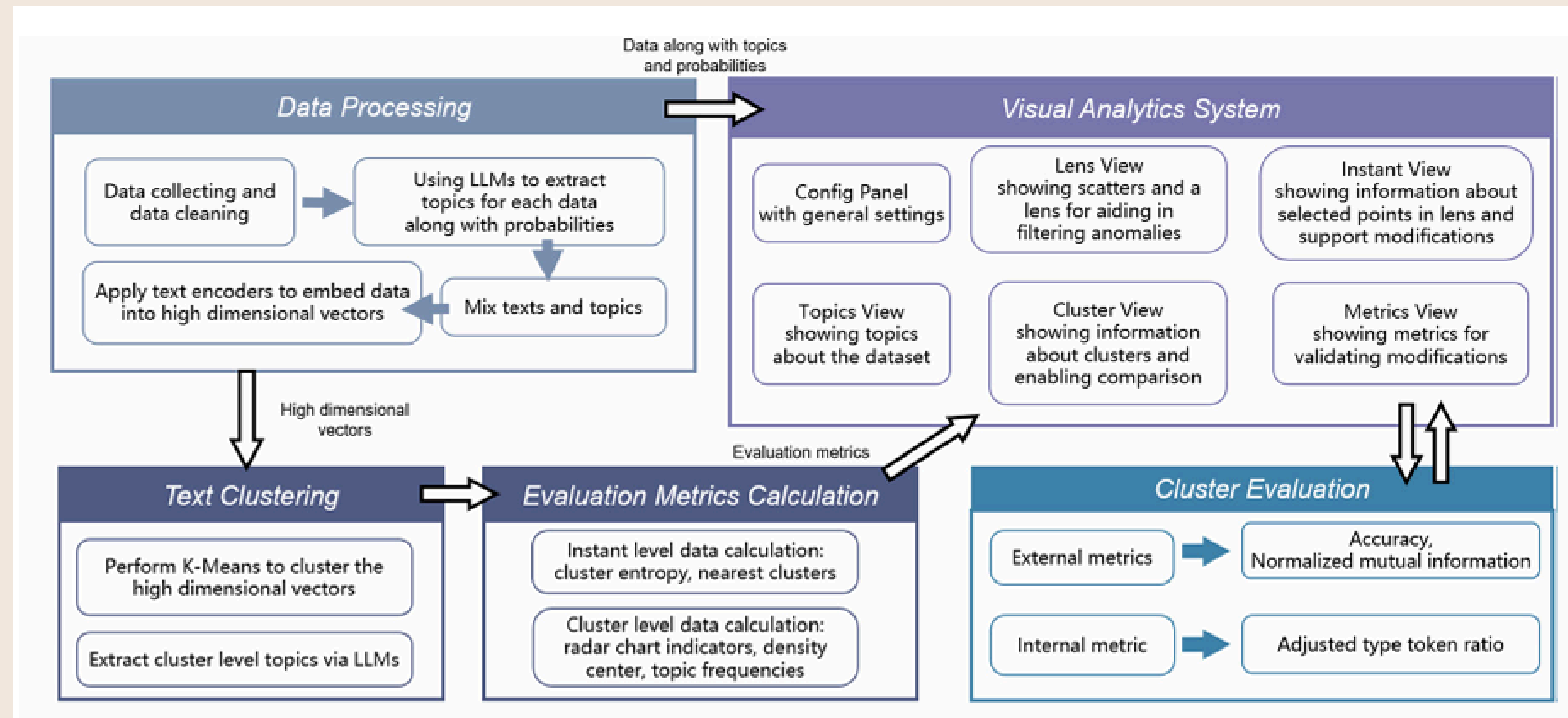


## 03. Cálculo de Métricas y Detección de Anomalías

Aquí entra en juego un sistema de analítica visual diseñado para identificar textos que no encajan bien en sus clusters (anomalías). Las anomalías detectadas son refinadas y corregidas con la ayuda de los LLMs, lo que sugiere una interacción entre la visualización y la capacidad del modelo para entender el contexto.

## 04. Evaluación del Cluster

Por último, se calculan métricas de evaluación para cuantificar la calidad de los resultados del clustering y realizar una evaluación general de cómo se han agrupado los datos.



# PROCESAMIENTO DE DATOS



Dado que codificar el texto en bruto no da resultados óptimos, este módulo se centra en enriquecer los datos antes de la codificación. Los pasos son:

01

Extracción de Tópicos con LLMs: Utilizando un LLM (específicamente GPT-3.5 Turbo), se extraen entre uno y tres intenciones por cada punto de datos. A cada intención se le asigna una probabilidad, lo que permite un filtrado posterior de datos de baja calidad. A diferencia de métodos anteriores que extraían un solo tópico, este enfoque considera la ambigüedad inherente de los textos

02

Mezcla de Texto e Intenciones : Una vez extraídos, las intenciones se concatenan directamente con el texto original antes de la codificación. Esta estrategia es clave porque:

- Evita el aumento de dimensionalidad: A diferencia de codificar texto e intenciones por separado y luego concatenar los vectores (lo que duplicaría la dimensionalidad, ej., de 784 a 1568 dimensiones), este método mantiene la dimensionalidad original.
- Mejora la fuerza temática: Al mezclar texto y las intenciones antes de codificar, se potencia la "fuerza temática" de cada punto de datos. Esto hace que las diferencias entre textos de distintos temas sean más claras, mejorando el rendimiento de algoritmos.

