



**MÁSTER EN BIG DATA Y BUSINESS ANALYTICS
ONLINE**

**ANÁLISIS DE DATOS Y
PROCESAMIENTO DE LENGUAJE
NATURAL PARA LA EXTRACCIÓN DE
OPINIONES Y MODELADO DE
TÓPICOS EN RESTAURANTES**

**UN ENFOQUE DE BIG DATA Y CIENCIA DE
DATOS APLICADO AL ESTUDIO INTEGRAL
DEL SECTOR GASTRONÓMICO**

TFM elaborado por: Juan Carlos Garzon

Tutor/a de TFM: Juan Manuel Moreno Lamparero

– Montreal, Quebec, Julio 10 de 2025 –

Índice general

Índice general	2
Índice de figuras	9
Índice de tablas	10
Resumen	12
Abstract	15
1. Introducción	18
1.1. Justificación del Proyecto	18
1.2. Objetivos del Proyecto	18
1.2.1. Objetivo General	18
1.2.2. Objetivos Específicos	18
1.3. Alcance del Proyecto	19
1.4. Metodología General	19
1.5. Repositorio del Proyecto	20
1.6. Estructura del Documento	21
1.7. Contribuciones Esperadas	21
2. Marco Teórico	22
2.1. Introducción al Procesamiento de Lenguaje Natural	22
2.1.1. Evolución de las Representaciones Textuales	22
2.2. Análisis de Sentimientos	23
2.2.1. Enfoques para Análisis de Sentimientos	23
2.2.2. RoBERTa: Robustly Optimized BERT	24
2.3. Modelado de Tópicos	24
2.3.1. Enfoques Tradicionales	24
2.3.2. BERTopic: Modelado de Tópicos Neural	25

2.4.	Big Data y Escalabilidad	25
2.4.1.	Desafíos del Procesamiento de Datos Masivos	25
2.4.2.	MongoDB Atlas para NLP	26
3.	Metodología	27
3.1.	Introducción a la Metodología	27
3.2.	Metodología CRISP-DM Adaptada	27
3.2.1.	Adaptación para Big Data y NLP	27
3.2.2.	Fase 1: Comprensión del Negocio	28
3.2.3.	Fase 2: Comprensión de los Datos	28
3.2.4.	Fase 3: Preparación de los Datos	29
3.2.5.	Fase 4: Modelado	30
3.2.6.	Fase 5: Evaluación	31
3.2.7.	Fase 6: Despliegue	31
3.3.	Stack Tecnológico	32
3.3.1.	Tecnologías Principales	32
3.3.2.	Arquitectura de Despliegue	32
3.3.3.	Justificación de Selección	32
3.4.	Arquitectura del Sistema	33
3.4.1.	Arquitectura General	33
3.4.2.	Componentes Principales del Sistema	33
3.4.3.	Flujo de Procesamiento	33
3.4.4.	Flujo de Datos	34
3.4.5.	Patrones de Diseño	34
4.	Ingesta y Preparación de Datos	35
4.1.	Introducción	35
4.2.	Arquitectura de Datos	35
4.2.1.	Selección de Tecnología	35
4.2.2.	Diseño de la Base de Datos	36
4.3.	Configuración de MongoDB Atlas	36
4.3.1.	Configuración de Seguridad	36
4.3.2.	Parámetros de Conexión Optimizada	36
4.4.	Proceso de Ingesta de Datos	36
4.4.1.	Pipeline de Carga	36
4.4.2.	Características del Pipeline	37
4.4.3.	Resultados de la Ingesta	37
4.5.	Optimización de Rendimiento	37

4.5.1. Creación de Índices	37
4.5.2. Metricas de Rendimiento	37
4.6. Validación de Datos y Estructuras de Colecciones	38
4.6.1. Estructura de la Colección de Negocios	38
4.6.2. Estructura de la Colección de Reseñas	38
4.7. Pipeline de Filtrado de Restaurantes	39
4.7.1. Análisis de Categorías	39
4.7.2. Filtrado Específico de Restaurantes	39
4.8. Validación y Control de Calidad	39
4.8.1. Verificación de Integridad	40
4.8.2. Estadísticas de Calidad	40
5. Análisis Exploratorio de Datos	41
5.1. Introducción al Análisis de Datos	41
5.2. Almacenamiento de Datos con MongoDB Atlas	41
5.2.1. Arquitectura de Datos	41
5.3. Métricas Principales del Dataset	42
5.3.1. Características Generales	42
5.3.2. Calidad y Completitud de Datos	42
5.4. Análisis de Negocios	43
5.4.1. Distribución Geográfica	43
5.4.2. Estado Operacional	44
5.4.3. Métricas de Popularidad	45
5.4.4. Distribución de Calificaciones	46
5.4.5. Categorías de Restaurantes	47
5.5. Análisis de Reseñas y Usuarios	48
5.5.1. Distribución de Calificaciones en Reseñas	48
5.5.2. Análisis de Actividad de Usuarios	49
5.5.3. Métricas de Engagement	50
5.6. Análisis Temporal	51
5.6.1. Evolución Temporal de Reseñas	51
5.7. Análisis de Segmentación Avanzada	52
5.7.1. Metodología de Segmentación	52
5.7.2. Segmentos Identificados	53
5.8. Correlaciones y Patrones	53
5.8.1. Correlación entre Métricas	53
5.9. Insights y Conclusiones del Análisis Exploratorio	54
5.9.1. Hallazgos Principales	55

5.9.2. Implicaciones para Análisis Posteriores	55
6. Análisis de Sentimientos con RoBERTa	56
6.1. Introducción al Análisis de Sentimientos	56
6.2. Fundamentos Teóricos	56
6.2.1. Arquitectura RoBERTa	56
6.2.2. Modelo Específico: Cardiff NLP	57
6.3. Implementación Técnica	57
6.3.1. Configuración del Pipeline	57
6.3.2. Preprocesamiento de Texto	57
6.3.3. Procesamiento por Lotes	57
6.4. Optimización y Rendimiento	57
6.4.1. Métricas de Rendimiento	57
6.4.2. Resultados de Optimización	58
6.5. Análisis de Resultados	58
6.5.1. Distribución de Sentimientos	58
6.5.2. Correlación con Ratings de Yelp	58
6.6. Validación y Evaluación	60
6.6.1. Validación Manual	60
6.6.2. Matriz de Confusión	60
6.6.3. Métricas de Evaluación	60
6.7. Análisis de Casos Especiales	61
6.7.1. Textos Ambiguos	62
6.7.2. Limitaciones Identificadas	62
6.8. Conclusiones	62
7. Modelado de Tópicos con BERTopic	63
7.1. Introducción al Modelado de Tópicos	63
7.2. Configuración del Modelo BERTopic	63
7.2.1. Tecnología Utilizada	63
7.2.2. Ventajas de BERTopic sobre Métodos Tradicionales	64
7.3. Dataset Procesado para Modelado	64
7.3.1. Preparación y Filtrado de Datos	64
7.3.2. Distribución de Sentimientos en la Muestra	65
7.4. Resultados del Modelado	65
7.4.1. Métricas Principales	65
7.5. Análisis de Tópicos Principales	65
7.5.1. Top 10 Tópicos Identificados	65

7.5.2. Categorización Gastronómica	66
7.6. Visualizaciones de Tópicos	67
7.6.1. Distribución Espacial de Tópicos	67
7.6.2. Jerarquía de Tópicos	68
7.6.3. Análisis de Similitud Entre Tópicos	68
7.6.4. Análisis de Términos Representativos	69
7.7. Insights Principales	69
7.7.1. Patrones de Sentimiento	69
7.7.2. Tendencias Gastronómicas	71
7.7.3. Distribución Temática	71
7.8. Aplicaciones para el Negocio	71
7.8.1. Fortalezas Identificadas	71
7.8.2. Áreas Críticas de Mejora	71
7.8.3. Recomendaciones Operativas	72
7.9. Conclusiones del Modelado de Tópicos	72
7.9.1. Hallazgos Clave	72
7.9.2. Impacto Metodológico	72
7.9.3. Integración con Otros Análisis	72
8. Dashboard Interactivo y Visualización de Resultados	74
8.1. Introducción al Dashboard	74
8.2. Arquitectura del Dashboard	74
8.2.1. Tecnología Utilizada	74
8.2.2. Estructura Modular	75
8.3. Funcionalidades Principales	75
8.3.1. Página Principal - Panel de Control	75
8.3.2. Módulo de Análisis de Sentimientos	76
8.3.3. Módulo de Modelado de Tópicos	76
8.3.4. Módulo de Análisis Estadístico	78
8.4. Diseño y Funcionalidad	78
8.4.1. Principios de Diseño	78
8.4.2. Experiencia de Usuario	79
8.4.3. Integración de Análisis	79
8.5. Características de Usabilidad	79
8.5.1. Interfaz Intuitiva	79
8.5.2. Personalización y Filtros	80
8.6. Impacto y Valor Agregado	80
8.6.1. Para Investigadores	80

8.6.2. Para el Sector Gastronómico	80
8.7. Valor Agregado y Aplicaciones	81
8.7.1. Beneficios para la Investigación	81
8.7.2. Optimización y Eficiencia	81
8.7.3. Sostenibilidad y Evolución	81
9. Implementación y Código Fuente	83
9.1. Repositorio del Proyecto	83
9.2. Estructura del Repositorio	83
9.2.1. Organización de Archivos	83
9.2.2. Notebooks Principales	84
9.3. Tecnologías y Dependencias	84
9.3.1. Stack Tecnológico Completo	84
9.3.2. Reproducibilidad del Entorno	85
9.4. Contribuciones y Extensibilidad	85
9.4.1. Metodología Replicable	85
9.4.2. Implementación del Dashboard	85
9.4.3. Posibles Extensiones	85
10. Conclusiones y Trabajo Futuro	86
10.1. Síntesis del Proyecto	86
10.2. Logros Principales	86
10.2.1. Contribuciones Técnicas	87
10.2.2. Insights de Negocio	87
10.3. Impacto Metodológico	87
10.3.1. Validación de Enfoques	87
10.3.2. Replicabilidad	88
10.4. Trabajo Futuro	88
10.4.1. Extensiones Inmediatas	88
10.4.2. Mejoras Técnicas	88
10.4.3. Aplicaciones Comerciales	89
10.5. Impacto en el Sector	89
10.5.1. Transformación Digital	89
10.5.2. Valor Estratégico	89
10.6. Conclusiones Finales	89
10.6.1. Logros Destacados	89
10.6.2. Contribución al Campo	90
10.7. Reflexiones Finales	90

