Análisis Visual de Problemas de Granularidad y Ambigüedad en el Agrupamiento de Intenciones con LLMs: Caso CLINC150

Cecilia del Pilar Vilca Alvites
Escuela Profesional Ciencia de la Computación
Universidad Nacional de San Agustín
Arequipa, Perú
cvilcaal@unsa.edu.pe

I. INTRODUCCIÓN

En la actualidad, la creciente cantidad de información en forma de texto representa un reto importante para el procesamiento de lenguaje natural (NLP). Esto es especialmente cierto en tareas no supervisadas, como el agrupamiento automático de frases según la intención del usuario, que es común en sistemas como asistentes virtuales o chatbots. Organizar y entender grandes volúmenes de texto sin etiquetas previas requiere métodos efectivos y comprensibles.

Los Modelos de Lenguaje de Gran Escala (LLMs) han revolucionado la manera en que representamos el significado del texto, generando embeddings que capturan el contexto y la semántica de las frases [2]. Estas representaciones permiten agrupar textos similares con mayor precisión. Sin embargo, los métodos basados en LLMs suelen funcionar como "cajas negras", lo que dificulta entender por qué ciertos textos se agrupan de una forma determinada y limita la interpretación de los resultados [1].

Un desafío importante surge de la alta granularidad presente en muchos conjuntos de datos de intenciones. Estos datos suelen contener numerosas categorías que, aunque diferentes, son semánticamente muy próximas [3]. Los embeddings generados por LLMs pueden agrupar frases con significados similares, pero a menudo no logran separar claramente intenciones que difieren de forma sutil, provocando solapamientos entre las clases. Por otro lado, la ambigüedad inherente al lenguaje natural añade complejidad, ya que frases idénticas o muy parecidas pueden estar asociadas a intenciones distintas, reflejando la riqueza y variabilidad del lenguaje humano [5]. Esta ambigüedad representa un reto difícil de resolver con métodos automáticos, pues requiere un análisis contextual profundo.

La dificultad para diagnosticar estos problemas se agrava debido a la falta de herramientas visuales interactivas que permitan a los analistas explorar y comprender los agrupamientos. Los sistemas existentes rara vez combinan visualización analítica con retroalimentación humana, lo que limita la detección de errores, la identificación de consultas anómalas y el ajuste manual de clústeres [1], [4]. Contar con recursos que faciliten esta interpretación es crucial para mejorar la calidad de los agrupamientos.

Ante estos retos, el presente proyecto propone el desarrollo de una herramienta visual interactiva que aproveche los embeddings generados por LLMs para analizar el dataset CLINC150, compuesto por conversaciones de usuarios. El objetivo es facilitar la exploración y diagnóstico de la estructura de los clústeres, identificar patrones problemáticos como la granularidad excesiva y la ambigüedad, y permitir la mejora iterativa de los resultados. Para ello, la herramienta integrará técnicas de análisis temático, incluirá métricas internas para evaluar la calidad de los grupos y ofrecerá una interfaz dinámica que permita al usuario modificar y refinar la cantidad y composición de los clústeres.

De esta manera, el proyecto busca aportar una solución que combine la potencia de los modelos avanzados de lenguaje con la capacidad humana para interpretar y mejorar los agrupamientos, contribuyendo a superar los desafíos inherentes al procesamiento no supervisado de intenciones en lenguaje natural.

II. TRABAJOS RELACIONADOS

El análisis visual de problemas de granularidad y ambigüedad en el agrupamiento de intenciones, especialmente en el contexto de Modelos de Lenguaje Grandes (LLMs) se basa en diversas áreas de investigación. Esta sección revisa los trabajos más relevantes que contribuyen a la comprensión y abordaje de estos desafíos.

A. Análisis Visual para la Interpretación de Modelos y Datos

La visualización juega un papel crucial en la interpretación y validación de los resultados del agrupamiento y la comprensión de los modelos de lenguaje. [1], en su trabajo sobre TextLens, proponen un marco de análisis visual que aprovecha los LLMs para mejorar el agrupamiento de texto. El diseño de TextLens se centra en facilitar la exploración interactiva de los clústeres generados por LLMs. Esto se logra mediante una interfaz que permite a los usuarios interactuar con las incrustaciones de LLM para refinar el agrupamiento, explorar la composición de los clústeres y comprender las relaciones entre

ellos. La capacidad de TextLens para visualizar y manipular estas incrustaciones de alta dimensión es clave para su utilidad en la interpretación de los resultados del agrupamiento.

De manera similar, [3] presenta SemLa (Semantic Landscape), un marco de análisis visual diseñado para modelos y conjuntos de datos de clasificación de texto de grano fino. SemLa aborda la complejidad semántica y la dificultad para diferenciar clases en entornos de grano fino. Su diseño de análisis visual permite a los usuarios explorar el espacio semántico de los datos y las decisiones del modelo a través de múltiples vistas interconectadas. Específicamente, SemLa incluye una vista de mapa para visualizar la distribución de las muestras y su proximidad semántica, una vista de lista para examinar muestras individuales y sus características, y vistas a nivel de etiqueta que muestran cómo las clases se relacionan entre sí y cómo se comportan los errores del modelo. La interacción entre estas vistas permite una comprensión profunda de las áreas de confusión y granularidad dentro del conjunto de datos.