03

Codificación de Datos: El texto enriquecido (texto original + intención) es luego codificado en vectores de alta dimensión utilizando el modelo Instructor-large. Este modelo es elegido por su rendimiento superior y escala moderada, y requiere una "directiva de ajuste" para definir el objetivo de codificación (aquí, "Representar las expresiones para la clasificación de intenciones").

04

Estadísticas Preliminares de Palabras Clave: Además de la codificación, se calculan estadísticas sobre las combinaciones de dos palabras clave más frecuentes. Esto da a los usuarios una idea inicial del número de clusters apropiado y ayuda a configurar los hiperparámetros del algoritmo de clustering.



# CLUSTERING DE TEXTO: AGRUPACIÓN Y COMPRENSIÓN SEMÁNTICA

Una vez que los datos están vectorizados el proceso avanza a la agrupación y la interpretación:

- **Aplicación del Algoritmo de Clustering:** Se utiliza el algoritmo K-means para agrupar los vectores de datos en clusters, obteniendo etiquetas de cluster predichas.
- **Comprensión Semántica de los Clusters:** Para entender las connotaciones precisas de cada cluster, se emplea nuevamente un LLM. Este modelo interpreta las palabras clave de todos los puntos de datos dentro de un cluster para extraer una representación temática más matizada y específica para ese grupo. Se usa un prompt directo ("Your task is to extract the topic that best describes the group of keywords provided in the list. Please return only this topic") para obtener una visión semántica clara de cada cluster.



## CÁLCULO DE MÉTRICAS DE EVALUACIÓN PARA CLUSTERING

Se calculan indicadores numéricos para evaluar el proceso de clustering, tanto a nivel de punto de datos individual como a nivel de cluster completo. Estas métricas son cruciales para el sistema de análisis visual, ayudando a entender, filtrar y modificar datos anómalos.

- **Nivel de Punto de Datos (Instantáneo): Entropía del Cluster.** Se introduce la entropía del cluster para medir la perplejidad o incertidumbre de la asignación de cada punto de datos a un cluster. Una entropía más alta indica mayor confusión o incertidumbre en la asignación de un punto de datos, lo que sugiere una posible anomalía o una asignación ambigua. También se registran los tres clusters más cercanos para uso futuro.



# CÁLCULO DE MÉTRICAS DE EVALUACIÓN PARA CLUSTERING



Nivel de Cluster: Evaluación Integral (Compactación Numérica y Análisis de Texto)

Se utilizan cinco métricas para evaluar la calidad de los clusters desde dos perspectivas: compactación numérica y análisis a nivel de texto.

Métrica de Análisis a Nivel de Texto:

1. Relación Tipo-Token (TTR): Evalúa la diversidad del vocabulario dentro de cada cluster. Se calcula como la relación entre el número de palabras únicas (intención) y el número total de palabras (tokens). Un TTR más alto sugiere un vocabulario más diverso, lo que podría indicar dispersión o la presencia de datos anómalos a nivel textual.

Métricas de Compactación Numérica:

1. Distancia Promedio Dentro de una Clase: Mide la distancia media entre todos los pares de puntos de datos dentro de un mismo cluster. Una distancia mayor indica que el cluster está más disperso.
2. Distancia Máxima Dentro de una Clase: Identifica la distancia más grande entre dos puntos cualquiera dentro de un cluster. Una distancia máxima grande puede señalar la presencia de datos anómalos significativos dentro del cluster.
3. Desviación Dentro de una Clase: Calcula la desviación estándar de las distancias entre pares de puntos en un cluster. Proporciona una indicación de la dispersión de los puntos de datos.
4. Métrica de Densidad: Evalúa la proporción de puntos de datos dentro de un círculo (centrado en el centro de densidad del cluster y con radio igual a la distancia promedio). Una densidad más alta indica un cluster más compacto; una densidad baja puede sugerir anomalías.

# EVALUACIÓN DEL CLUSTER



Para evaluar el rendimiento del clustering, se utilizan tres métricas principales: Accuracy (ACC), Normalized Mutual Information (NMI) y Adjusted Type-Token Ratio (a-TTR).

- Si el conjunto de datos tiene etiquetas reales, se comparan los clusters generados con las etiquetas verdaderas usando NMI y ACC (ajustando con el algoritmo húngaro).
- Si no hay etiquetas (como suele ocurrir en la práctica), se recurre a métricas internas. Las métricas tradicionales como el silhouette score no son efectivas en datos textuales.
- Por ello, se propone a-TTR, una nueva métrica basada en texto, que mide la diversidad léxica ajustada para evaluar la calidad del clustering sin necesidad de etiquetas.