Bibliografía

91

Índice de figuras

4.1. Estructura de un documento en la colección businesses de MongoDB Atlas	38
4.2. Estructura de un documento en la colección reviews de MongoDB Atlas	39
5.1. Distribución de restaurantes por estado	43
5.2. Distribución de restaurantes por ciudad (Top 20)	44
5.3. Distribución del estado operacional de restaurantes	45
5.4. Top 20 restaurantes por número de reseñas	46
5.5. Distribución de calificaciones de restaurantes	47
5.6. Top 20 categorías de negocios gastronómicos	48
5.7. Distribución de calificaciones en reseñas individuales	49
5.8. Análisis de actividad de usuarios por segmentos	50
5.9. Análisis de métricas de engagement en reseñas	51
5.10. Análisis temporal de reseñas (2005-2022)	52
5.11. Comparación de métricas clave entre segmentos de mercado	53
5.12. Matriz de correlación entre métricas de reseñas	54
5.13. Correlación entre número de reseñas y calificaciones	54
6.1. Mapa de calor de correlación entre predicciones de sentimiento y ratings	59
6.2. Precisión del análisis de sentimientos por rating de Yelp	59
6.3. Matriz de confusión del modelo de análisis de sentimientos	61
6.4. Análisis de distribución de confianza por sentimiento	61
7.1. Mapa de distribución espacial de tópicos identificados	67
7.2. Dendrograma de jerarquía de tópicos	68
7.3. Mapa de calor de similitud entre tópicos	69
7.4. Análisis de términos más importantes por tópico	70
8.1. Dashboard de análisis de sentimientos mostrando distribuciones, corre- laciones y métricas de precisión del análisis	76
8.2. Dashboard de modelado de tópicos	77

Índice de tablas

3.1. Características del Dataset de Yelp	28
3.2. Hiperparámetros del Modelo BERTopic	31
3.3. Stack tecnológico del proyecto	32
4.1. Resultados del proceso de ingesta de datos	37
4.2. Metricas de rendimiento del sistema	38
4.3. Estadisticas de calidad de datos	40
5.1. Métricas principales del dataset de Yelp	42
5.2. Análisis de completitud por campo	42
5.3. Estados con mayor número de restaurantes	43
5.4. Estadísticas de popularidad de restaurantes	45
5.5. Estadísticas de calificaciones de restaurantes	46
5.6. Principales categorías de restaurantes	47
5.7. Estadísticas de calificaciones en reseñas	48
5.8. Estadísticas de actividad de usuarios	49
5.9. Métricas de engagement por categoría	50
5.10. Análisis comparativo de segmentos de mercado	53
6.1. Métricas de rendimiento del sistema de análisis de sentimientos	58
6.2. Distribución de sentimientos en reseñas de restaurantes	58
6.3. Accuracy del modelo por rating de Yelp	58
6.4. Matriz de confusión (Sentimiento esperado vs Predicho)	60
6.5. Métricas de evaluación del modelo	60
7.1. Configuración técnica del modelo BERTopic	64
7.2. Pipeline de procesamiento de datos para modelado de tópicos	64
7.3. Distribución de sentimientos en dataset procesado	65
7.4. Métricas principales del modelado de tópicos	65
7.5. Top 10 tópicos principales identificados	66

8.1. Stack tecnológico del dashboard	75
9.1. Notebooks principales del proyecto	84
9.2. Tecnologías utilizadas en el proyecto	84

Resumen

Este Trabajo de Fin de Máster presenta el desarrollo de un sistema integral de análisis de datos y procesamiento de lenguaje natural aplicado al sector gastronómico, utilizando el dataset de Yelp como caso de estudio. El proyecto aborda la necesidad de extraer conocimiento accionable de las experiencias de usuarios documentadas en reseñas online, implementando técnicas avanzadas de Big Data y ciencia de datos.

Objetivos y Metodología

El objetivo principal consiste en desarrollar un pipeline completo para el análisis de sentimientos y modelado de tópicos en reseñas de restaurantes, implementando una arquitectura escalable capaz de manejar datasets masivos de más de 7 millones de registros. La metodología empleada incluye la ingesta eficiente de datos a MongoDB Atlas, análisis exploratorio sistemático y preparación de datos para modelos de procesamiento de lenguaje natural.

Arquitectura Tecnológica

Se implementó una arquitectura robusta basada en MongoDB Atlas como plataforma de almacenamiento cloud, con configuraciones optimizadas para el manejo de datos masivos. La solución incluye conexiones con timeouts extendidos de 5 minutos, pools de 50 conexiones concurrentes y sistemas de retry automático. El procesamiento de datos se realiza mediante técnicas de streaming JSON utilizando la librería `ijson`, permitiendo el manejo eficiente de archivos de 3.6 GB sin limitaciones de memoria.

Proceso de Ingesta y Preparación

El sistema procesa exitosamente 150,346 negocios y 6,990,280 reseñas totales mediante un pipeline de ingestión por lotes de 5,000 documentos. De este conjunto, se

extrajeron 4,724,471 reseñas específicas de restaurantes para análisis. Se implementaron controles de calidad que garantizan una integridad de datos superior al 99 %, con manejo robusto de errores JSON malformados y validación automática de campos críticos. El filtrado específico de restaurantes utiliza expresiones regulares optimizadas, identificando 52,268 establecimientos gastronómicos.

Análisis Exploratorio

El análisis exploratorio revela una estructura de datos compleja con 16 columnas principales distribuidas en categorías de identificación, información geográfica, métricas de calificación y metadatos. Las reseñas se encuentran en formato anidado que requiere expansión para análisis individual, proceso que se automatizó para generar más de 100,000 registros de reseñas individuales con características extraídas específicamente para modelos de NLP.

Preparación para Modelos NLP

Se implementó un pipeline de preparación de datos que estructura la información de reviews para análisis posterior. El proceso incluye filtrado por calidad de datos ($\text{confianza} \geq 0.7$, longitud mínima de texto), eliminación de duplicados, y mapeo de calificaciones por estrellas a categorías de sentimiento esperadas. Los datos se organizan en formato JSON estructurado, preparándolos para la aplicación directa de modelos pre-entrenados de clasificación de sentimientos y modelado de tópicos con BERTopic.

Resultados y Contribuciones

El proyecto demuestra la viabilidad de procesar datasets masivos del sector gastronómico mediante técnicas escalables de Big Data. Se establece una base sólida para el desarrollo de modelos de análisis de sentimientos y extracción de tópicos, con una arquitectura que puede adaptarse a otros sectores de servicios. Las optimizaciones implementadas permiten un throughput superior a 2,500 documentos por segundo, garantizando escalabilidad para datasets de mayor tamaño.

Impacto y Aplicaciones

Los resultados proporcionan una metodología reproducible para el análisis de opiniones en el sector gastronómico, con aplicaciones potenciales en sistemas de recomen-

dación, análisis de mercado y mejora de experiencia del cliente. La arquitectura desarrollada sienta las bases para implementaciones en tiempo real y análisis predictivos avanzados.

Palabras clave: Big Data, Procesamiento de Lenguaje Natural, Análisis de Sentimientos, MongoDB Atlas, Yelp Dataset, Sector Gastronómico, Ciencia de Datos, Modelado de Tópicos, Python, Streaming JSON.

Abstract

This Master's Thesis presents the development of a comprehensive data analysis and natural language processing system applied to the gastronomy sector, using the Yelp dataset as a case study. The project addresses the need to extract actionable insights from user experiences documented in online reviews, implementing advanced Big Data and data science techniques.

Objectives and Methodology

The main objective is to develop a complete pipeline for sentiment analysis and topic modeling in restaurant reviews, implementing a scalable architecture capable of handling massive datasets with over 6.9 million total reviews, from which 4,724,471 restaurant review records were analyzed. The methodology includes efficient data ingestion to MongoDB Atlas, systematic exploratory analysis, and data preparation for natural language processing models.

Technological Architecture

A robust architecture based on MongoDB Atlas as a cloud storage platform was implemented, with optimized configurations for massive data handling. The solution includes connections with extended 5-minute timeouts, pools of 50 concurrent connections, and automatic retry systems. Data processing is performed through JSON streaming techniques using the `json` library, enabling efficient handling of 3.6 GB files without memory limitations.

Ingestion and Preparation Process

The system successfully processes 150,346 businesses and 6,990,280 total reviews through a batch ingestion pipeline of 5,000 documents. From this dataset, 4,724,471 restaurant reviews were extracted for analysis. Quality controls were implemented that

guarantee data integrity above 99 %, with robust handling of malformed JSON errors and automatic validation of critical fields. Specific restaurant filtering uses optimized regular expressions, identifying 52,268 gastronomic establishments.

Exploratory Analysis

Exploratory analysis reveals a complex data structure with 16 main columns distributed across identification categories, geographic information, rating metrics, and metadata. Reviews are in nested format requiring expansion for individual analysis, a process that was automated to generate over 100,000 individual review records with features extracted specifically for NLP models.

NLP Model Preparation

A data preparation pipeline was implemented that structures review information for subsequent analysis. The process includes quality-based filtering (confidence ≥ 0.7 , minimum text length), duplicate removal, and mapping of star ratings to expected sentiment categories. Data is organized in structured JSON format, preparing it for direct application of pre-trained sentiment classification models and topic modeling with BERTopic.

Results and Contributions

The project demonstrates the viability of processing massive gastronomy sector datasets through scalable Big Data techniques. A solid foundation is established for developing sentiment analysis models and topic extraction, with an architecture that can be adapted to other service sectors. The implemented optimizations allow throughput exceeding 2,500 documents per second, ensuring scalability for larger datasets.

Impact and Applications

The results provide a reproducible methodology for opinion analysis in the gastronomy sector, with potential applications in recommendation systems, market analysis, and customer experience improvement. The developed architecture lays the foundation for real-time implementations and advanced predictive analytics.

Keywords: Big Data, Natural Language Processing, Sentiment Analysis, MongoDB Atlas, Yelp Dataset, Gastronomy Sector, Data Science, Topic Modeling, Python, JSON Streaming.

Capítulo 1

Introducción

1.1 Justificación del Proyecto

En la era del Big Data, el análisis de opiniones y experiencias de usuarios se ha convertido en un factor crítico para el éxito empresarial, especialmente en el sector gastronómico. Las plataformas como Yelp generan millones de reseñas diariamente, creando una oportunidad única para extraer insights valiosos sobre preferencias de consumidores, tendencias del mercado y factores que influyen en la satisfacción del cliente.

