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. MOTIVACIÓN Y JUSTIFICACIÓN

En la actualidad, donde cada vez se genera más información en forma de texto, organizar y entender grandes cantidades de datos textuales se ha vuelto un reto importante dentro del campo del procesamiento de lenguaje natural (NLP). Esto es especialmente difícil en tareas que no tienen datos previamente etiquetados, como es el caso del agrupamiento automático de frases por intención, algo muy común en sistemas como asistentes virtuales o chatbots.

El avance de los Modelos de Lenguaje Grandes (LLMs) ha revolucionado la representación semántica del texto, generando embeddings potentes que codifican el significado contextual de las frases de manera efectiva [2]. Estos embeddings han demostrado un potencial prometedor para mejorar la efectividad general del agrupamiento de texto. Sin embargo, a pesar de estas capacidades, los enfoques de clustering impulsados por LLMs a menudo operan como ”cajas negras”, lo que resulta en una baja interpretabilidad de los procesos de agrupamiento y de los resultados obtenidos [1]. La comprensión humana de por qué ciertos textos se agrupan de una manera específica aún requiere un esfuerzo considerable.

Frente a este problema de interpretabilidad, se presentan desafíos más profundos que están directamente relacionados con la naturaleza de los datos utilizados en la clasificación de intenciones. En este contexto, una intención se refiere al objetivo o propósito subyacente de una consulta formulada por el usuario, es decir, lo que el usuario realmente quiere lograr al interactuar con un sistema conversacional. Un desafío crítico es la alta granularidad de estas intenciones, presente en muchos conjuntos de datos que recopilan conversaciones de usuarios. Estos datasets suelen incluir un gran número de categorías de intención que, aunque distintas, son semánticamente muy cercanas entre sí [3]. Aunque los embeddings generados por modelos de lenguaje de gran escala (LLMs) pueden agrupar frases con significados similares, a menudo no logran distinguir adecuadamente entre intenciones que difieren de forma sutil, lo que provoca un solapamiento significativo de clases dentro de los clusters. Además, el lenguaje natural humano presenta una ambigüedad semántica intrínseca, en la que frases idénticas o muy similares pueden estar asociadas a intenciones completamente diferentes. Este fenómeno plantea un gran desafío para cualquier enfoque de categorización automática [5], ya que no se trata de un defecto en los datos, sino de un reflejo auténtico de la complejidad y riqueza del lenguaje humano.

La dificultad de diagnosticar y comprender estas problemáticas (la granularidad que lleva al solapamiento y la ambigüedad inherente) se ve agravada por la falta de herramientas visuales interactivas. Los sistemas de análisis actuales rara vez ofrecen interfaces que combinen una visualización analítica robusta con la retroalimentación humana, lo que dificulta la detección de errores de agrupamiento, la identificación de consultas anómalas o el refinamiento manual de los clusters a nivel de diagnóstico [1]. En este contexto, la necesidad de lograr un agrupamiento humanamente interpretable se vuelve primordial [4].

En respuesta a estos desafíos, se justifica la creación de un enfoque de análisis visual que, potenciado por los LLMs, permita explorar y diagnosticar de manera efectiva la estructura y las inconsistencias de los agrupamientos de intenciones. Este proyecto busca proporcionar una herramienta que facilite la comprensión de los problemas de granularidad y ambigüedad en los datos de intenciones, permitiendo a los analistas discernir patrones problemáticos y la calidad de los clusters generados.

II. PROBLEMA

El agrupamiento de textos en lenguaje natural, específicamente las intenciones de usuario, representa un desafío significativo en el ámbito del procesamiento de lenguaje no supervisado. Este desafío se fundamenta en dos propiedades de los conjuntos de datos de intenciones: su alta granularidad y la ambigüedad semántica presente en sus consultas.

En primer lugar, los conjuntos de datos de intenciones a menudo se caracterizan por incluir un elevado número de categorías de intención distintas, muchas de las cuales son específicas y semánticamente muy próximas entre sí. Esta alta granularidad de intenciones constituye una dificultad

considerable. Si bien los modelos de lenguaje grandes (LLMs) son capaces de generar embeddings potentes que codifican el significado contextual del texto, estas representaciones no siempre logran discriminar de forma óptima entre intenciones tan sutilmente diferenciadas. Como consecuencia, las visualizaciones de embeddings a menudo revelan un solapamiento considerable entre clases relacionadas.