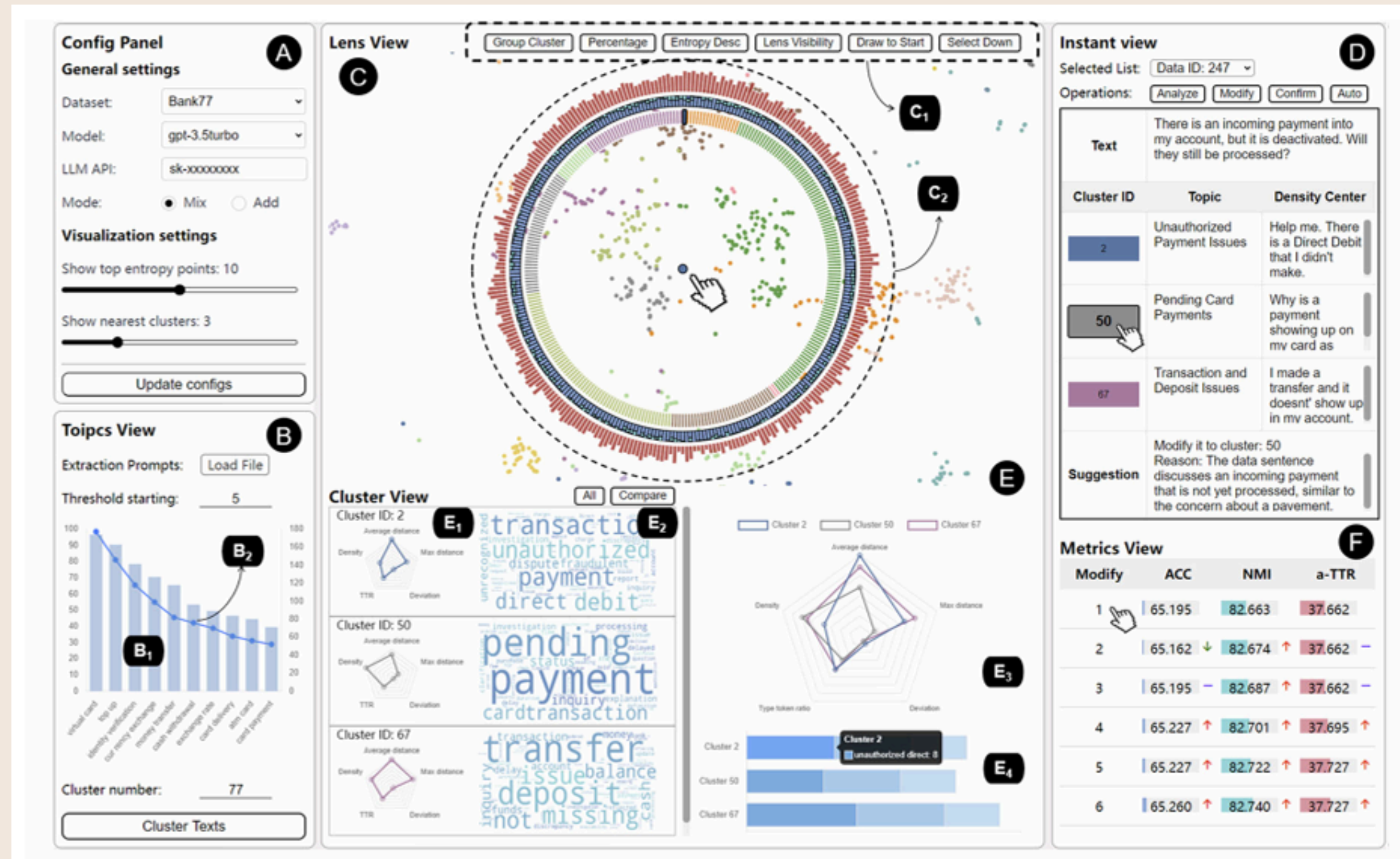
## LENS VIEW

- Muestra los resultados del clustering en una proyección t-SNE interactiva.
- Usa una lente con tres capas:
- Interna: Agrupación de puntos por ID de clúster.
- Media: Distribución de tópicos por punto, según un LLM.
- Externa: Entropía del clúster por punto, como indicador de ambigüedad.
- Soporta interacción avanzada: mover, hacer zoom, seleccionar con clic o forma libre.
- Botones permiten reorganizar la vista según agrupación, porcentaje de tópicos o entropía.

## INSTANT VIEW

- Muestra información detallada de los puntos seleccionados desde la lente.
- Permite ver el texto original y compararlo con los 3 clústeres más cercanos.
- El modelo de lenguaje puede sugerir modificaciones al agrupamiento con base en similitud semántica.
- Se pueden modificar etiquetas manualmente o automáticamente con la opción Auto.