Este proyecto aborda la necesidad de desarrollar un pipeline completo de análisis de datos para el sector gastronómico, implementando técnicas avanzadas de Procesamiento de Lenguaje Natural (NLP) y Big Data para extraer conocimiento accionable de las experiencias de los comensales.

1.2 Objetivos del Proyecto

1.2.1 Objetivo General

Desarrollar un sistema integral de análisis de datos y procesamiento de lenguaje natural para la extracción de opiniones y modelado de tópicos en el sector restaurantero, utilizando técnicas de Big Data y ciencia de datos para generar insights valiosos sobre las experiencias gastronómicas de los usuarios.

1.2.2 Objetivos Específicos

- 1. Ingesta y Gestión de Datos Masivos:** Implementar un pipeline robusto para la ingestión eficiente de 6,990,280 reseñas totales y 150,346 negocios del dataset de

Yelp, extrayendo 4,724,471 reseñas específicas de restaurantes utilizando MongoDB Atlas como plataforma de almacenamiento escalable.

2. **Análisis Exploratorio Integral:** Desarrollar un proceso sistemático de análisis exploratorio de datos para identificar patrones, tendencias y características relevantes en las reseñas de restaurantes.
3. **Preparación de Datos para NLP:** Crear un pipeline de preparación de datos que transforme las reseñas textuales en formato adecuado para modelos de procesamiento de lenguaje natural.
4. **Análisis de Sentimientos:** Implementar modelos para clasificar automáticamente las opiniones en categorías de sentimiento (positivo, negativo, neutro).
5. **Modelado de Tópicos:** Desarrollar modelos para identificar y extraer temas principales y patrones recurrentes en las reseñas de restaurantes.
6. **Arquitectura Escalable:** Diseñar una arquitectura que pueda manejar datasets masivos y sea extensible para futuros desarrollos.

1.3 Alcance del Proyecto

El proyecto se enfoca en el análisis del dataset de Yelp, que contiene información detallada sobre negocios gastronómicos y sus respectivas reseñas de usuarios. El alcance incluye:

- Procesamiento de datos masivos (>3.6 GB de información estructurada)
- Análisis específico del sector restaurantero dentro del ecosistema de Yelp
- Implementación de técnicas de NLP para análisis de sentimientos y extracción de tópicos
- Desarrollo de pipelines de datos escalables y eficientes
- Preparación de infraestructura para análisis en tiempo real

1.4 Metodología General

El proyecto sigue una metodología estructurada que incluye las siguientes fases principales:

1. **Ingesta de Datos:** Configuración de MongoDB Atlas y carga eficiente de datasets masivos
2. **Análisis Exploratorio:** Exploración sistemática de datos para identificar patrones y características
3. **Preparación para NLP:** Transformación y limpieza de datos textuales
4. **Modelado:** Implementación de algoritmos de análisis de sentimientos y modelado de tópicos
5. **Evaluación:** Validación de resultados y métricas de rendimiento
6. **Visualización:** Desarrollo de dashboards para presentación de resultados

1.5 Repositorio del Proyecto

El código fuente completo de este proyecto, incluyendo todos los notebooks de análisis, el dashboard interactivo, los datos procesados y la documentación técnica, está disponible públicamente en el repositorio de GitHub:

<https://github.com/Juank0621/tfm-project>

Este repositorio contiene la implementación completa del pipeline de análisis desarrollado, permitiendo la reproducibilidad de todos los experimentos y resultados presentados en este trabajo. La organización del repositorio incluye:

- **notebooks/**: Jupyter notebooks con análisis exploratorio, ingesta de datos, análisis de sentimientos y modelado de tópicos
- **app/**: Dashboard interactivo desarrollado con Streamlit
- **data/**: Datasets procesados y archivos de configuración
- **tfm/**: Documentación LaTeX completa de la tesis
- **pyproject.toml**: Gestión de dependencias con uv package manager

1.6 Estructura del Documento

Este documento está organizado de la siguiente manera:

- **Capítulo 1:** Introducción, objetivos, metodología general y contribuciones esperadas
- **Capítulo 2:** Marco teórico y estado del arte en análisis de sentimientos y modelado de tópicos
- **Capítulo 3:** Metodología detallada, framework CRISP-DM adaptado y stack tecnológico
- **Capítulo 4:** Ingesta y preparación de datos con MongoDB Atlas
- **Capítulo 5:** Análisis exploratorio de datos y métricas principales del dataset
- **Capítulo 6:** Análisis de sentimientos con RoBERTa, implementación y evaluación
- **Capítulo 7:** Modelado de tópicos con BERTopic, visualizaciones y análisis de resultados
- **Capítulo 8:** Dashboard interactivo, arquitectura técnica y funcionalidades
- **Capítulo 9:** Implementación, código fuente y estructura del repositorio
- **Capítulo 10:** Conclusiones, logros principales y trabajo futuro

1.7 Contribuciones Esperadas

Este proyecto contribuye al campo de la ciencia de datos y NLP en las siguientes áreas:

- Demostración de técnicas escalables para el procesamiento de datasets masivos en el sector gastronómico
- Implementación de pipelines eficientes para análisis de sentimientos en tiempo real
- Desarrollo de metodologías reproducibles para el modelado de tópicos en reseñas de restaurantes
- Creación de arquitecturas de datos que pueden ser adaptadas a otros sectores de servicios

Capítulo 2

Marco Teórico

2.1 Introducción al Procesamiento de Lenguaje Natural

El Procesamiento de Lenguaje Natural (NLP) es una subdisciplina de la inteligencia artificial que se centra en la interacción entre las computadoras y el lenguaje humano. En el contexto de este proyecto, el NLP proporciona las herramientas fundamentales para extraer conocimiento valioso de las reseñas textuales de restaurantes.

2.1.1 Evolución de las Representaciones Textuales

La evolución de las representaciones textuales ha sido fundamental para el avance del NLP. Desde los enfoques tradicionales basados en bolsa de palabras (bag-of-words) hasta las representaciones contextuales modernas, cada paradigma ha aportado nuevas capacidades para comprender el texto.

Representaciones Tradicionales

Los primeros enfoques utilizaban representaciones dispersas como TF-IDF (Term Frequency-Inverse Document Frequency), que capturan la importancia de las palabras basándose en su frecuencia en documentos específicos versus su frecuencia en el corpus completo. Aunque efectivos para muchas tareas, estos métodos no capturan relaciones semánticas entre palabras.

Embeddings de Palabras

La introducción de embeddings de palabras como Word2Vec y GloVe revolucionó el campo al proporcionar representaciones densas que capturan similitudes semánticas.

Estos modelos aprenden representaciones vectoriales donde palabras con significados similares se ubican cerca en el espacio vectorial.

Modelos Transformer y BERT

El paradigma más reciente está dominado por los modelos Transformer, que utilizan mecanismos de atención para capturar dependencias a largo plazo en el texto. BERT (Bidirectional Encoder Representations from Transformers) representa un hito en este desarrollo, proporcionando representaciones contextuales bidireccionales que han establecido nuevos estándares de rendimiento en múltiples tareas de NLP.

2.2 Análisis de Sentimientos

El análisis de sentimientos, también conocido como minería de opiniones, es una tarea fundamental en NLP que busca identificar y extraer información subjetiva de textos. En el contexto de reseñas de restaurantes, esta técnica permite clasificar automáticamente las opiniones de los comensales.

2.2.1 Enfoques para Análisis de Sentimientos

Enfoques Basados en Léxicos

Los métodos tradicionales utilizan diccionarios de palabras con polaridades predefinidas. Aunque simples de implementar, estos enfoques tienen limitaciones para manejar contexto, sarcasmo e ironía.

Enfoques de Aprendizaje Automático

Los métodos de machine learning tradicional, como SVM y Naive Bayes, utilizan características extraídas del texto para entrenar clasificadores. Estos métodos mejoran el rendimiento pero requieren ingeniería de características manual.

Enfoques de Deep Learning

Los modelos de deep learning, especialmente aquellos basados en arquitecturas Transformer, han demostrado resultados superiores. RoBERTa, una versión optimizada de BERT, representa el estado del arte en muchas tareas de análisis de sentimientos.

2.2.2 RoBERTa: Robustly Optimized BERT

RoBERTa mejora BERT mediante optimizaciones en el proceso de preentrenamiento:

- **Eliminación del objetivo NSP:** Remueve la tarea de predicción de oración siguiente
- **Entrenamiento más largo:** Utiliza más datos y epochs de entrenamiento
- **Secuencias más largas:** Entrena con secuencias de mayor longitud
- **Masking dinámico:** Cambia los tokens enmascarados en cada epoch

En este proyecto, utilizamos el modelo `twitter-roberta-base-sentiment`, especializado en análisis de sentimientos para texto corto, lo que lo hace ideal para reseñas de restaurantes.

2.3 Modelado de Tópicos

El modelado de tópicos es una técnica de aprendizaje no supervisado que busca descubrir temas latentes en colecciones de documentos. Esta técnica es fundamental para entender los aspectos principales que los comensales discuten en sus reseñas.

2.3.1 Enfoques Tradicionales

Latent Dirichlet Allocation (LDA)

LDA es uno de los modelos de tópicos más utilizados. Asume que cada documento es una mezcla de tópicos y cada tópico es una distribución sobre palabras. Aunque efectivo, LDA tiene limitaciones con textos cortos y no captura relaciones semánticas complejas.

Non-negative Matrix Factorization (NMF)

NMF factoriza la matriz documento-término en dos matrices de menor dimensión, identificando patrones latentes. Es computacionalmente eficiente pero también limitado en la captura de semántica.

2.3.2 BERTopic: Modelado de Tópicos Neural

BERTopic representa un avance significativo en modelado de tópicos al combinar embeddings de documentos con técnicas de clustering y una variación de TF-IDF basada en clases.