En segundo lugar, estos conjuntos de datos también presentan cierta ambigüedad en el significado de algunas frases. Hemos encontrado casos en los que dos consultas son iguales o muy similares, pero están etiquetadas con intenciones distintas. Esto refleja una característica común del lenguaje humano: una misma frase puede tener diferentes significados según el contexto. Esta situación representa un desafío para cualquier modelo automático, ya que sin información adicional, distinguir entre esas intenciones se vuelve muy difícil.

La dificultad para resolver estos problemas —como el solapamiento entre intenciones similares y la ambigüedad en algunas frases— se complica aún más porque los resultados del agrupamiento automático no siempre son fáciles de interpretar. Además, no existen muchas herramientas visuales e interactivas que ayuden a un analista a explorar y entender estos agrupamientos de manera clara. En un escenario sin etiquetas, como en el clustering no supervisado, es muy importante contar con recursos que permitan ver fácilmente estos errores o confusiones, para así identificar patrones problemáticos como intenciones mezcladas o frases ambiguas dentro del conjunto de datos.

III. OBJETIVOS

El objetivo principal de este proyecto es crear una herramienta visual interactiva usando D3.js. Esta herramienta se basará en modelos de lenguaje avanzados (LLMs) para analizar el dataset CLINC150, que contiene conversaciones de usuarios. Buscamos que esta herramienta nos ayude a entender mejor los datos de texto, evaluar cómo se agrupan las intenciones de los usuarios y permitir que se puedan mejorar esos grupos.

Este visualizador está diseñado para solucionar los problemas más comunes al agrupar textos de forma automática en el procesamiento de lenguaje natural. Entre estos problemas, destacamos: que es difícil entender por qué los textos se agrupan de cierta manera (baja interpretabilidad), la dificultad para encontrar los temas principales de forma automática, y la falta de formas interactivas para detectar errores o ajustar los resultados manualmente. Para lograr esto, la herramienta usará los ”embeddings” (representaciones numéricas de los textos) generados por los LLMs, podrá encontrar los temas ocultos en los textos, incluirá maneras de medir la calidad de los grupos, y mostrará la información de forma dinámica. Todo esto con el fin de que los usuarios puedan analizar la estructura de los temas de los textos agrupados de forma sencilla y efectiva.

A. OBJETIVOS GENERALES

Para alcanzar nuestro objetivo general, nos hemos propuesto los siguientes pasos específicos:

- Usar LLMs para encontrar temas importantes en los textos y así mejorar la forma en que se representan antes de agruparlos.
- Identificar textos inusuales dentro de los grupos, usando tanto el análisis de temas como la visualización.
- Implementar métricas internas para evaluar la calidad de los clústeres, considerando coherencia semántica y separación.
- Diseñar e implementar una interfaz visual en D3.js que permita explorar los grupos de textos creados, cambiar cuántos grupos queremos ver y mejorar los resultados de manera interactiva.

IV. 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”]).

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.

TABLE I
ATRIBUTOS PRINCIPALES DEL DATASET CLINC150 Y SU SIGNIFICADO

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.

El dataset CLINC150, tras su preprocesamiento y enriquecimiento para este estudio, consta de los siguientes atributos principales, los cuales se describen detalladamente en la Tabla I.

V. PREGUNTAS

A. ¿Qué problemas identifican en el dataset?

Durante la fase de exploración y data wrangling del dataset CLINC150, hemos identificado desafíos clave que impactan el agrupamiento de intenciones, los cuales buscamos abordar.

Uno de estos problemas es la ambigüedad semántica del lenguaje natural. Hemos encontrado casos donde consultas de texto son idénticas en su formulación, pero corresponden a intenciones completamente distintas (ej., la misma frase para "modo susurro" y "cambiar volumen"). Esto presenta un reto significativo para los algoritmos de agrupamiento, ya que distinguir estas intenciones basándose solo en el texto es muy difícil, lo que puede llevar a errores en el clustering o a que estas consultas ambiguas se sitúen incorrectamente en los límites entre grupos.