# DISEÑO VISUAL





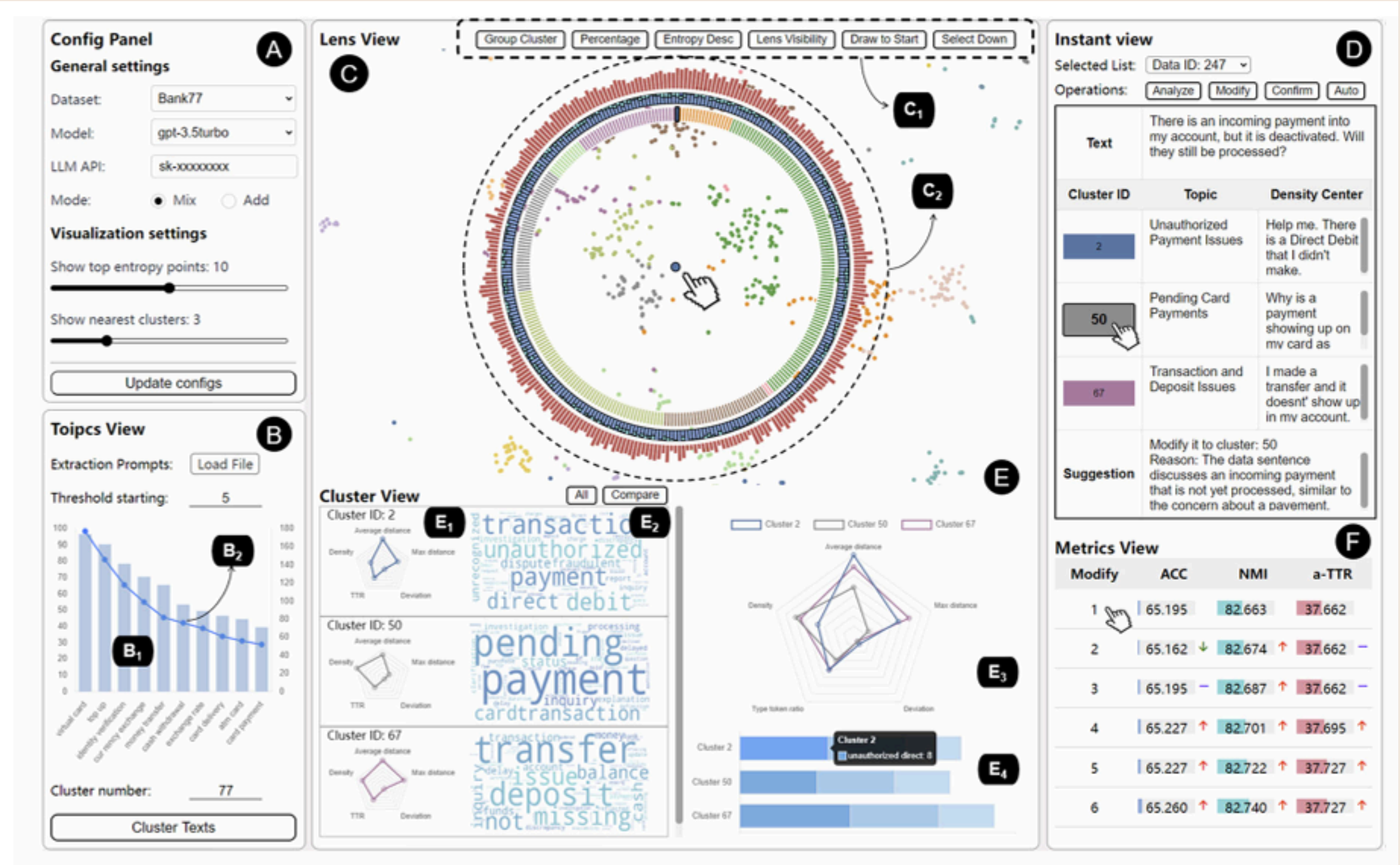
CLUSTER VIEW

- Vista izquierda: Muestra todos los clústeres mediante radar charts (5 métricas normalizadas) y nubes de palabras.
- Vista derecha: Compara los tres clústeres más cercanos al punto seleccionado, tanto en métricas como en tópicos frecuentes.
- Ayuda a analizar la coherencia interna de los clústeres y detectar anomalías.

METRICS VIEW

- Visualiza los cambios en métricas (ACC, NMI, a-TTR) tras cada modificación.
- Usa barras de progreso y flechas de comparación entre filas para facilitar la evaluación del impacto.
- Permite evaluar si las modificaciones mejoraron la calidad del clustering.

DISEÑO VISUAL





# CONCLUSIONES


Se midió la efectividad con métricas como accuracy y NMI, obteniendo resultados superiores a métodos previos como KMeans, PCKMeans y Keyphrase Clustering.

En cuanto a costos, TextLens reduce significativamente el uso de la API de GPT-3.5-turbo, ya que lo emplea solo para tareas específicas como extracción de temas y sugerencias sobre datos anómalos. Comparado con otros métodos que usan GPT en todo el proceso, los costos se reducen hasta 10 veces.