Ambos trabajos son pertinentes para nuestro estudio, ya que se alinean con la meta de desarrollar herramientas de análisis visual para comprender mejor los desafíos de granularidad, solapamiento y ambigüedad. La capacidad de TextLens para interactuar con incrustaciones de LLM y facilitar la interpretación de clústeres, junto con la aproximación multifacética de SemLa para visualizar relaciones semánticas y errores de modelo, ofrecen metodologías y enfoques que pueden adaptarse para un análisis de agrupamiento de intenciones.

B. Agrupamiento de Texto y LLMs

El agrupamiento de texto es una tarea fundamental para organizar volúmenes crecientes de contenido digital y descubrir patrones ocultos. La efectividad de esta tarea depende en gran medida de la selección de las incrustaciones textuales y los algoritmos de agrupamiento. Avances recientes en los LLMs han demostrado un potencial significativo para mejorar esta tarea. Por ejemplo, [2] investigan cómo diferentes incrustaciones textuales, particularmente las utilizadas en LLMs, y varios algoritmos de agrupamiento influyen en los resultados del agrupamiento de conjuntos de datos de texto. Sus hallazgos indican que las incrustaciones de LLM son superiores, lo que subraya la importancia de estos modelos en el agrupamiento de texto. Complementariamente, [4] demuestran que los LLMs pueden superar las limitaciones de los enfoques de agrupamiento tradicionales al generar incrustaciones que capturan los matices semánticos de textos cortos. Su trabajo es especialmente relevante, ya que cuantifica el éxito del enfoque de agrupamiento utilizando revisores humanos y un LLM generativo, sugiriendo este último como un medio para cerrar la ”brecha de validación” entre la producción e interpretación de clústeres.

C. Granularidad y Ambigüedad en la Clasificación y Agrupamiento de Texto

A medida que las tareas de clasificación de texto se vuelven más finas, los conjuntos de datos se fragmentan en un mayor

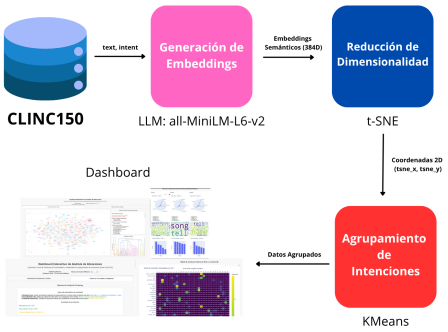


Fig. 1: Pipeline del Sistema de Análisis Visual de Agrupamiento de Intenciones

número de clases que son más difíciles de diferenciar, lo que conduce a estructuras semánticas más complejas y decisiones de modelo más difíciles de explicar. [3] abordan este desafío en la clasificación de texto de grano fino, destacando la necesidad de herramientas que puedan manejar la creciente complejidad semántica y la dificultad de diferenciar clases. Si bien su enfoque se centra en la clasificación, los problemas de granularidad y la dificultad de diferenciar categorías son directamente aplicables al agrupamiento de intenciones, donde la separación clara de intenciones similares es crucial.

La ambigüedad semántica es un problema persistente en el procesamiento del lenguaje natural, que afecta directamente la calidad del agrupamiento. El trabajo de [5] sobre el procesamiento de la polisemia semántica basada en la ontología de las palabras resalta la necesidad de un tratamiento automático de este fenómeno para mejorar las aplicaciones del procesamiento del lenguaje. Aunque su enfoque se centra en el árabe, los principios subyacentes de la ambigüedad y la polisemia son universalmente relevantes para el agrupamiento de intenciones, donde una única expresión puede tener múltiples interpretaciones intencionales. La superación de esta ambigüedad es esencial para formar clústeres de intenciones coherentes y bien definidos.

III. VISIÓN GENERAL

A. Descripción del Dataset

Para este proyecto, se utiliza el dataset CLINC150, una colección de consultas de usuario diseñada específicamente para la tarea de clasificación de intenciones en sistemas de diálogo. Publicado por investigadores de la Universidad de Michigan en 2019, este dataset se ha convertido en un referente estándar en la investigación de Procesamiento de Lenguaje Natural (NLP). Es ampliamente reconocido por su alta granularidad de intenciones, ya que cuenta con 150 categorías distintas (incluyendo la clase out_of_scope o OOS), lo que lo convierte en un desafío significativo para los modelos de comprensión del lenguaje. La tarea principal es la clasificación de intención, y todas las consultas están en idioma inglés, originalmente estructuradas en formato JSON ([”texto de la consulta”, ”etiqueta de la intención”]).