Otro desafío importante es la granularidad fina y el solapamiento semántico de las 150 intenciones del dataset. Muchas categorías son muy específicas y semánticamente muy cercanas (ej., "enviar dinero" y "transferir dinero"). Aunque los embeddings generados por LLMs logran agrupar frases por significado, la visualización con t-SNE revela que los clusters resultantes a menudo mezclan consultas de diferentes intenciones reales, o que "nubes" de intenciones distintas se superponen. Esto significa que los algoritmos de clustering pueden tener dificultades para diferenciar claramente entre estas clases sutiles, generando agrupamientos menos puros y bien definidos.

B. ¿Qué descubrimos al analizar los datos?

Confirmamos que el dataset está bien estructurado en sus subconjuntos (entrenamiento, validación y prueba), con distribuciones de longitud de consultas notablemente similares entre ellos, lo que garantiza una evaluación imparcial del modelo. Además, se verificó la completitud de los datos, sin presencia de valores nulos, lo que simplifica el preprocesamiento. Un hallazgo crucial es el balance de clases altamente uniforme: las 150 intenciones están distribuidas equitativamente en todos los subconjuntos (aproximadamente 100 ejemplos en entrenamiento), lo que previene el sesgo del modelo hacia clases mayoritarias y es ideal para el entrenamiento y la evaluación de modelos de clasificación.

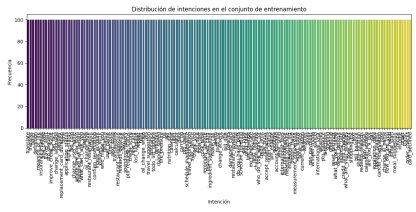

Fig. 1. Distribución de intenciones en el subconjunto train

En cuanto a la variabilidad de las consultas, se observó que la longitud del texto (text_length) oscila entre 2 y 136 caracteres, con la mayoría agrupada alrededor de los 39 caracteres, pero con una dispersión considerable. La distribución de la longitud es relativamente simétrica, aunque presenta algunas consultas excepcionalmente más largas (outliers). Este análisis de la longitud no reveló una correlación directa con la intención, lo que subraya la importancia de las representaciones semánticas profundas (como los embeddings de LLMs) sobre las características superficiales del texto para la comprensión de las intenciones.

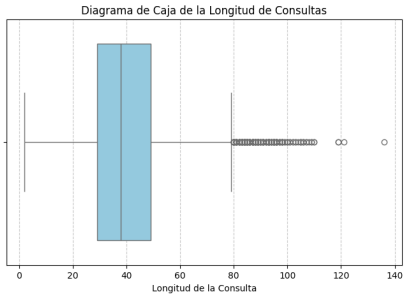

Fig. 2. Diagrama de Caja de la Longitud de Consultas en Train

C. *¿Que reflejan los patrones de tendencia?*

El análisis de los datos de CLINC150 no solo nos permitió identificar problemas, sino también observar patrones de tendencia clave que informan sobre la naturaleza del dataset y la complejidad de la tarea:

1) **Variabilidad en la Longitud de las Consultas:** Se observa una considerable dispersión en la longitud de las consultas (desviación estándar de aproximadamente 15.76 caracteres). Esto significa que, aunque hay una mayoría de frases de longitud media (con la moda en 30 y la media en 39.31 caracteres), el dataset incluye una gama diversa de interacciones, desde muy cortas hasta significativamente más largas. Esta variabilidad es un reflejo del uso natural y variado del lenguaje por parte de los usuarios. La distribución de longitudes presenta un ligero sesgo positivo (hacia la derecha) debido a la presencia de algunas consultas outlier que son considerablemente más largas que el promedio, como se muestra en la Figura 2.
2) **Equidad en la Representación de Intenciones (Balance de Clases):**Una tendencia notable es la fuerte uniformidad en la representación de las 150 intenciones. Cada clase está equilibrada, apareciendo aproximadamente el mismo número de veces en los diferentes subconjuntos del dataset (100 ejemplos en entrenamiento, 20 en validación y 30 en prueba). Este patrón asegura que los modelos no estén sesgados hacia intenciones más frecuentes, lo cual es fundamental para una evaluación justa y robusta, como se muestra en la Figura 1.
3) **Tendencia al Solapamiento Semántico de Intenciones:** El patrón más desafiante y relevante para este proyecto es la tendencia al solapamiento y la mezcla de intenciones (representadas por diferentes colores en las visualizaciones de embeddings) dentro y entre los clústeres. A pesar de que los LLMs logran agrupar frases con significado similar, la alta granularidad de las 150 intenciones —muchas de ellas sutiles o semánticamente muy cercanas— provoca que consultas de intenciones diferentes pero relacionadas se agrupen en la misma región del espacio vectorial. Este patrón de solapamiento es la justificación principal para la necesidad de una herramienta de analítica visual como TextLens, que permita "desenredar" y refinar estas agrupaciones complejas de manera interactiva, como se observa en la Figura 3.