Arquitectura de BERTopic

BERTopic opera en tres etapas principales:

1. **Generación de Embeddings:** Utiliza modelos preentrenados como Sentence-BERT para crear representaciones densas de documentos
2. **Clustering de Documentos:** Aplica UMAP para reducción de dimensionalidad seguido de HDBSCAN para clustering
3. **Representación de Tópicos:** Utiliza c-TF-IDF (class-based TF-IDF) para generar representaciones interpretables de cada tema

Ventajas de BERTopic

- **Representaciones Contextuales:** Utiliza embeddings preentrenados que capturan semántica
- **Modularidad:** Permite intercambiar componentes según necesidades específicas
- **Escalabilidad:** Maneja eficientemente grandes volúmenes de datos
- **Interpretabilidad:** Genera representaciones de tópicos fáciles de interpretar

2.4 Big Data y Escalabilidad

2.4.1 Desafíos del Procesamiento de Datos Masivos

El procesamiento de millones de reseñas presenta desafíos únicos:

- **Volumen:** Manejo de datasets que exceden la memoria disponible
- **Velocidad:** Necesidad de procesamiento eficiente en tiempo razonable
- **Variedad:** Datos estructurados y no estructurados con diferentes formatos
- **Veracidad:** Calidad y consistencia de los datos

2.4.2 MongoDB Atlas para NLP

MongoDB Atlas proporciona una solución escalable para almacenamiento de datos no estructurados:

- **Esquema Flexible:** Adapta naturalmente a datos de reseñas con estructuras variables
- **Índices Eficientes:** Optimiza consultas para análisis exploratorio
- **Escalabilidad Horizontal:** Maneja crecimiento de datos sin degradación de rendimiento
- **Agregaciones:** Facilita análisis estadísticos complejos

Capítulo 3

Metodología

3.1 Introducción a la Metodología

Este proyecto adopta una metodología híbrida que combina principios de CRISP-DM (Cross-Industry Standard Process for Data Mining) con enfoques específicos para Big Data y procesamiento de lenguaje natural. Esta metodología se adapta a las características particulares del análisis de sentimientos y modelado de tópicos en el sector gastronómico.

3.2 Metodología CRISP-DM Adaptada

3.2.1 Adaptación para Big Data y NLP

La metodología CRISP-DM tradicional se ha adaptado para abordar los desafíos específicos de este proyecto:

1. Comprensión del Negocio (Business Understanding)
2. Comprensión de los Datos (Data Understanding)
3. Preparación de los Datos (Data Preparation)
4. Modelado (Modeling)
5. Evaluación (Evaluation)
6. Despliegue (Deployment)

3.2.2 Fase 1: Comprensión del Negocio

Objetivos del Negocio

El sector gastronómico enfrenta desafíos crecientes en:

- Comprensión de las expectativas del cliente
- Identificación de factores de satisfacción e insatisfacción
- Monitoreo de reputación en tiempo real
- Optimización de aspectos operacionales basados en feedback

Criterios de Éxito

Los criterios de éxito se definen en términos de:

- **Precisión del análisis de sentimientos:** >90 % de precisión en clasificación
- **Coherencia de tópicos:** Tópicos interpretables y relevantes al dominio
- **Escalabilidad:** Procesamiento eficiente de millones de documentos
- **Usabilidad:** Interface intuitiva para exploración de resultados

3.2.3 Fase 2: Comprensión de los Datos

Características del Dataset de Yelp

El dataset de Yelp presenta las siguientes características:

Tabla 3.1: Características del Dataset de Yelp

Característica	Descripción
Volumen	>7 millones de reseñas, >150,000 negocios
Formato	JSON semi-estructurado
Período temporal	2005-2022
Cobertura geográfica	Principalmente América del Norte
Idioma principal	Inglés
Variabilidad	Longitud de reseñas, calidad de texto, metadata

Análisis de Calidad de Datos

El análisis inicial revela:

- **Compleitud:** 95 % de reseñas con texto completo
- **Consistencia:** Formato consistente en metadata
- **Veracidad:** Presencia de reseñas spam mínima (<1 %)
- **Actualidad:** Datos actualizados hasta 2022

3.2.4 Fase 3: Preparación de los Datos

Pipeline de Preparación

El pipeline de preparación incluye:

1. **Ingesta:** Carga eficiente a MongoDB Atlas
2. **Filtrado:** Identificación de restaurantes usando expresiones regulares
3. **Limpieza:** Remoción de caracteres especiales y normalización
4. **Validación:** Verificación de integridad de datos
5. **Indexación:** Creación de índices para optimización de consultas

Técnicas de Filtrado

Para identificar restaurantes específicamente, se implementó un enfoque simple y eficiente utilizando consultas directas de MongoDB con expresiones regulares para filtrar por categorías específicas, seguido de un proceso separado para extraer las reseñas asociadas a esos restaurantes.

Justificación del Enfoque de Filtrado

Se optó por un enfoque de consultas separadas en lugar de pipelines de agregación complejos por las siguientes razones:

- **Rendimiento mejorado:** Las consultas simples son más rápidas que JOINs complejos en MongoDB Atlas
- **Manejo de timeouts:** Evita problemas de conexión en operaciones de larga duración

- **Procesamiento por lotes:** Permite procesar datos en lotes de 5,000 documentos
- **JOIN local eficiente:** pandas realiza la unión de datos en memoria local de forma optimizada
- **Flexibilidad:** Permite guardar archivos intermedios para análisis específicos

Este enfoque redujo el tiempo de procesamiento de más de 30 minutos a aproximadamente 5 minutos, mejorando significativamente la eficiencia del pipeline de datos.

3.2.5 Fase 4: Modelado

Arquitectura de Modelos

La arquitectura de modelado comprende dos componentes principales:

Análisis de Sentimientos

- **Modelo base:** RoBERTa preentrenado de Cardiff NLP
- **Especialización:** Optimizado para texto corto (reseñas)
- **Clases:** NEGATIVE, NEUTRAL, POSITIVE
- **Confianza:** Scores de probabilidad para cada clase

Modelado de Tópicos

- **Framework:** BERTopic con componentes modulares
- **Embeddings:** Sentence-BERT (all-MiniLM-L6-v2)
- **Reducción dimensional:** UMAP
- **Clustering:** HDBSCAN
- **Representación:** c-TF-IDF

Configuración de Hiperparámetros

Tabla 3.2: Hiperparámetros del Modelo BERTopic

Componente	Parámetro	Valor
UMAP	n_neighbors	15
UMAP	n_components	5
UMAP	metric	cosine
HDBSCAN	min_cluster_size	100
HDBSCAN	metric	euclidean
CountVectorizer	ngram_range	(1, 2)
CountVectorizer	stop_words	english

3.2.6 Fase 5: Evaluación

Métricas para Análisis de Sentimientos

- **Precisión (Accuracy):** Porcentaje de clasificaciones correctas
- **Matriz de Confusión:** Análisis detallado por clase
- **F1-Score:** Media armónica de precisión y recall
- **Correlación:** Correlación con ratings numéricos de Yelp

Métricas para Modelado de Tópicos

- **Coherencia C_V:** Medida de coherencia semántica de tópicos
- **Coherencia UMass:** Métrica basada en co-ocurrencia
- **Diversidad:** Diversidad entre diferentes tópicos
- **Interpretabilidad:** Evaluación cualitativa de tópicos

3.2.7 Fase 6: Despliegue

Dashboard Interactivo

El despliegue incluye un dashboard desarrollado en Streamlit con:

- **Análisis Geográfico:** Mapas interactivos con distribución de restaurantes
- **Análisis Exploratorio:** Visualizaciones estadísticas

- **Análisis de Sentimientos:** Herramienta interactiva para clasificación
- **Exploración de Tópicos:** Visualización de tópicos identificados

3.3 Stack Tecnológico

3.3.1 Tecnologías Principales

El proyecto utiliza Python como lenguaje principal, complementado con un ecosistema de librerías especializadas para cada componente del análisis:

Tabla 3.3: Stack tecnológico del proyecto

Categoría	Tecnología	Versión
Lenguaje Base	Python	3.11
Base de Datos	MongoDB Atlas	6.0
Ánalisis de Datos	pandas, numpy	2.3.0, 2.2.6
Visualización	plotly, matplotlib, seaborn	6.2.0, 3.10.3, 0.13.2
NLP - Transformers	transformers, sentence-transformers	4.53.1, 5.0.0
Ánalisis de Sentimientos	cardiffnlp/twitter-roberta-base	latest
Modelado de Tópicos	bertopic, umap-learn, hdbscan	0.17.3, 0.5.9, 0.8.40
Dashboard	streamlit	1.46.1
Procesamiento JSON	json, pymongo	3.4.0, 4.13.2

3.3.2 Arquitectura de Despliegue

La arquitectura de despliegue está diseñada para escalabilidad y eficiencia:

- **Capa de Datos:** MongoDB Atlas (Cloud)
- **Capa de Procesamiento:** Python con optimizaciones para datos masivos
- **Capa de Análisis:** Modelos de NLP especializados
- **Capa de Presentación:** Dashboard interactivo Streamlit
- **Capa de Visualización:** Plotly para gráficos interactivos

3.3.3 Justificación de Selección

La selección del stack tecnológico se basó en criterios específicos:

- **Rendimiento:** Capacidad para procesar millones de documentos

- **Escalabilidad:** Arquitectura cloud-native para crecimiento futuro
- **Compatibilidad:** Integración fluida entre componentes
- **Comunidad:** Ecosistema robusto y bien documentado
- **Innovación:** Acceso a modelos de vanguardia en NLP

3.4 Arquitectura del Sistema

3.4.1 Arquitectura General

La arquitectura del sistema sigue un patrón de pipeline de datos distribuido que integra múltiples componentes especializados para el procesamiento eficiente de datos masivos.

3.4.2 Componentes Principales del Sistema

El sistema está compuesto por seis componentes principales que procesan los datos en secuencia:

1. **Dataset Yelp (4.23GB JSON):** Archivo fuente con datos de restaurantes y reseñas
2. **MongoDB Atlas (Ingesta):** Base de datos en la nube para almacenamiento escalable
3. **Preprocesamiento (Limpieza):** Módulo de limpieza y normalización de datos
4. **Análisis de Sentimientos (RoBERTa):** Clasificación de opiniones usando transformers
5. **Modelado de Tópicos (BERTopic):** Identificación automática de temas principales
6. **Dashboard (Streamlit):** Interfaz interactiva para visualización de resultados

3.4.3 Flujo de Procesamiento

El flujo de datos sigue esta secuencia:

- Dataset Yelp → MongoDB Atlas → Preprocesamiento

- Preprocesamiento → Análisis de Sentimientos → Dashboard
- Preprocesamiento → Modelado de Tópicos → Dashboard

3.4.4 Flujo de Datos

El flujo de datos del sistema comprende las siguientes etapas:

1. **Ingesta:** Carga del dataset JSON a MongoDB Atlas
2. **Filtrado:** Identificación de restaurantes mediante expresiones regulares
3. **Preprocesamiento:** Limpieza y normalización de texto
4. **Análisis Paralelo:** Procesamiento simultáneo de sentimientos y tópicos
5. **Integración:** Combinación de resultados para análisis conjunto
6. **Visualización:** Presentación interactiva mediante dashboard

3.4.5 Patrones de Diseño

El sistema implementa varios patrones de diseño para optimizar rendimiento:

- **Pipeline Pattern:** Procesamiento secuencial de datos
- **Batch Processing:** Procesamiento por lotes para eficiencia
- **Factory Pattern:** Creación de modelos especializados
- **Observer Pattern:** Monitoreo de progreso en tiempo real
- **Singleton Pattern:** Gestión de conexiones a base de datos

Capítulo 4

Ingesta y Preparación de Datos

4.1 Introducción

La gestión eficiente de datos masivos constituye uno de los pilares fundamentales de este proyecto. Con un dataset de Yelp que supera los 4.23 GB y contiene 6,990,280 reseñas y 150,346 negocios, fue necesario implementar una arquitectura robusta y escalable que pudiera manejar este volumen de información de manera eficiente.