Limitaciones identificadas incluyen:

- Pérdida de información al usar t-SNE para visualizar los datos en 2D.
- Dificultades para manejar textos largos, lo que afecta la calidad de las respuestas de GPT y eleva los costos.

Futuro trabajo:

- Mejorar las técnicas de visualización para minimizar pérdida de información.
  - Adaptar el sistema a textos más extensos y variados para ampliar su aplicabilidad.
- 

# ANÁLISIS VISUAL PARA MODELOS DE CLASIFICACIÓN DE TEXTO DE GRAN FINEZA Y SUS CONJUNTOS DE DATOS

## Introducción

En NLP, la clasificación de texto se ha vuelto más compleja por la creciente cantidad y sutileza de las etiquetas, dificultando la interpretación de los modelos. Para abordar este reto, se desarrolló SemLa, un sistema de análisis visual creado junto a expertos, que facilita la comprensión de modelos de clasificación fina mediante visualizaciones interactivas y explicaciones detalladas.

El sistema SemLa permite:

- Identificar discrepancias entre los datos reales y lo que el modelo ha aprendido.
- Detectar patrones léxicos y conceptuales, incluidos sesgos presentes en los datos.
- Ofrecer explicaciones detalladas a nivel de muestra que muestran el significado de etiquetas sutiles.
- Proporcionar información sobre las relaciones entre clases.

Las contribuciones principales del trabajo son:

- El diseño de SemLa y sus componentes para la clasificación de texto de alta granularidad.
- La documentación del proceso de diseño iterativo con reflexiones y discusión.
- Una evaluación detallada basada en estudios de caso y retroalimentación de expertos.



# REQUERIMIENTOS

01

Comprensión jerárquica del razonamiento del modelo. Lograr una comprensión multinivel de las predicciones y errores del modelo, utilizando información general y específica.

02

Explicación de los aspectos específicos que distinguen etiquetas similares.

03

Descubrimiento de debilidades como características falsas, sesgos y la causa raíz de confusiones frecuentes.

04

Visualización de características semánticas de etiquetas.

Comprender relaciones entre etiquetas, conceptos detallados y subgrupos dentro de una misma etiqueta.

05

Evitar interpretaciones erróneas.

Representar fielmente el razonamiento del modelo, minimizando suposiciones y apoyando una evaluación crítica por parte del usuario.



# PIPELINE DE SEMLA

## 01. Nivel Global

Estas tareas abordan el análisis del modelo en su conjunto:

- T1: Identificar patrones generales en el espacio de codificación: Relacionado con R1. Permite observar cómo el modelo organiza la información a nivel general para validar su comportamiento y detectar patrones sistemáticos.
- T2: Detectar áreas de debilidad del modelo: Enfocado en R3. Esta tarea ayuda a ubicar fallos del modelo, como errores frecuentes, sesgos o patrones erróneos, para facilitar su depuración.
- T3: Comparar modelos: También relacionado con R1. Permite comparar diferentes modelos entre sí a un nivel alto para identificar cuál tiene un mejor rendimiento o una organización más coherente de los datos.



## 02. Nivel de Etiqueta

Estas tareas se centran en el análisis detallado de clases o etiquetas específicas:

- T4: Explicar y resaltar fronteras de decisión entre etiquetas: Apoya R4 (visualización de características semánticas de etiquetas). Esta tarea permite entender cómo el modelo diferencia entre etiquetas similares, lo cual es clave para la depuración y validación.
- T5: Explicar cómo el modelo interpreta (o malinterpreta) una etiqueta: Relacionada con R3. Permite detectar en qué casos el modelo falla al entender correctamente una clase, y por qué ocurre esa confusión.
- T6: Identificar similitudes entre grupos de etiquetas: Apoya también R4. Ayuda a detectar etiquetas que son semánticamente parecidas, lo cual es útil para reorganizar o reanotar el dataset.
- T7: Identificar subgrupos dentro de una misma etiqueta: También basado en R4. Permite descubrir sub-clases o patrones internos dentro de una sola etiqueta, lo cual puede revelar sesgos o necesidades de refinamiento en los datos.

# PIPELINE DE SEMLA

## 01. Nivel Global

Estas tareas abordan el análisis del modelo en su conjunto:

- T1: Identificar patrones generales en el espacio de codificación: Relacionado con R1. Permite observar cómo el modelo organiza la información a nivel general para validar su comportamiento y detectar patrones sistemáticos.
- T2: Detectar áreas de debilidad del modelo: Enfocado en R3. Esta tarea ayuda a ubicar fallos del modelo, como errores frecuentes, sesgos o patrones erróneos, para facilitar su depuración.
- T3: Comparar modelos: También relacionado con R1. Permite comparar diferentes modelos entre sí a un nivel alto para identificar cuál tiene un mejor rendimiento o una organización más coherente de los datos.