TABLE I: Atributos del Dataset CLINC150

Atributo	Descripción	Tipo de Dato	Rango / Valores Posibles
`text`	Representa la **consulta original del usuario** en lenguaje natural. Es la entrada textual que el sistema debe procesar para inferir la intención.	String	Cadenas de texto que varían en longitud desde 2 hasta 136 caracteres. Ejemplos incluyen "what is my account balance" o "can you please tell me how much money I have left in my primary checking account after deducting all pending transactions?".
`intent`	Es la **etiqueta de la intención** subyacente asociada a la consulta de texto. Sirve como la verdad fundamental (*ground truth*) para la clasificación.	String	**150 categorías de intención** distintas y finamente granularizadas. Incluye intenciones como `pay_bill`, `transfer`, `balance`, `greeting`, `goodbye`, `translate`, `money_transfer`, y la categoría `out_of_scope` (OOS).
`text_length`	Atributo derivado que representa la **longitud de la consulta de texto** en número de caracteres. Se utiliza para análisis exploratorio.	Entero	Valores entre 2 y 136 caracteres. La mayoría de las consultas se agrupan alrededor de los 39 caracteres, con una mediana de 37.
`split`	Indica a qué subconjunto pertenece la instancia, utilizado para la división estándar del dataset para entrenamiento, validación y prueba de modelos.	String	`train` (entrenamiento), `val` (validación), `test` (prueba).
`embeddings`	Atributo derivado, no original del dataset, pero crucial para este proyecto. Son las **representaciones numéricas densas** de cada consulta de texto, generadas por un Large Language Model (LLM) (específicamente, `all-MiniLM-L6-v2`).	Vector	Vector de 384 dimensiones. Cada valor en el vector es un número flotante.
`tsne_x`, `tsne_y`	Atributos derivados de la reducción de dimensionalidad de los *embeddings*. Representan las **coordenadas 2D** de cada consulta en un espacio visual, obtenidas mediante t-SNE, facilitando su visualización e interpretación.	Flotante	Valores que varían según la distribución en el espacio 2D, generalmente en un rango de números reales.

CLINC150 simula interacciones de usuarios con asistentes virtuales en diversos dominios como banca, viajes, clima, entretenimiento y más. Cada consulta representa una petición o una pregunta formulada por un usuario. El objetivo principal de este dataset es proporcionar un recurso robusto para entrenar y evaluar modelos capaces de comprender la intención subyacente detrás de estas consultas, a pesar de la variabilidad y ambigüedad natural del lenguaje humano.

La entidad fundamental de estudio en este dataset es la "consulta de usuario" o "frase de texto". Cada instancia en el dataset representa una interacción individual y aislada de un usuario con un sistema conversacional, cuya intención subyacente ha sido previamente etiquetada por expertos.

El dataset CLINC150 se distribuye en tres subconjuntos para facilitar el ciclo de vida del aprendizaje automático y asegurar una evaluación robusta:

- **Entrenamiento (train):** Contiene 15,000 consultas, con aproximadamente 100 ejemplos por cada una de las 150 intenciones.
- **Validación (val):** Contiene 3,000 consultas, con aproximadamente 20 ejemplos por cada intención.
- **Prueba (test):** Contiene 4,500 consultas, con aproximadamente 30 ejemplos por cada intención.

Esta distribución equitativa de las intenciones en todos los subconjuntos, sumado a un número considerable de muestras, asegura que los modelos puedan ser entrenados y evaluados de manera justa y representativa, evitando el sesgo hacia clases mayoritarias.

B. Pre-pocesamiento de datos

El dataset CLINC150, en su formato original, provee consultas de usuario (text) y sus respectivas intenciones (intent). Para adecuar estos datos a las necesidades del análisis visual y el agrupamiento basado en Modelos de Lenguaje Grandes (LLMs), se realizó un proceso de pre-procesamiento y enriquecimiento que implicó la derivación de nuevos atributos y la transformación de la información textual. Este proceso fue guiado por las observaciones iniciales de nuestro Análisis Exploratorio de Datos (AED).

Los pasos clave del pre-procesamiento fueron los siguientes:

1) Generación de Embeddings con LLMs: La entidad fundamental de estudio es la "consulta de usuario" (text). Dado que las características superficiales del texto, como la longitud, no se correlacionaron significativamente con la intención subyacente (como se observó en el AED, donde "una consulta más larga o más corta no predice con precisión su intención"), se determinó la necesidad de representaciones semánticas más ricas. Para ello, cada consulta de texto fue transformada en una representación numérica densa o embedding. Se utilizó el Large Language Model all-MiniLM-L6-v2 para generar estas incrustaciones. Este modelo fue elegido por su eficiencia y su capacidad para capturar los matices semánticos del texto, produciendo vectores de 384 dimensiones para cada consulta. Este paso es crucial ya que estos embeddings son la base numérica sobre la cual se realizará el agrupamiento.