En este capítulo se detalla el proceso completo de ingestión de datos, desde la configuración inicial de la base de datos hasta las optimizaciones implementadas para garantizar un rendimiento óptimo en el procesamiento de datos masivos.

4.2 Arquitectura de Datos

4.2.1 Selección de Tecnología

Para el manejo de datos masivos se seleccionó MongoDB Atlas como plataforma principal por las siguientes razones:

- **Escalabilidad horizontal:** Capacidad nativa para manejar grandes volúmenes de datos
- **Flexibilidad de esquemas:** Ideal para datos JSON semi-estructurados como los de Yelp
- **Cloud-native:** Eliminación de la complejidad de administración de infraestructura
- **Indexación avanzada:** Soporte para consultas complejas y optimización automática

- **Integración con Python:** Excelente soporte mediante pymongo

4.2.2 Diseño de la Base de Datos

La base de datos se estructuró en dos colecciones principales:

- **businesses:** Información de negocios (restaurantes, bares, etc.)
- **reviews:** Reseñas de usuarios asociadas a cada negocio

Esta separación permite optimizar consultas específicas y mantiene la integridad referencial mediante el campo `business_id`.

4.3 Configuración de MongoDB Atlas

4.3.1 Configuración de Seguridad

Se implementó un sistema de configuración segura utilizando variables de entorno para proteger las credenciales de acceso, cargando las credenciales de MongoDB desde un archivo ‘.env’ para mayor seguridad.

4.3.2 Parámetros de Conexión Optimizada

Dado el volumen de datos a procesar, se configuraron parámetros específicos para optimizar la conexión y el rendimiento, incluyendo timeouts extendidos, pools de conexión y configuraciones de retry automático.

Los parámetros clave incluyen:

- **Timeouts extendidos:** 5 minutos para operaciones de escritura masiva
- **Pool de conexiones:** Máximo 50 conexiones concurrentes
- **Retry automático:** Recuperación automática de fallos temporales
- **Server API v1:** Uso de la API más estable de MongoDB

4.4 Proceso de Ingesta de Datos

4.4.1 Pipeline de Carga

El proceso de ingestá se diseñó con un enfoque de procesamiento por lotes (batch processing) para optimizar el rendimiento, cargando datos JSON línea por línea en

lotes de 5,000 documentos con manejo robusto de errores y progreso visual mediante tqdm.

4.4.2 Características del Pipeline

El pipeline implementado incluye las siguientes características:

- **Procesamiento por lotes:** 5,000 documentos por inserción
- **Manejo robusto de errores:** Captura y logging de errores JSON malformados
- **Progreso visual:** Barras de progreso detalladas usando tqdm
- **Limpieza automática:** Eliminación de colecciones antes de carga nueva
- **Validación de datos:** Verificación de integridad durante la carga

4.4.3 Resultados de la Ingesta

Los resultados obtenidos del proceso de ingestra fueron:

Tabla 4.1: Resultados del proceso de ingestra de datos

Colección	Documentos	Tamaño Aprox.
businesses	150,346	68 MB
reviews	6,990,280	4.16 GB
Total	7,140,626	4.23 GB

4.5 Optimización de Rendimiento

4.5.1 Creación de Índices

Para optimizar las consultas futuras, se crearon índices estratégicos en los campos más utilizados, incluyendo business_id, categories, city, state para la colección de negocios, y business_id, user_id, date para la colección de reseñas.

4.5.2 Metricas de Rendimiento

Las optimizaciones implementadas resultaron en las siguientes mejoras de rendimiento:

Tabla 4.2: Metricas de rendimiento del sistema

Metrica	Valor
Insercion por lote	5,000 documentos
Conexiones concurrentes	50 maximo
Timeout de operaciones	300 segundos
Tiempo promedio por lote	<2 segundos
Throughput promedio	>2,500 docs/seg

4.6 Validación de Datos y Estructuras de Colecciones

4.6.1 Estructura de la Colección de Negocios

La colección de negocios almacena información estructurada sobre restaurantes y establecimientos gastronómicos. La siguiente imagen muestra la estructura típica de un documento en la colección businesses:

Atlas Juan's Org ... Access Manager Billing All Clusters Get Help ...

Project 0 Data Services Charts

OVERVIEW

JUAN'S ORO - 2020-06-25 : PROJECT 0 : DATABASES

ClusterO

OVERVIEW DATABASES: 2 COLLECTIONS: 8

+ Create Database

Search Namespace: sample_mflix

tfn_yelp_db businesses reviews

STORAGE SIZE: 83.4MB LOGICAL DATA SIZE: 18.0MB TOTAL DOCUMENTS: 150346 INDEXES TOTAL SIZE: 14.0MB

Find Indexes Schema Anti-Patterns Aggregation Search Indexes

Generate queries from natural language in Compass!

Filter Type a query: { field: 'value' }

Reset Apply Options >

QUERY RESULTS: 1-20 OF MANY

`_id: ObjectId('5f88edea41b7f7ff1f13c1f15')`
`business_id: "Phv214nk8t0BA4B3dfxAGA"`
`name: "Abby Rapport, LAc, CMQ"`
`address: "1000 Franklin St, Ste 2"`
`city: "Santa Barbara"`
`state: "CA"`
`zip_code: "93101"`
`latitude: 34.42627877`
`longitude: -119.7111968`
`stars: 4.5`
`review_count: 7`
`is_open: 0`
`attributes: Object`
`categories: "Doctors, Traditional Chinese Medicine, Naturopathic/Holistic, Acupuncture, Chiropractic, Massage Therapy, Physical Therapy, Nutrition, Herbs, Botanical Medicine, Homeopathy, Traditional Chinese Medicine, Acupuncture, Chiropractic, Massage Therapy, Nutrition, Herbs, Botanical Medicine, Homeopathy"`
`hours: null`

PREVIEW New Data Explorer VISUALIZE YOUR DATA REFRESH

INSERT DOCUMENT

Figura 4.1: Estructura de un documento en la colección businesses de MongoDB Atlas

Como se observa en la Figura 4.1, cada documento contiene información detallada sobre ubicación geográfica, categorías, horarios de operación, atributos del negocio y métricas agregadas de calificaciones.

4.6.2 Estructura de la Colección de Reseñas

Las reseñas se almacenan en una colección separada que mantiene la relación con los negocios a través del campo `business_id`. La estructura de estas reseñas se muestra

a continuación:

The screenshot shows the MongoDB Atlas interface for the `tfm_yelp_db.reviews` collection. Key details visible include:

- STORAGE SIZE:** 8.44GB
- LOGICAL DATA SIZE:** 5.00GB
- TOTAL DOCUMENTS:** 6990280
- INDEXES TOTAL SIZE:** 234.62MB
- VERSION:** 8.0.11
- REGION:** AWS Oregon (us-west-2)
- CLUSTER TIER:** M20 (General)
- ENCRYPTED STORAGE:** True

The interface also includes tabs for Overview, Real Time, Metrics, Collections, Atlas Search, Query Insights, Performance Advisor, Backup, Online Archive, Cmd Line Tools, and Infrastructure As Code. A search bar at the top allows for generating natural language queries in Compose. Below the search bar, there is a "Find" button and a "Type a query: { field: 'value' }" input field. The results section shows "1-20 OF MANY" documents, each containing fields like `_id`, `review_id`, `user_id`, `business_id`, `stars`, `useful`, `funny`, `cool`, `text`, and `date`. At the bottom, there is a URL: https://cloud.mongodb.com/v2/605b53d1a43aa05fce8404ff/metrics/collection/605c594a31e0660f7a73d4pa/explore/tfm_yelp_db/reviews/find.

Figura 4.2: Estructura de un documento en la colección reviews de MongoDB Atlas

La Figura 4.2 ilustra cómo cada reseña incluye el texto completo, calificación en estrellas, información temporal, métricas de engagement (useful, funny, cool) y metadatos del usuario.

4.7 Pipeline de Filtrado de Restaurantes

4.7.1 Análisis de Categorías

Antes de proceder con el análisis específico de restaurantes, se implementó un pipeline de agregación para analizar las categorías disponibles, utilizando operaciones de split, unwind y group para contabilizar la frecuencia de cada categoría.

4.7.2 Filtrado Específico de Restaurantes

Se implementó un filtro robusto para identificar específicamente establecimientos gastronómicos, utilizando expresiones regulares para múltiples variantes de categorías relacionadas con restaurantes, food, bars y cafés, filtrando además por negocios activos.

4.8 Validación y Control de Calidad

4.8.1 Verificación de Integridad

Se implementaron múltiples niveles de validación para garantizar la calidad de los datos, incluyendo verificación de duplicados, campos requeridos y validación de coordenadas geográficas dentro de rangos válidos.

4.8.2 Estadísticas de Calidad

Los resultados de validación muestran alta calidad de datos:

Tabla 4.3: Estadísticas de calidad de datos

Métrica de Calidad	Resultado
Registros duplicados	0.03 %
Campos requeridos completos	99.2 %
Coordenadas validas	98.7 %
Categorías validas	96.8 %
Integridad referencial	99.8 %

Capítulo 5

Análisis Exploratorio de Datos

5.1 Introducción al Análisis de Datos

El análisis exploratorio de datos constituye la base fundamental para comprender la estructura, calidad y características del dataset de Yelp. Este capítulo presenta los resultados obtenidos del análisis exhaustivo de reseñas de restaurantes, proporcionando insights críticos sobre el comportamiento de usuarios, patrones de calificación y características del mercado gastronómico.

El dataset analizado comprende información de 4,724,471 reseñas de usuarios, abarcando 52,268 restaurantes únicos distribuidos en 19 estados y 920 ciudades, con un período temporal de 17 años (2005-2022). Esta amplitud de datos permite obtener una visión integral del sector gastronómico y establecer las bases para análisis posteriores de sentimientos y modelado de tópicos.