## 03. Nivel de Muestra

T8 Explicar la importancia de cada palabra a través de múltiples métricas: Esta tarea, crucial para la depuración del modelo, se alinea con la explicación de muestras individuales (R2) y la protección contra posibles impresiones falsas de una única métrica (R5).

T9 Proporcionar explicaciones contrastivas detalladas: En apoyo del requisito de explicaciones detalladas a nivel de muestra (R2) y la protección contra posibles impresiones falsas de la explicación general de la importancia de las características (R5), esta tarea se centra en el análisis para la depuración y la validación.

## 02. Nivel de Etiqueta

Estas tareas se centran en el análisis detallado de clases o etiquetas específicas:

- T4: Explicar y resaltar fronteras de decisión entre etiquetas: Apoya R4 (visualización de características semánticas de etiquetas). Esta tarea permite entender cómo el modelo diferencia entre etiquetas similares, lo cual es clave para la depuración y validación.
- T5: Explicar cómo el modelo interpreta (o malinterpreta) una etiqueta: Relacionada con R3. Permite detectar en qué casos el modelo falla al entender correctamente una clase, y por qué ocurre esa confusión.
- T6: Identificar similitudes entre grupos de etiquetas: Apoya también R4. Ayuda a detectar etiquetas que son semánticamente parecidas, lo cual es útil para reorganizar o reanotar el dataset.
- T7: Identificar subgrupos dentro de una misma etiqueta: También basado en R4. Permite descubrir sub-clases o patrones internos dentro de una sola etiqueta, lo cual puede revelar sesgos o necesidades de refinamiento en los datos.

## MAP VIEW

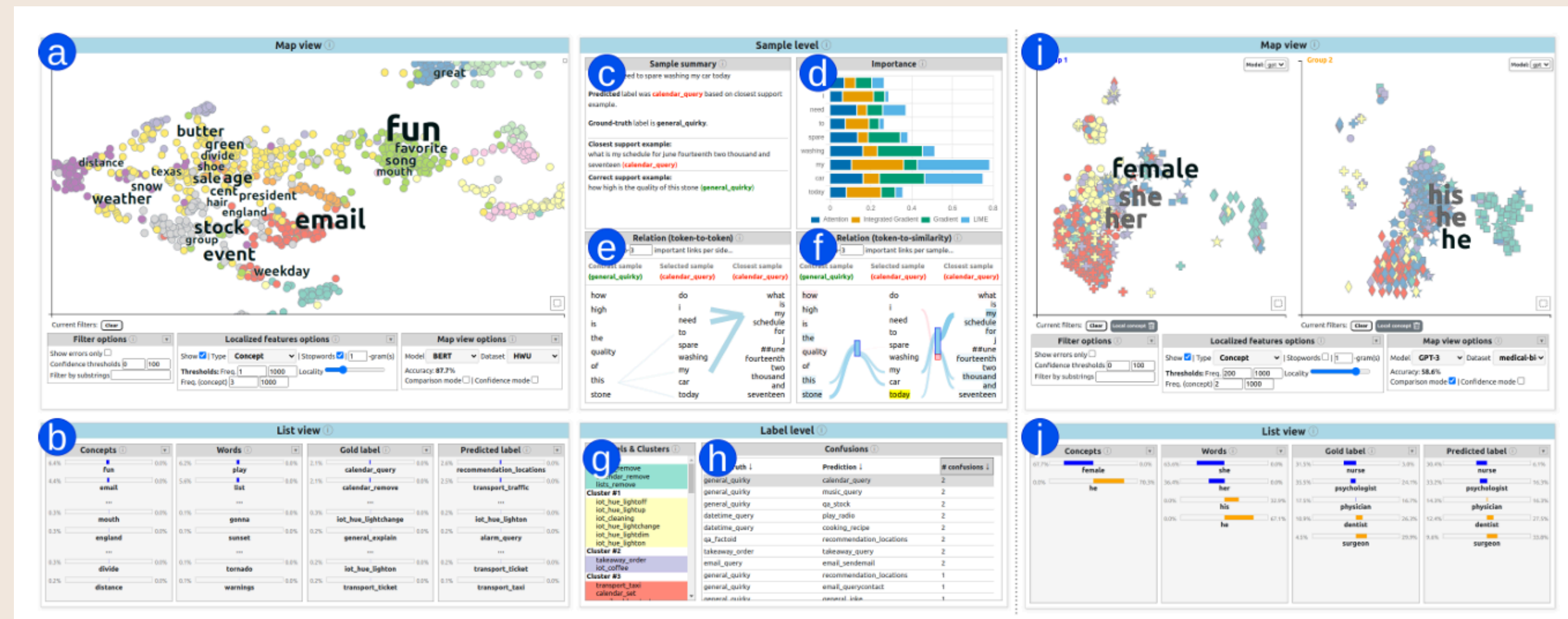
- Proyecta un corpus textual en un gráfico de dispersión 2D usando embeddings de un modelo y técnicas de reducción de dimensionalidad.
- Cada muestra se representa como un círculo (o con distintas formas para distinguir etiquetas).
- Para manejar el exceso de etiquetas, estas se agrupan en clústeres, codificados por colores.
- Incluye herramientas interactivas como zoom, paneo, filtros, y una visualización local de palabras para ayudar a entender el contenido semántico de diferentes zonas del mapa.

