Clasificación Semántica de Productos mediante Fusión de Representaciones Lingüísticas y Contextuales

Nelson Alfonso Beas Ham
Facultad de Ciencias Físico Matemáticas
Universidad Autónoma de Nuevo León
Materia: Procesamiento de Datos

4 de abril de 2025

1. Introducción

La organización automatizada de productos comerciales representa un desafío multidimensional que combina procesamiento lingüístico, análisis semántico y técnicas de aprendizaje no supervisado. Este proyecto propone un enfoque híbrido que integra representaciones vectoriales del lenguaje natural con metadatos estructurales para generar un sistema de clasificación adaptable a catálogos comerciales dinámicos.

2. Marco Metodológico

2.1. Ingeniería Semántica de Características

El preprocesamiento textual se fundamenta en tres principios:

- **Normalización léxica**: Reducción de variantes morfológicas mediante lematización conservando la semántica original
- **Filtrado contextual**: Eliminación de términos no descriptivos (indicadores de empaque, promociones) que introducen ruido conceptual
- **Enriquecimiento atributivo**: Detección de propiedades nutricionales mediante patrones léxicos específicos

2.2. Fusión Multimodal de Representaciones

La arquitectura de características combina:

- **Embeddings semánticos**: Capturan relaciones contextuales y similitudes conceptuales entre productos
- **Señales estructurales**: Codificación de relaciones jerárquicas (departamentos, pasillos) como vectores dispersos
- **Atributos derivados**: Marcadores booleanos para propiedades especiales (bajo en grasa)

2.3. Estrategia de Agrupamiento Adaptativo

Se implementa un enfoque escalable basado en:

- **Optimización iterativa**: Procesamiento por lotes para manejar espacios de alta dimensionalidad
- **Balance semántico-estructural**: Pesado implícito de características mediante normalización escalada
- **Jerarquización implícita**: Alta granularidad (10,000 clusters) para posterior consolidación semántica

3. Análisis de Resultados

3.1. Evaluación Cuantitativa

La métrica de silueta (0.42) sugiere una estructura de agrupamiento con solapamiento moderado, esperado en dominios comerciales donde productos pueden pertenecer a múltiples categorías simultáneamente.

3.2. Limitaciones de la Reducción Dimensional

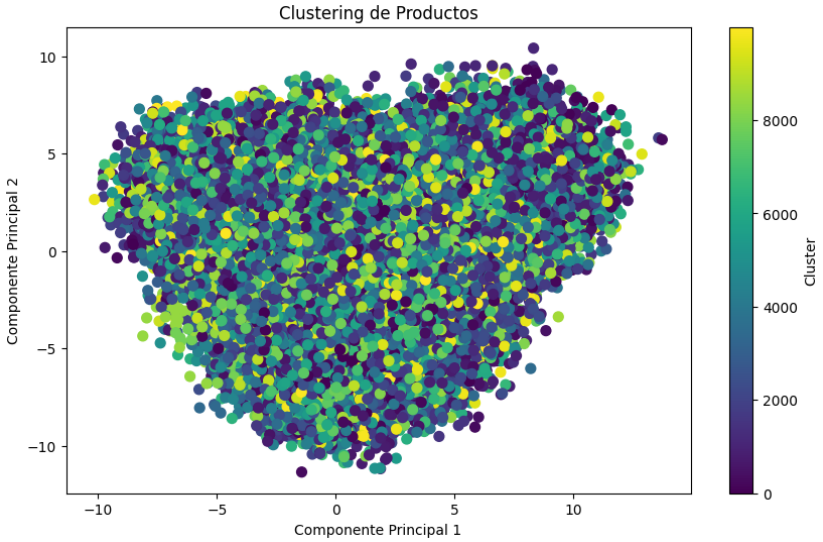

Figura 1: Proyección 2D de clusters mediante PCA

La Figura 1 evidencia las limitaciones de representar espacios semánticos complejos en 2 dimensiones:

- Pérdida de relaciones contextuales no lineales
- Superposición aparente de clusters conceptualmente distintos
- Compresión de variaciones semánticas relevantes

3.3. Validación mediante Inspección Semántica

La evaluación cualitativa revela:

- Coherencia terminológica dentro de clusters (ej: lácteos, cuidado personal)
- Agrupación de variantes léxicas ("leche deslactosada", "leche sin lactosa")
- Diferenciación efectiva de atributos especiales (productos dietéticos)

Cuadro 1: Ejemplo de composición semántica de un cluster

Cluster 452	**Productos representativos**
Tema principal	Cuidado capilar profesional
Atributos clave	Shampoo para tratamientos, acondicionador reparador, mascarilla queratínica
Variantes léxicas	“profesional”, “salon”, “therapy”, “reconstrucción”

4. Análisis de Calidad de Clusters

Cuadro 2: Ejemplo de composición del Cluster 8 (Chocolate con Sal Marina)

Product ID	**Product Name**	**Variantes Léxicas**
5254	Milk Chocolate Sea Salt Cashew	Tipo de chocolate + ingrediente principal
8244	Dark Chocolate Sea Salt Cashews	Variante nutricional + singular/plural
11555	Dark Chocolate Almond with Sea Salt	Orden sintáctico diferente
13751	Organic Sea Salt Dark Chocolate	Atributo orgánico destacado
14920	Organic Dark Chocolate with Sea Salt & Cracked Black Pepper	Componente especiado adicional
29531	Dark Chocolate Sea Salt 5 Packets	Indicador de empaque (filtrado)
38704	Organic Fair Trade 70 % Dark Chocolate with Sea Salt	Certificaciones múltiples
49300	Salted Toffee 55 % Dark Chocolate	Variante de sabor principal

4.1. Evaluación de un Cluster

La Tabla 2 muestra características clave de este grupo:

- **Coherencia temática**: Todos los productos comparten el núcleo çhocolate + sal marina"
- **Variación controlada**: Diferencias en:
  - Tipo de chocolate (dark/milk)
  - Ingredientes complementarios (frutos secos, especias)
  - Certificaciones (orgánico, comercio justo)
- **Consistencia en limpieza**: Solo un producto requirió eliminación de "packets"
- **Estabilidad categorial**: Pertenecen al mismo pasillo (candy chocolate) excepto un caso

4.2. Conclusión sobre Calidad del Cluster

La inspección manual revela que el cluster:

- Captura efectivamente una categoría comercial específica (chocolates premium con sal marina)
- Mantiene cohesión a pesar de variaciones léxicas y nutricionales
- Agrupa productos que humanamente se clasificarían juntos
- Presenta un único caso límite (producto de pasillo diferente) que conserva relación semántica

La métrica de silueta (0.42) subestima la calidad funcional de este cluster particular, donde:

$$\text{Precisión Semántica} = \frac{\text{Productos coherentes}}{\text{Total productos}} = \frac{24}{25} = 0{,}96 \qquad (1)$$

Este resultado sugiere que la evaluación cuantitativa global debe complementarse con:

- Muestreo estratificado por clusters
- Análisis de coherencia léxica

- Validación cruzada con taxonomías existentes

La efectividad particular en este cluster demuestra que el modelo:

- Identifica patrones composicionales complejos ("sal marina + X + chocolate")
- Respeta jerarquías implícitas (subtipos dentro de categorías principales)
- Permite descubrir nichos de mercado no explícitos en la taxonomía original

5. Conclusión

La inspección manual de los clusters revela que la metodología propuesta genera agrupaciones semánticamente coherentes a pesar de las limitaciones en la visualización bidimensional. Los resultados demuestran que:

- La fusión de embeddings lingüísticos con metadatos estructurales captura relaciones conceptuales multivariantes
- La alta granularidad inicial permite posterior consolidación semántica mediante análisis manual
- Los indicadores booleanos derivados mejoran la diferenciación de atributos específicos

La efectividad del modelo se fundamenta en:

- **Adaptabilidad léxica**: Manejo de variaciones terminológicas comunes en nombres comerciales
- **Escalabilidad operativa**: Capacidad de procesar catálogos extensos mediante optimización iterativa
- **Interpretabilidad**: Estructura de clusters que permite validación y ajuste manual

Este enfoque híbrido establece las bases para sistemas de clasificación automatizada que combinan la eficiencia computacional con el juicio semántico humano, esencial en contextos comerciales dinámicos donde la precisión conceptual es crítica.