2) Reducción de Dimensionalidad para Visualización (t-SNE): El análisis exploratorio visual de los embeddings de 384 dimensiones es inviable directamente. Para permitir la visualización de la estructura de agrupamiento y la identificación de problemas de granularidad y ambigüedad, se aplicó una técnica de reducción de dimensionalidad no lineal. Se utilizó t-Distributed Stochastic Neighbor Embedding (t-SNE) para proyectar los embeddings de 384 dimensiones a un espacio bidimensional (2D). Las coordenadas resultantes, denominadas tsne_x y tsne_y, permiten representar visualmente la proximidad semántica de las consultas. Aunque se observó una superposición significativa entre clases en el espacio t-SNE durante el AED, esta visualización es fundamental para identificar las áreas de granularidad fina y ambigüedad, lo que valida la necesidad de herramientas de análisis visual.
3) Cálculo de la Longitud del Texto: Como parte del análisis exploratorio, se derivó el atributo text_length, que representa la longitud de cada consulta en número de caracteres. Aunque el AED concluyó que este atributo no es un predictor directo de la intención, es útil para análisis exploratorios adicionales y para entender la distribución general de las consultas. La mayoría de las consultas se agrupan alrededor de los 39 caracteres, con una mediana de 37.
4) Preservación de la Estructura del Dataset: El dataset original CLINC150 ya viene pre-dividido en subconjuntos de train, val y test, con una distribución equilibrada de intenciones (aproximadamente 100, 20 y 30 ejemplos por intención respectivamente). Esta estructura se mantuvo intacta durante el pre-procesamiento para asegurar que las fases posteriores de análisis y evaluación se realicen sobre particiones consistentes y representativas, evitando sesgos.

El resultado de este proceso de pre-procesamiento es un dataset enriquecido que incluye los atributos originales (text, intent, split) y los atributos derivados (text_length, embeddings, tsne_x, tsne_y), listos para ser utilizados en el análisis visual y el estudio del agrupamiento de intenciones. La tabla I detalla la descripción de cada uno de estos atributos procesados.

C. Visión general del sistema

El sistema propuesto está diseñado para facilitar el análisis visual interactivo del agrupamiento de intenciones, con un enfoque particular en la identificación y comprensión de los desafíos asociados a la granularidad y la ambigüedad en datasets de lenguaje natural. La arquitectura del sistema se concibe como un pipeline que transforma las consultas de usuario en perspectivas interpretables, integrando capacidades de Modelos de Lenguaje Grandes (LLMs) con técnicas avanzadas de visualización.

El flujo de datos y procesamiento se estructura en las siguientes etapas principales:

1) Ingesta de Datos: El proceso inicia con la carga del dataset CLINC150, el cual proporciona consultas de usuario (text) y sus correspondientes intenciones (intent) como verdad fundamental. Este dataset se carga en sus divisiones predefinidas de entrenamiento, validación y prueba para mantener la consistencia en el análisis.
2) Representación Semántica (Embeddings de LLM): Para capturar las complejas relaciones semánticas inherentes a las consultas de usuario, cada text es transformado en una representación numérica densa o embedding. Este paso es fundamental, ya que los embeddings generados por un LLM (específicamente, all-MiniLM-L6-v2) permiten trasladar el texto a un espacio vectorial de alta dimensión (384 dimensiones), donde la proximidad entre vectores indica similitud semántica. Esta elección se basa en la capacidad probada de los LLMs para comprender los matices del lenguaje, superando las limitaciones de características superficiales.
3) Reducción de Dimensionalidad: Dado que los embeddings de alta dimensión no son directamente visualizables, se aplica la técnica de Reducción de Dimensionalidad de t-Distributed Stochastic Neighbor Embedding (t-SNE). t-SNE proyecta estos vectores de 384 dimensiones a un espacio bidimensional (2D), generando las coordenadas tsne_x y tsne_y. El objetivo de este paso es preservar la estructura local y global de los datos en el espacio de alta dimensión, permitiendo que las similitudes semánticas se reflejen espacialmente en la visualización, a pesar de la posible superposición entre clases observada en la fase exploratoria.
4) Agrupamiento Automático (KMeans): Sobre las representaciones bidimensionales generadas por t-SNE, se aplica el algoritmo de agrupamiento KMeans. Este algoritmo organiza las consultas en un número predefinido de clusters, asignando a cada consulta un cluster_id. El propósito de esta etapa es identificar agrupaciones naturales en el espacio semántico, las cuales pueden ser comparadas y contrastadas con las intenciones reales para evaluar la coherencia del agrupamiento y detectar soplamientos.
5) Análisis Visual Interactivo: La etapa final y central del pipeline es un dashboard interactivo diseñado para la exploración y la interpretación de los resultados del agrupamiento. Esta herramienta permite a los usuarios:

- Visualizar la distribución de las consultas en el espacio 2D, coloreadas por intención real o por el cluster_id asignado.
- Inspeccionar detalles de consultas individuales o conjuntos seleccionados, mostrando su texto, intención y cluster.
- Generar dendrogramas jerárquicos para analizar la similitud y granularidad a un nivel más fino dentro de regiones específicas.
- Evaluar la composición de clusters específicos mediante métricas, nubes de palabras y gráficos de

Fig. 2: Dashboard Interactivo de Análisis de Intenciones

tópicos principales, facilitando la identificación de ambigüedad y granularidad.