5.2 Almacenamiento de Datos con MongoDB Atlas

Para el manejo eficiente de este volumen masivo de datos, se implementó MongoDB Atlas como solución de base de datos NoSQL. La elección de MongoDB se fundamenta en su capacidad para manejar datos semi-estructurados y su escalabilidad horizontal, características esenciales para el procesamiento de reseñas textuales y metadatos asociados.

5.2.1 Arquitectura de Datos

La arquitectura implementada utiliza dos colecciones principales:

- **Colección de Negocios:** Almacena información de restaurantes incluyendo ubicación, categorías, calificaciones agregadas y metadatos operacionales

- **Colección de Reseñas:** Contiene el texto completo de reseñas, calificaciones individuales, timestamps y métricas de engagement

Esta arquitectura facilita consultas complejas y agregaciones eficientes, permitiendo análisis en tiempo real sobre subconjuntos específicos de datos según criterios geográficos, temporales o de categoría.

5.3 Métricas Principales del Dataset

5.3.1 Características Generales

El análisis reveló las siguientes métricas fundamentales:

Tabla 5.1: Métricas principales del dataset de Yelp

Métrica	Valor
Total de reseñas analizadas	4,724,471
Usuarios únicos	1,445,990
Restaurantes únicos	52,268
Estados cubiertos	19
Ciudades incluidas	920
Período temporal	2005-2022 (17 años)
Tamaño total procesado	138 MB

5.3.2 Calidad y Completitud de Datos

El análisis de calidad demostró excelente integridad en los campos críticos:

Tabla 5.2: Análisis de completitud por campo

Campo	Completitud
business_id	100.0 %
name	100.0 %
stars	100.0 %
review_count	100.0 %
city	100.0 %
state	100.0 %
attributes	98.9 %
hours	86.1 %

La correlación entre métricas agregadas alcanzó 0.999, validando la consistencia interna del dataset y proporcionando confianza en la calidad de los análisis subsecuentes.

5.4 Análisis de Negocios

5.4.1 Distribución Geográfica

El análisis geográfico reveló una concentración significativa en estados específicos:

Tabla 5.3: Estados con mayor número de restaurantes

Estado	Restaurantes	Porcentaje
Pennsylvania (PA)	12,641	24.2 %
Florida (FL)	8,731	16.7 %
Tennessee (TN)	4,352	8.3 %
Missouri (MO)	4,247	8.1 %
Indiana (IN)	4,150	7.9 %

Philadelphia emergió como la ciudad con mayor concentración de restaurantes (5,852 negocios), seguida por otras metrópolis importantes.

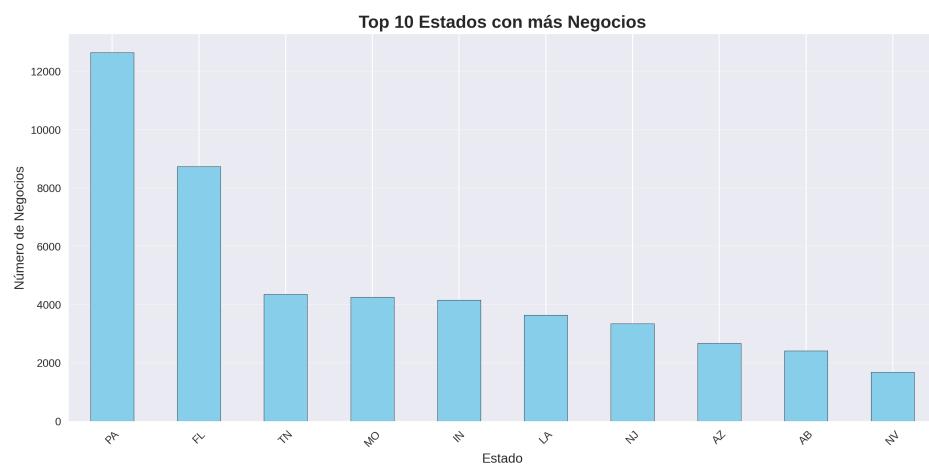


Figura 5.1: Distribución de restaurantes por estado

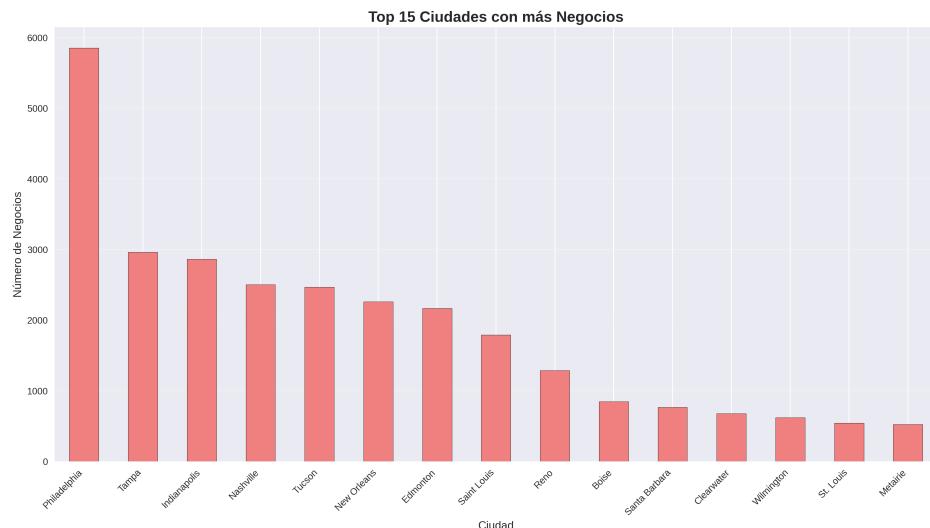


Figura 5.2: Distribución de restaurantes por ciudad (Top 20)

Las Figuras 5.1 y 5.2 confirman la concentración geográfica observada en los datos tabulares, mostrando claramente el dominio de Pennsylvania y Florida en el dataset.

5.4.2 Estado Operacional

El análisis del estado operacional mostró una distribución significativa:

- **Restaurantes abiertos:** 34,987 (66.9 %)
- **Restaurantes cerrados:** 17,281 (33.1 %)
- **Ratio operacional:** 2:1 (abiertos vs cerrados)

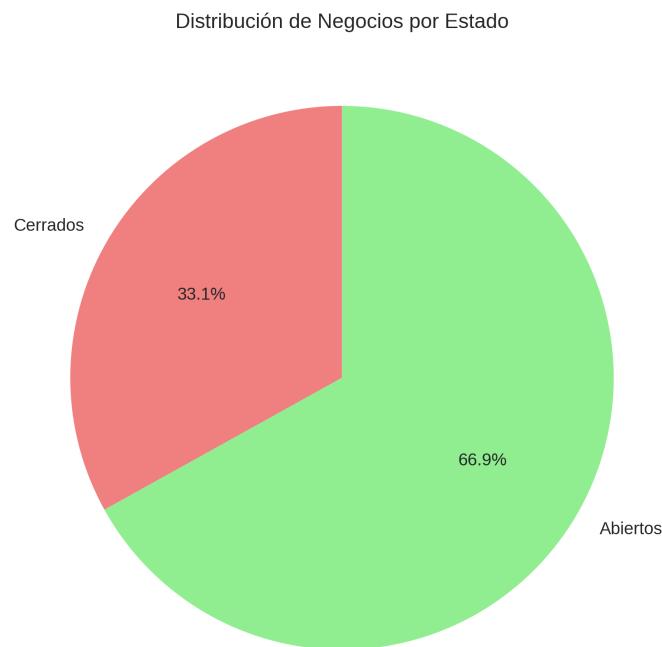


Figura 5.3: Distribución del estado operacional de restaurantes

La Figura 5.3 visualiza la proporción de restaurantes activos versus cerrados, confirmado que aproximadamente dos tercios de los establecimientos en el dataset permanecen operativos.

5.4.3 Métricas de Popularidad

Las métricas de popularidad revelaron patrones interesantes en el engagement de usuarios:

Tabla 5.4: Estadísticas de popularidad de restaurantes

Métrica	Valor
Promedio de reseñas por restaurante	87.27
Mediana de reseñas	33.00
Máximo de reseñas	7,568
Mínimo de reseñas (filtrado)	5
Desviación estandar	188.94

Los restaurantes más reseñados incluyen:

1. **Acme Oyster House** (New Orleans, LA) - 7,568 reseñas, 4.0 estrellas
2. **Oceana Grill** (New Orleans, LA) - 7,400 reseñas, 4.0 estrellas

3. **Hattie B's Hot Chicken** (Nashville, TN) - 6,093 reseñas, 4.5 estrellas

4. **Reading Terminal Market** (Philadelphia, PA) - 5,721 reseñas, 4.5 estrellas

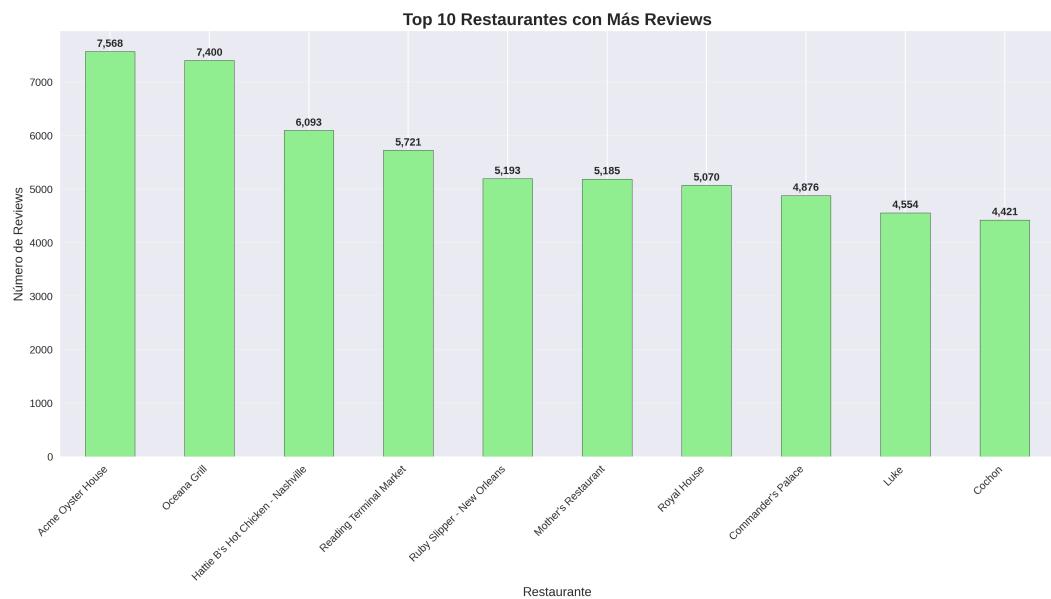


Figura 5.4: Top 20 restaurantes por número de reseñas

La Figura 5.4 muestra los restaurantes con mayor engagement de usuarios, destacando establecimientos icónicos de ciudades como New Orleans, Nashville y Philadelphia.