## LISTS VIEW

- Muestra un resumen clasificado de conceptos, palabras y etiquetas presentes en las muestras visibles actualmente en el Mapa.
- Soporta un modo de comparación, donde:
- El Mapa muestra dos gráficos de dispersión separados, permitiendo comparar dos grupos de muestras o el mismo corpus bajo dos modelos diferentes.
- La Vista de Listas indica qué elementos son comunes o exclusivos entre ambos grupos.

# DISEÑO VISUAL

SemLa es un sistema de análisis visual (Visual Analytics, VA) compuesto por cuatro vistas coordinadas que ayudan a explorar, entender y explicar modelos de clasificación de texto de grano fino





## SAMPLE-LEVEL VIEW

# DISEÑO VISUAL

SemLa es un sistema de análisis visual (Visual Analytics, VA) compuesto por cuatro vistas coordinadas que ayudan a explorar, entender y explicar modelos de clasificación de texto de grano fino

- Se activa cuando el usuario selecciona una muestra individual.
- Explica la predicción del modelo a través de tres visualizaciones:

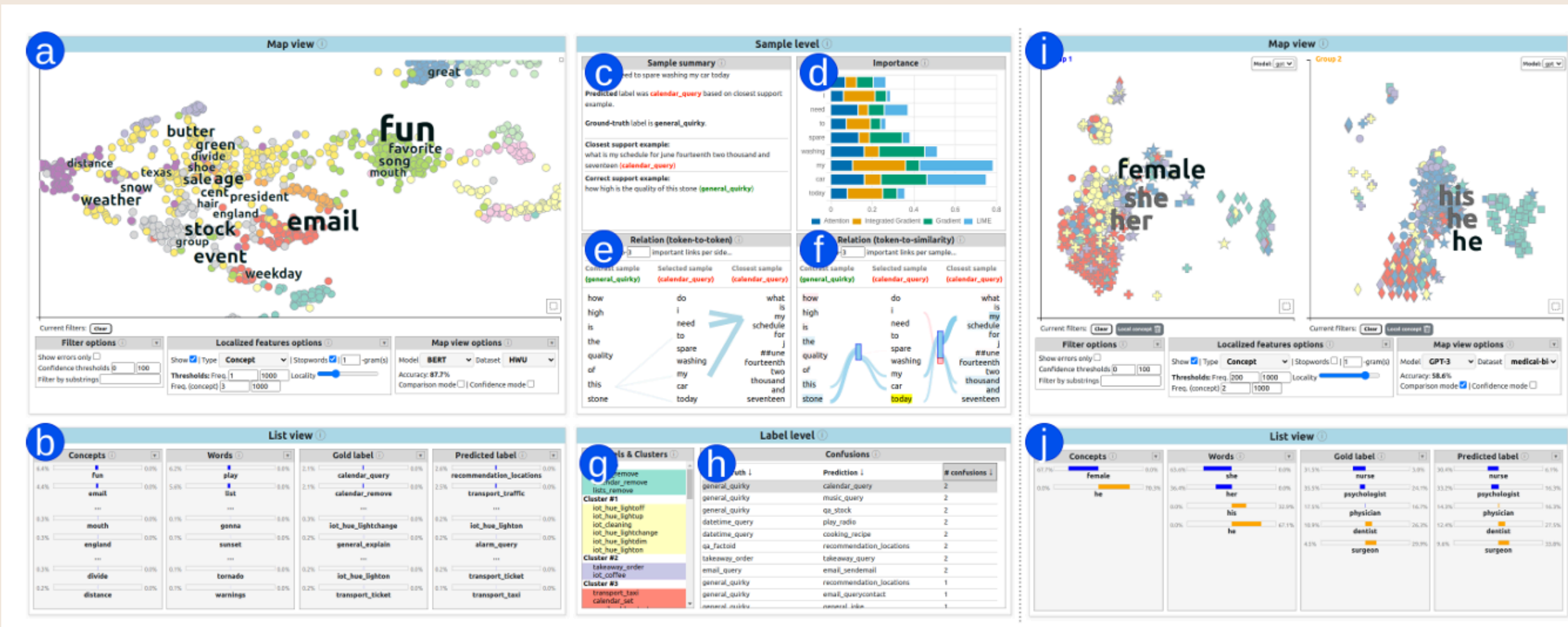
Las visualizaciones son:

- Resumen en lenguaje natural (Fig. 1c): explica la diferencia entre las dos etiquetas en lenguaje simple.
- Grafo de relaciones token a token (Fig. 1e): muestra conexiones directas entre las palabras de las tres muestras.
- Visualización de contribuciones (Fig. 1f): indica cuánto contribuye cada token a la similitud o diferencia entre las muestras.