- Acceder a métricas de calidad de clustering y una matriz de confusión para una evaluación cuantitativa y cualitativa del rendimiento del agrupamiento.

IV. Diseño del Sistema

La herramienta central de este proyecto es un dashboard interactivo, implementado utilizando el framework Dash de Plotly. Este dashboard está meticulosamente diseñado para permitir a los usuarios explorar, analizar y obtener perspectivas sobre el agrupamiento de intenciones, abordando directamente los desafíos de granularidad y ambigüedad presentes en el dataset CLINC150. El diseño se enfoca en proporcionar múltiples vistas interconectadas, cada una con un propósito analítico específico y un esquema de codificación visual claro.

El dashboard se organiza en varias pestañas y paneles, facilitando una exploración estructurada de los datos. A continuación, se describe cada una de sus secciones principales y el encoding visual aplicado:

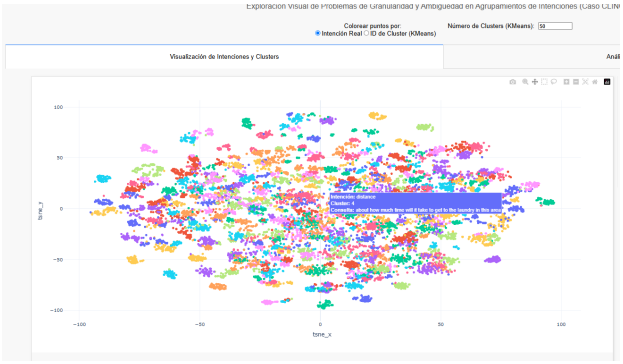
A. *Visualización Principal: Gráfico de Dispersión (t-SNE)*

Este es el punto de entrada visual para el análisis, representando la estructura semántica del dataset en un espacio bidimensional reducido. Ofrece una visión general de la distribución de las consultas de usuario, permitiendo identificar patrones de agrupamiento y áreas de superposición entre intenciones o clusters. Facilita la detección visual de granularidad (intenciones muy cercanas) y ambigüedad (puntos mezclados entre diferentes etiquetas).

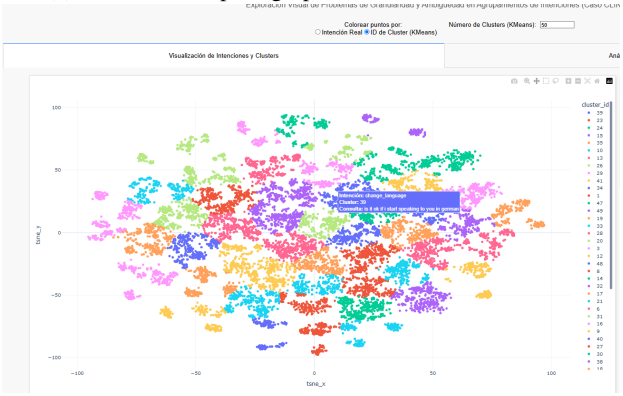
Las coordenadas tsne_x y tsne_y se mapean directamente a las posiciones horizontal y vertical de cada punto en el gráfico. La proximidad espacial en este gráfico indica similitud semántica de las consultas originales. Si se selecciona la opción ”Intención Real”, el color de cada punto representa la intención original del dataset, permitiendo observar cómo se agrupan y distribuyen las clases verdaderas en el espacio proyectado. Si se selecciona ”ID de Cluster (KMeans)”, el color codifica el cluster_id asignado por el algoritmo KMeans, mostrando los agrupamientos detectados. Al pasar el cursor sobre un punto, se despliega una tarjeta que codifica información detallada de la consulta original: text (la consulta completa), intent (su intención real), cluster_id (el cluster asignado), y id (identificador único del punto). Esto permite la inspección granular de cada instancia.

El gráfico permite la selección de puntos mediante arrastre (selección de lazo o rectangular) o clic individual, lo que activa el panel de detalles de selección y el dendrograma. El modo de arrastre (dragmode=’pan’) facilita la navegación.

Un campo de entrada (num-clusters-input) permite al usuario ajustar dinámicamente el número de clusters para el algoritmo KMeans (de 2 a 150), observando cómo los agrupamientos y la matriz de confusión se adaptan a este cambio.



(a) Visualización por agrupamiento de Intención Real



(b) Visualización por agrupamiento de ID de Cluster

Fig. 3: Comparación de resultados de detección.