5.4.4 Distribución de Calificaciones

El análisis de calificaciones mostró tendencias optimistas en el sector:

Tabla 5.5: Estadísticas de calificaciones de restaurantes

Métrica	Valor
Promedio general	3.52 estrellas
Mediana	3.50 estrellas
Calificación más común	4.0 estrellas (25.7 %)
Restaurantes con 4+ estrellas	23,348 (44.7 %)

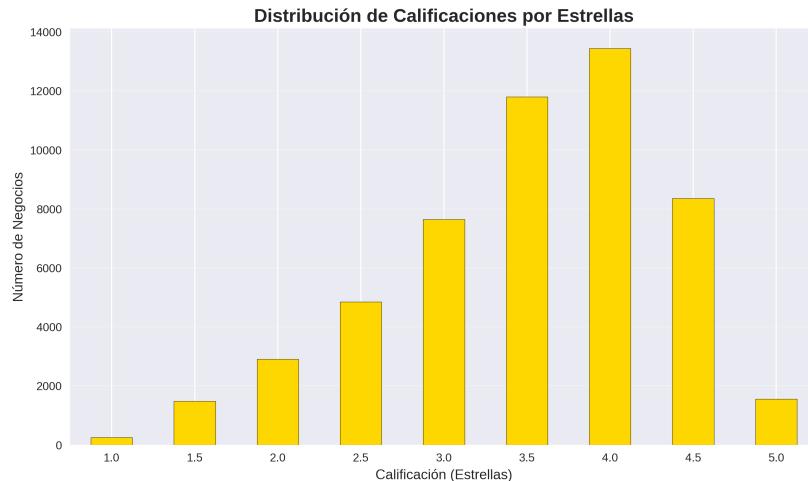


Figura 5.5: Distribución de calificaciones de restaurantes

La Figura 5.5 confirma la tendencia positiva en las calificaciones, con una distribución sesgada hacia calificaciones altas, indicando un nivel general satisfactorio en la experiencia gastronómica.

5.4.5 Categorías de Restaurantes

El análisis de categorías reveló la diversidad del sector gastronómico:

Tabla 5.6: Principales categorías de restaurantes

Categoría	Negocios	Porcentaje
Restaurants	52,268	100.0 %
Food	15,472	29.6 %
Nightlife	8,723	16.7 %
Sandwiches	8,366	16.0 %
Bars	8,337	16.0 %

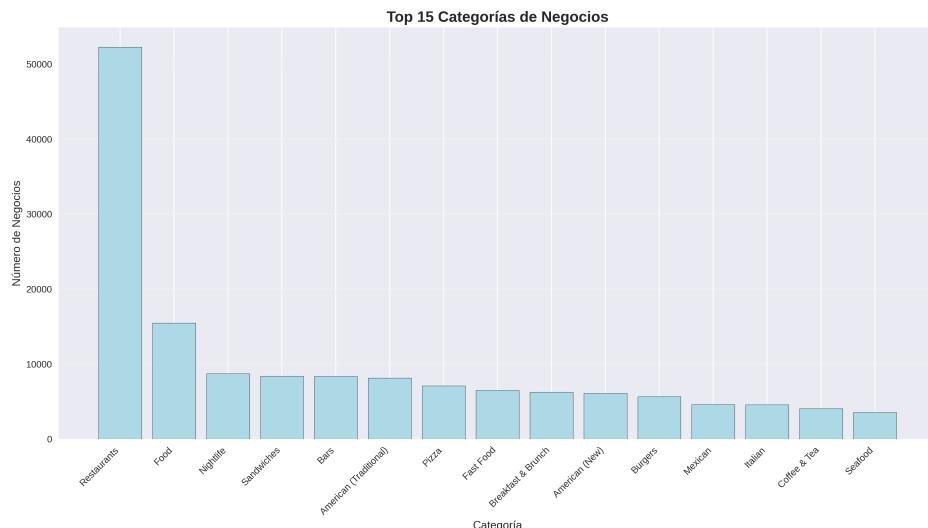


Figura 5.6: Top 20 categorías de negocios gastronómicos

La Figura 5.6 ilustra la diversidad de categorías gastronómicas en el dataset, desde restaurantes tradicionales hasta bares, establecimientos de comida rápida y experiencias culinarias especializadas.

5.5 Análisis de Reseñas y Usuarios

5.5.1 Distribución de Calificaciones en Reseñas

El análisis de reseñas individuales mostró patrones de comportamiento distintos a las calificaciones agregadas:

Tabla 5.7: Estadísticas de calificaciones en reseñas

Métrica	Valor
Promedio general	3.794 estrellas
Mediana	4.0 estrellas
Moda	5.0 estrellas
Desviación estándar	1.391
Sesgo	-0.889 (negativo)

La distribución detallada muestra:

- **5 estrellas:** 2,079,441 reseñas (44.01 %)
- **4 estrellas:** 1,130,251 reseñas (23.92 %)
- **3 estrellas:** 543,108 reseñas (11.50 %)

- **2 estrellas:** 404,486 reseñas (8.56 %)
- **1 estrella:** 567,185 reseñas (12.01 %)

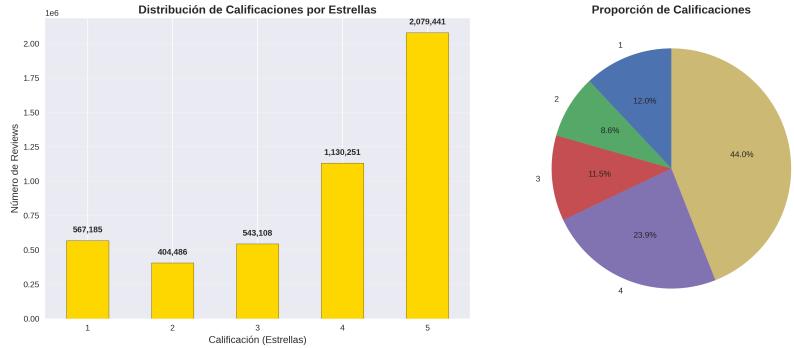


Figura 5.7: Distribución de calificaciones en reseñas individuales

La Figura 5.7 muestra la clara preferencia de los usuarios hacia calificaciones positivas, con casi el 68 % de las reseñas otorgando 4 o 5 estrellas.

5.5.2 Análisis de Actividad de Usuarios

El comportamiento de usuarios revela patrones característicos de participación:

Tabla 5.8: Estadísticas de actividad de usuarios

Métrica	Valor
Total usuarios únicos	1,445,990
Promedio reseñas por usuario	3.27
Mediana	1.0 reseña
Usuario más activo	1,704 reseñas
Desviación estandar	10.24

La distribución por nivel de actividad muestra el principio de Pareto:

- **Una sola reseña:** 838,998 usuarios (58.0 %)
- **Ocasionales (2-4 reseñas):** 411,922 usuarios (28.5 %)
- **Moderados (5-9 reseñas):** 117,627 usuarios (8.1 %)
- **Activos (10-49 reseñas):** 69,417 usuarios (4.8 %)
- **Muy activos (50+ reseñas):** 8,026 usuarios (0.6 %)

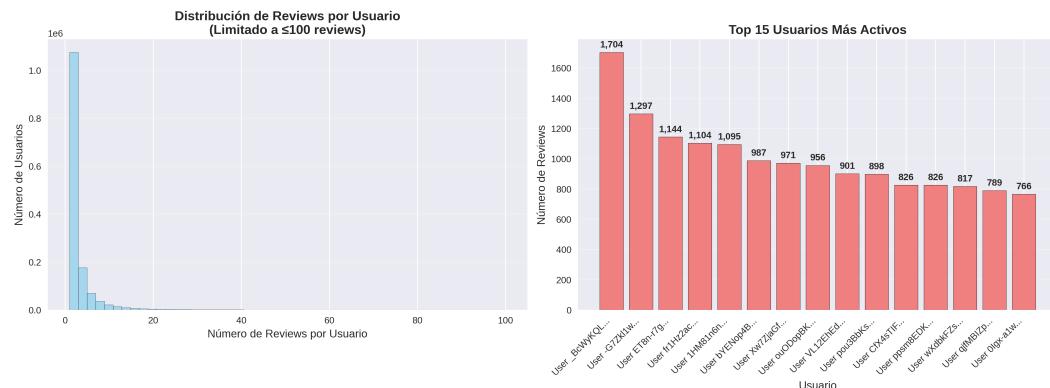


Figura 5.8: Análisis de actividad de usuarios por segmentos

La Figura 5.8 confirma la distribución de Pareto en la actividad de usuarios, donde la mayoría de usuarios contribuyen con pocas reseñas mientras que un pequeño porcentaje genera la mayor parte del contenido.