## LABEL-LEVEL VIEW

## Incluye:

- Una lista de clústeres de etiquetas agrupadas por similitud.
- Una tabla de confusión que muestra qué pares de etiquetas son confundidos con mayor frecuencia.
- Esta tabla puede ordenarse por frecuencia de confusión, ayudando a identificar errores comunes del modelo.
- El usuario puede seleccionar etiquetas de esta vista para analizarlas en las vistas de Mapa y Listas.



# PALABRAS LOCALIZADAS



El análisis semántico de modelos con estructuras complejas y jerárquicas requiere identificar patrones en diferentes niveles. Las técnicas tradicionales, como el modelado de tópicos basado en BERT, dependen del agrupamiento (clustering) para identificar tópicos y luego extraer palabras clave mediante tf-idf. Sin embargo, este enfoque tiene dos problemas principales:

1. Alto costo computacional del clustering.
2. Suposiciones arbitrarias (como número de clústeres) que pueden distorsionar la interpretación.

LWC busca palabras que estén localizadas geográficamente en el espacio del modelo, es decir, que aparezcan frecuentemente en una zona específica, pero no en otras.

Cómo funciona:

- Toma como entrada un conjunto de muestras (D) y sus posiciones (PD) en el espacio del modelo.
- Para cada palabra, LWC:
- Registra las posiciones donde aparece.
- Calcula su “localidad” (área donde ocurre).
- Filtra las palabras por frecuencia (T) y tamaño de la localidad (R).
- Calcula el centro de esa localidad para posicionarla en el mapa (C).
- El resultado es un conjunto de palabras representativas de regiones específicas, que pueden visualizarse directamente sobre el gráfico de dispersión 2D.

Ventajas del enfoque:

- Eficiencia espacial: se evita generar múltiples nubes de palabras separadas.
- Evita repeticiones: no se repiten las mismas palabras en distintas regiones.
- Mejor integración visual: se superpone directamente sobre el mapa de muestras, manteniendo el contexto espacial.
- Aplicación recursiva: se puede usar LWC varias veces para obtener conceptos más abstractos a partir de palabras locales iniciales.

# CASO DESTACADO: CLINC150



La confusión más frecuente fue entre las etiquetas aparentemente no relacionadas: "vaccines" y "cancel\_reservation".

Exploración con SemLa:

- En la vista Map, no se halló una conexión evidente entre ambas etiquetas observando las local words.
- Cambiando a la vista de local concepts, se detectó una relación semántica oculta:
  - En las muestras con errores, aparecía la palabra “cuba”.
  - Aunque “cuba” no pertenecía a ninguna muestra con la etiqueta cancel\_reservation, el modelo la asociaba con esa etiqueta porque:
    - Aparecen menciones frecuentes de otros países en cancel\_reservation (ej.: “spain”, “mexico”, “china”, “zimbabwe”).
    - El modelo probablemente asoció “cuba” con otros nombres de países, generando así errores de clasificación.

Conclusión

- SemLa permitió descubrir automáticamente esta relación conceptual implícita (asociación de países), que de otro modo habría sido difícil detectar manualmente revisando los datos.
- Este caso evidencia cómo el sistema ayuda a revelar patrones semánticos sutiles que influyen en el comportamiento del modelo.



# CONCLUSIONES

## Fortalezas del enfoque

- Diseño iterativo centrado en expertos: La participación activa de expertos con distintos perfiles (desarrolladores, analistas, personas enfocadas en el cliente) aseguró que el sistema abordara necesidades diversas y reales.
- Múltiples niveles de análisis visual: SemLa ofrece vistas a nivel de mapa, lista, muestra y etiqueta, permitiendo un análisis detallado y flexible.
- Alta capacidad explicativa: A través de vistas como “Local Words” y “Relation Graphs”, se logran explicaciones contrastivas que superan a herramientas tradicionales como LIME o análisis de topics.
- Adaptabilidad y generalizabilidad: Los expertos valoraron que el sistema puede aplicarse a distintos dominios (diálogos, medicina, detección de abuso), gracias a su diseño modular y flexible.

## Desafíos y limitaciones identificadas

- Balance entre simplicidad y capacidad técnica: Algunos usuarios requerían mayor transparencia en los detalles técnicos detrás de las visualizaciones (por ejemplo, el cálculo de relaciones o la semántica de los ejes), lo que plantea el reto de mantener la usabilidad sin sacrificar profundidad.
- Necesidad de guía y documentación contextual: Aunque el sistema permite gran libertad exploratoria, se identificó la necesidad de guiar al usuario a través de tutoriales, configuraciones predeterminadas o flujos de trabajo sugeridos.
- Comparación de modelos limitada: Aunque se pueden cargar distintos modelos, los expertos sugirieron mejorar las capacidades para comparar directamente resultados entre checkpoints o versiones.





MUCHAS  
GRACIAS