B. *Panel de Detalles de Selección*

Este panel proporciona un resumen estadístico y textual de los puntos seleccionados en el gráfico principal, ya sea que provengan de una selección manual del usuario o de un cluster específico generado automáticamente. Su propósito principal es ofrecer una comprensión tanto cuantitativa como cualitativa de la composición del subconjunto de datos seleccionado. De esta manera, permite evaluar dos aspectos clave: la ambigüedad, entendida como la presencia de múltiples intenciones dentro de una misma selección o cluster, y la granularidad, que se refiere a la distribución fina de las distintas intenciones presentes.

Para lograrlo, se utilizan tres atributos fundamentales de cada punto seleccionado: la intención real (intent), el iden-

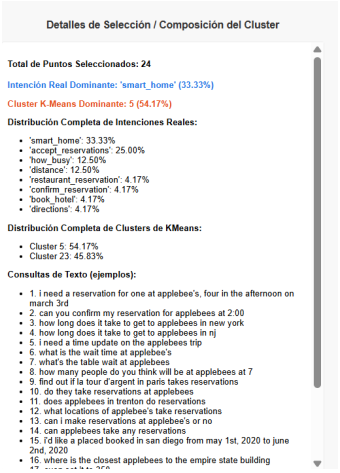


Fig. 4: Panel de Detalles de Selección

tificador de cluster (cluster_id) asignado por K-Means, y el texto original de la consulta (text). A partir de estos datos, el sistema genera distintos tipos de codificación visual que resumen y detallan la información.

En primer lugar, se muestra el número total de puntos seleccionados, brindando un indicador inmediato del tamaño del subconjunto analizado. Luego, se presentan las intenciones reales y clusters dominantes, es decir, aquellas categorías que aparecen con mayor frecuencia dentro de la selección. Estos se acompañan de sus respectivos porcentajes, facilitando una interpretación rápida del grado de homogeneidad o diversidad del grupo.

Adicionalmente, se incluyen listas completas de la distribución de intenciones reales y clusters K-Means, que indican tanto las etiquetas presentes como sus proporciones normalizadas. Esto permite al usuario identificar fácilmente patrones o excepciones dentro de la selección.

Finalmente, se presentan ejemplos textuales de las consultas correspondientes a los puntos seleccionados. Estos ejemplos permiten al usuario leer directamente las instancias originales del dataset, favoreciendo así una interpretación más rica y contextualizada de los patrones observados en la visualización estadística.

C. Dendrograma Jerárquico

Esta visualización se construye a partir de una selección de puntos específicos del gráfico principal y permite un análisis más detallado de las relaciones de similitud entre las consultas. Su principal propósito es revelar la estructura jerárquica de dichas similitudes, brindando una perspectiva de micro-granularidad. Esta es especialmente útil para identificar sub-agrupamientos o relaciones de proximidad semántica que podrían estar ocultas en vistas más agregadas. En contextos donde hay ambigüedad o mezcla de intenciones dentro de un grupo de puntos, el dendrograma actúa como una herramienta clave para descomponer estas complejidades.

Los datos utilizados para construir esta visualización provienen directamente de los textos (text) asociados a los

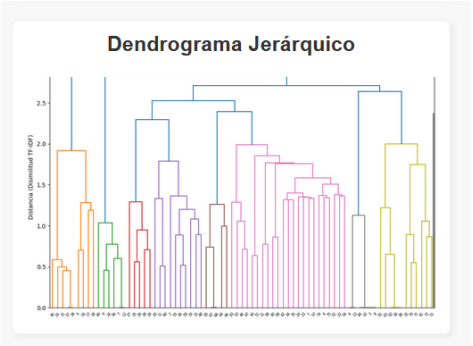


Fig. 5: Vista del dendograma Jerárquico

puntos seleccionados. A partir de ellos, se calcula una matriz de distancias semánticas empleando representaciones TF-IDF y se aplica un algoritmo de agrupamiento jerárquico.

Visualmente, el dendrograma se compone de ramas y nodos que representan las fusiones secuenciales entre pares de consultas o grupos de consultas. En el eje Y, la altura de cada fusión codifica el grado de disimilitud entre los elementos fusionados: cuanto más corta es la rama, mayor es la similitud entre los elementos unidos. Esto permite una lectura directa de la densidad semántica dentro de la selección. En el eje X, se encuentran las hojas del dendrograma, cada una representando una consulta individual. Por motivos de simplicidad visual, estas hojas están etiquetadas con índices numéricos, pero conceptualmente corresponden a las consultas originales seleccionadas.

Para la construcción del dendrograma se utiliza el método de enlace de Ward, una técnica que busca minimizar la varianza intra-cluster en cada etapa de la fusión. Este enfoque tiende a generar agrupamientos más compactos y balanceados, lo cual es especialmente adecuado para conjuntos de datos textuales con distribución densa y múltiples temas superpuestos. En conjunto, esta visualización proporciona una herramienta poderosa para explorar relaciones sutiles y estructuras emergentes en subconjuntos de datos seleccionados.