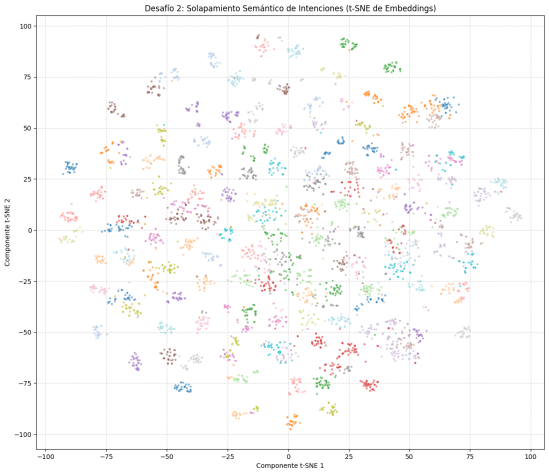

Fig. 3. Visualización del solapamiento semántico en el dataset CLINC150

D. *Como es afectado el comportamiento humano?*

El dataset CLINC150, siendo una colección de consultas de lenguaje natural generadas por humanos y posteriormente etiquetadas también por expertos humanos con intenciones específicas, actúa como un espejo directo de la complejidad y las particularidades del comportamiento humano en el uso del lenguaje.

Al analizar este dataset, hemos observado que la variabilidad en la longitud de las consultas, el vocabulario empleado y las distintas estructuras sintácticas reflejan fielmente cómo las personas formulan sus peticiones de maneras muy diversas. Hay usuarios que optan por frases muy concisas y directas, mientras que otros prefieren expresiones más elaboradas y detalladas. Esta diversidad inherente al lenguaje natural es un patrón común en la comunicación humana, lo que confiere al dataset un alto grado de realismo. Además, la presencia de consultas idénticas o muy similares asociadas a intenciones claramente diferentes (como el ejemplo de "turn up your volume" que puede significar tanto whisper_mode como change_volume) es una clara manifestación de la ambigüedad propia del lenguaje humano. Una misma frase puede poseer múltiples significados válidos dependiendo del contexto o de la verdadera intención del hablante. Este fenómeno no debe

ser interpretado como un ”error” en los datos, sino como una característica intrínseca y crucial del lenguaje que usamos diariamente. Reconocer esta ambigüedad es fundamental para el desarrollo de sistemas de comprensión de lenguaje natural, ya que la distinción de intenciones se vuelve un reto complejo que requiere no solo el análisis léxico, sino también la inferencia contextual o la interacción para su correcta resolución.

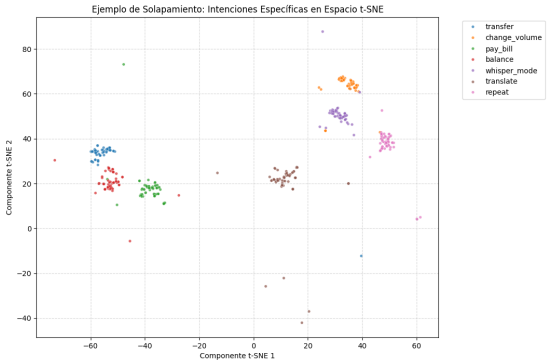

Fig. 4. Visualización de embeddings generados para un subconjunto de intenciones del dataset CLINC150. Se observa un solapamiento considerable entre categorías semánticamente cercanas, como whisper_mode, change_volume, repeat, y volume_up. A pesar de representar intenciones distintas, las frases asociadas tienden a ubicarse en regiones contiguas del espacio vectorial, lo que evidencia la ambigüedad del lenguaje natural y la dificultad de diferenciación entre intenciones sutilmente distintas.

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] N. Arias, P. Singh, and A. B. Imbert, “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.