D. Vista de los 3 Clusters más Cercanos (Radar Chart, Word Cloud, Bar Chart)

Esta sección se activa cuando el usuario hace clic en un punto del gráfico principal y tiene como objetivo proporcionar un análisis multifacético de los clusters más relevantes asociados a dicho punto. Su propósito es ofrecer una caracterización profunda tanto de la calidad como del contenido semántico de los agrupamientos, centrándose especialmente en aquellos que están más vinculados con la consulta seleccionada.

Esta exploración es esencial para entender por qué se han formado ciertos clusters y en qué medida estos se diferencian (o se solapan) semánticamente. A través de esta visualización, se puede evaluar la granularidad del agrupamiento (cuán finamente están segmentadas las intenciones) y detectar posibles ambigüedades (consultas similares ubicadas en clusters diferentes, o clusters heterogéneos en su contenido). En conjunto, esta herramienta aporta un contexto interpretativo clave para

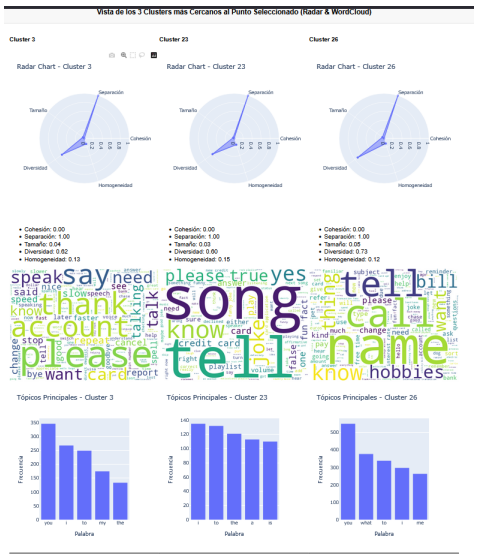


Fig. 6: Vista de los 3 Clusters más Cercanos (Radar Chart, Word Cloud, Bar Chart)

comprender la organización semántica del espacio de consultas y refinar los resultados del análisis.

1) Radar Chart (Gráfico de Araña): El gráfico de araña permite comparar visualmente múltiples métricas de calidad asociadas a un cluster específico. Se construye a partir de cinco indicadores principales: cohesión, separación, tamaño, diversidad y homogeneidad del cluster. Cada uno de estos indicadores se representa como un eje radial, y la distancia de los puntos desde el centro hasta los extremos codifica el valor normalizado de la métrica correspondiente (entre 0 y 1). Para mantener la coherencia interpretativa del gráfico bajo el criterio de ”más es mejor”, se invierte el valor de la cohesión. La forma y el área resultante al conectar los puntos en los ejes ofrecen una visión holística del perfil del cluster, facilitando la evaluación comparativa entre distintos agrupamientos.

2) Word Cloud (Nube de Palabras): La nube de palabras brinda una representación visual directa de los términos más frecuentes dentro de un cluster, facilitando una interpretación rápida de su contenido semántico. Se genera utilizando todos los textos de las consultas agrupadas. El tamaño de la fuente con que se representa cada palabra codifica su frecuencia relativa: términos más grandes aparecen con mayor frecuencia y, por lo tanto, son más representativos del tópico dominante. El color y la posición de las palabras se asignan principalmente con fines estéticos y de legibilidad, sin codificar variables adicionales.

3) Gráfico de Barras de Tópicos Principales: Para complementar la nube de palabras, se incluye un gráfico de barras que presenta cuantitativamente las cinco palabras más frecuentes del cluster junto a sus conteos absolutos. Este gráfico permite una comparación precisa entre los términos predominantes. En el eje X se ubican las palabras clave, mientras que el eje Y representa la frecuencia absoluta con la que cada una aparece en las consultas del cluster, reflejada en la altura de las barras. Esta visualización es especialmente útil para verificar rápidamente qué términos dominan el contenido textual de un cluster determinado.

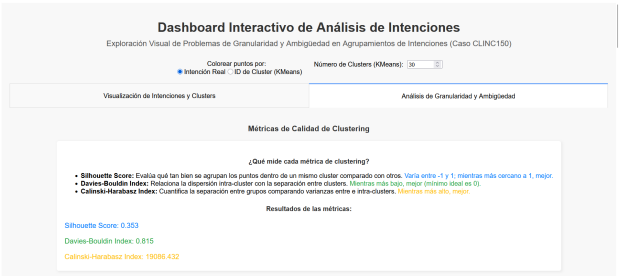
E. Pestaña de Análisis de Granularidad y Ambigüedad

Esta sección del sistema está diseñada para ofrecer una evaluación cuantitativa del rendimiento del agrupamiento, así como su alineación con las intenciones reales presentes en el dataset. Permite al usuario diagnosticar tanto la coherencia interna de los clusters formados como su fidelidad respecto a las etiquetas de verdad fundamental.

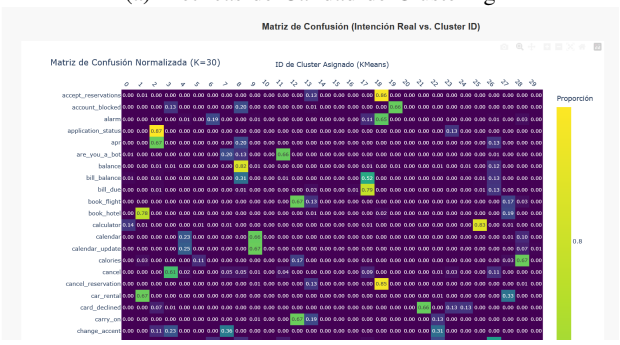
1) Métricas de Calidad de Clustering: En esta subsección se presentan métricas que evalúan la calidad intrínseca del agrupamiento realizado por el algoritmo KMeans, sin tener en cuenta las etiquetas reales. Para ello se utilizan las coordenadas proyectadas (tsne_x y tsne_y) y los identificadores de cluster (cluster_id) asignados a cada punto del dataset. Se calculan tres métricas fundamentales: el Silhouette Score, el Davies-Bouldin Index y el Calinski-Harabasz Index. Estos valores se muestran textualmente acompañados de una breve explicación interpretativa. Por ejemplo, un valor de Silhouette más cercano a 1 indica una mejor separación entre clusters, mientras que un Davies-Bouldin más bajo sugiere menor solapamiento entre ellos. Esta presentación facilita una comprensión rápida de la estructura general del agrupamiento sin necesidad de referencias externas.

2) Matriz de Confusión (Intención Real vs. Cluster ID): La matriz de confusión visualiza la correspondencia entre las intenciones reales (ground truth) y los clusters producidos por KMeans. Esta representación es esencial para analizar dos fenómenos clave en el análisis de agrupamiento semántico: la ambigüedad, que se presenta cuando una intención específica se fragmenta en múltiples clusters, y la granularidad, que ocurre cuando un único cluster contiene múltiples intenciones distintas. Cada fila de la matriz corresponde a una intención real, mientras que las columnas representan los IDs de clusters asignados. La intensidad del color en cada celda, usando una escala continua como ”Viridis”, codifica la proporción normalizada de ejemplos presentes en esa combinación de intención y cluster. Además, cada celda incluye el valor numérico correspondiente, formateado a dos decimales, que indica la proporción exacta. Esta codificación combinada (color y número) permite al usuario identificar rápidamente patrones de agrupamiento erróneo o difuso, y valorar la fidelidad semántica del sistema de clustering frente a la verdad fundamental del dataset.

Este diseño integral del dashboard, articulado en múltiples vistas interconectadas y sustentado por esquemas de codificación visual cuidadosamente seleccionados, proporciona una plataforma robusta para el análisis detallado de los agrupamientos generados por modelos de lenguaje de gran escala (LLMs). Más allá de visualizar los resultados, el sistema permite explorar de forma interactiva las relaciones semánticas entre consultas, identificar patrones de cohesión o dispersión en los clusters, y evaluar cuantitativamente su calidad mediante métricas internas y externas.



(a) Métricas de Calidad de Clustering



(b) Visualización por agrupamiento de ID de Cluster

Fig. 7: Matriz de Confusión (Intención Real vs. Cluster ID)

Particularmente valioso es el enfoque explícito en los fenómenos de granularidad y ambigüedad, dos desafíos recurrentes en tareas de clasificación y agrupamiento semántico. Al ofrecer herramientas visuales como el dendrograma jerárquico, la matriz de confusión y el análisis de perfiles de clusters, el sistema no solo revela los aciertos del modelo, sino que también expone sus limitaciones de forma comprensible y accionable. Esto ayuda a que el usuario pueda tomar decisiones informadas en la mejora iterativa del pipeline de NLP, ya sea ajustando el número de clusters, refinando representaciones vectoriales, o incorporando supervisión adicional.

REFERENCES

[1] R. Peng, Y. Dong, G. Li, D. Tian, and G. Shan, "TextLens: Large language models-powered visual analytics enhancing text clustering," *Journal of Intelligent & Fuzzy Systems*, DOI: 10.1007/s12650-025-01043-y, Feb. 2025.

[2] A. Petukhova, J. Carvalho, and N. Fachada, "Text Clustering with Large Language Model Embeddings," *International Journal of Cognitive Computing in Engineering*, vol. 6, pp. 100–108, Dec. 2025.

[3] M. Battogtokh, Y. Xing, C. Davidescu, A. Abdul-Rahman, M. Luck, R. Borgo "Visual Analytics for Fine-grained Text Classification Models and Datasets," *arXiv preprint arXiv:2405.02980*, 2024.

[4] L. K. Miller and C. P. Alexander, "Human-interpretable clustering of short text using large language models," *Royal Society Open Science*, vol. 12, no. 2, pp. 241088, 2025.

[5] S. Hamada, "Processing of Semantic Ambiguity Based on Words Ontology," *Journal of Computer Science*, vol. 16, no. 1, pp. 1–9, 2020.