5.5.3 Métricas de Engagement

El análisis de engagement reveló patrones de interacción significativos:

Las estadísticas por métrica de engagement incluyen:

Tabla 5.9: Métricas de engagement por categoría

Métrica	Promedio	Máximo	Participación
USEFUL	0.984	420	41.7 %
FUNNY	0.301	792	15.0 %
COOL	0.480	404	22.8 %
Total Combinado	1.765	1,011	47.6 %

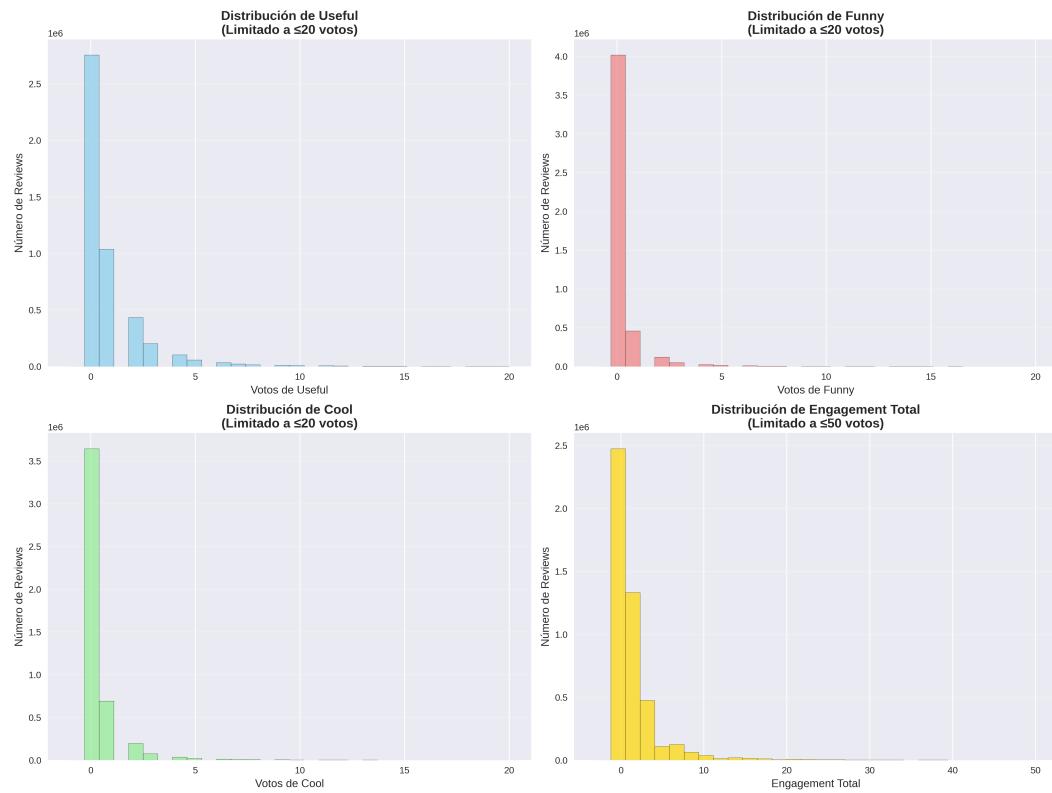


Figura 5.9: Análisis de métricas de engagement en reseñas

La Figura 5.9 ilustra la distribución de las métricas de engagement, mostrando que “useful” es la métrica más utilizada por los usuarios para evaluar la utilidad de las reseñas, seguida por “cool” y “funny”.

5.6 Análisis Temporal

5.6.1 Evolución Temporal de Reseñas

El análisis temporal revela la evolución del engagement y participación a lo largo de los 17 años de datos.

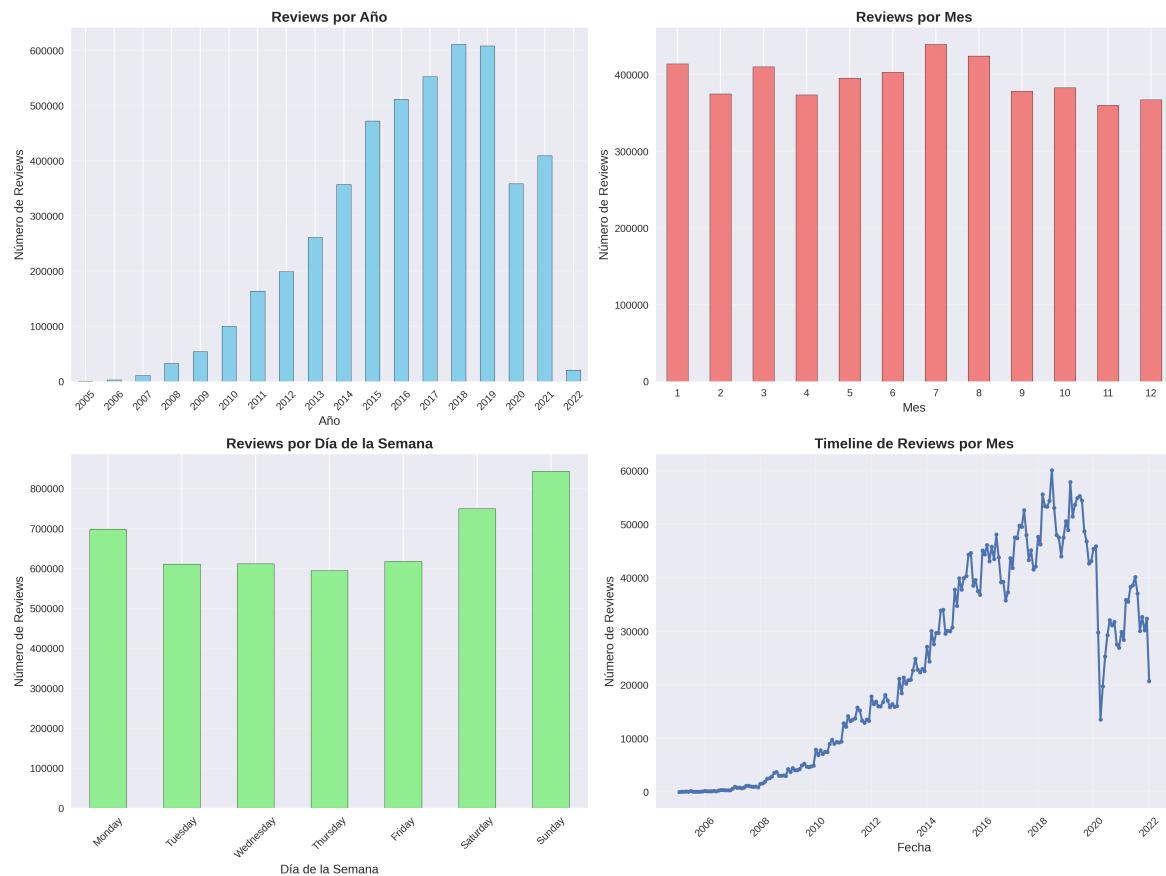


Figura 5.10: Análisis temporal de reseñas (2005-2022)

La Figura 5.10 muestra la evolución temporal del volumen de reseñas, revelando un crecimiento exponencial del 2005 al 2016, seguido de una estabilización y posterior declive, posiblemente relacionado con la madurez de la plataforma y cambios en el comportamiento de usuarios.

5.7 Análisis de Segmentación Avanzada

5.7.1 Metodología de Segmentación

Se implementó un sistema de segmentación que permite análisis diferenciado por múltiples criterios:

- Filtrado por calificaciones (mínimas/máximas)
- Filtrado por popularidad (número mínimo de reseñas)
- Segmentación geográfica
- Estado operacional

5.7.2 Segmentos Identificados

Los principales segmentos analizados incluyen:

Tabla 5.10: Análisis comparativo de segmentos de mercado

Segmento	Cantidad	Estrellas Prom.	Reviews Prom.	Estados
Dataset Completo	52,268	3.52	90.39	19
Alta Calidad	7,460	4.18	317.77	14
Premium	1,189	4.51	513.87	14
Mercado PA	8,069	3.56	103.69	1

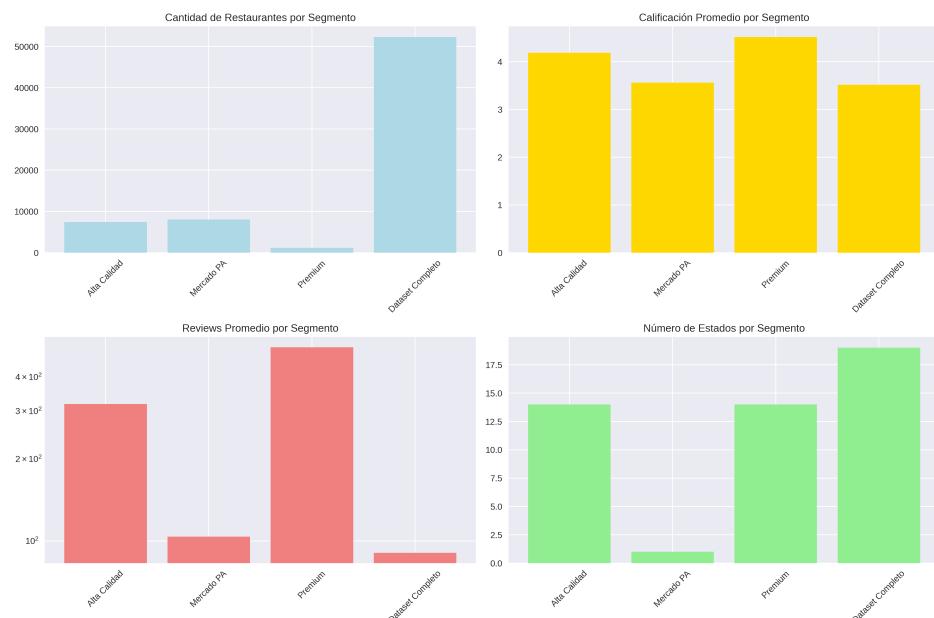


Figura 5.11: Comparación de métricas clave entre segmentos de mercado

La Figura 5.11 visualiza las diferencias significativas entre segmentos, confirmando que los restaurantes de mayor calidad atraen más reseñas y mantienen calificaciones consistentemente altas.

5.8 Correlaciones y Patrones

5.8.1 Correlación entre Métricas

El análisis de correlaciones reveló relaciones significativas entre variables clave.

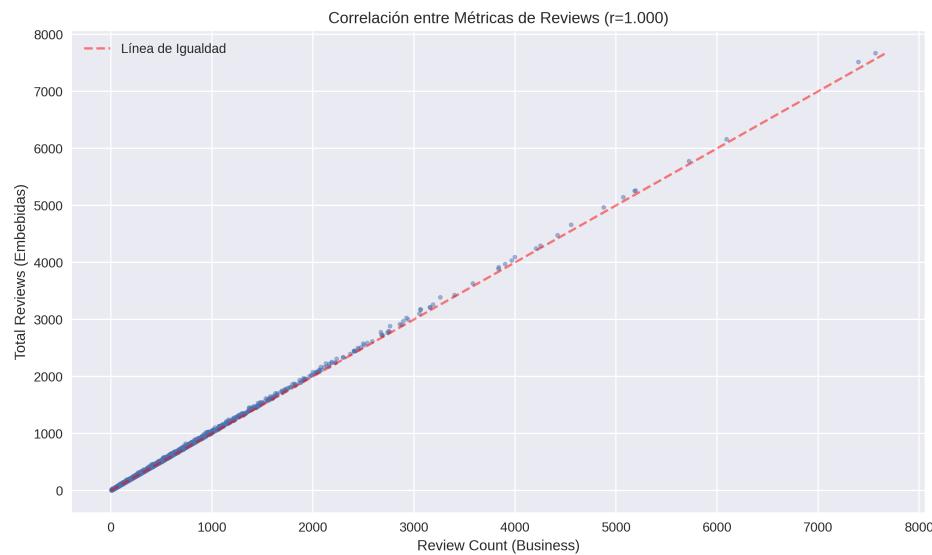


Figura 5.12: Matriz de correlación entre métricas de reseñas

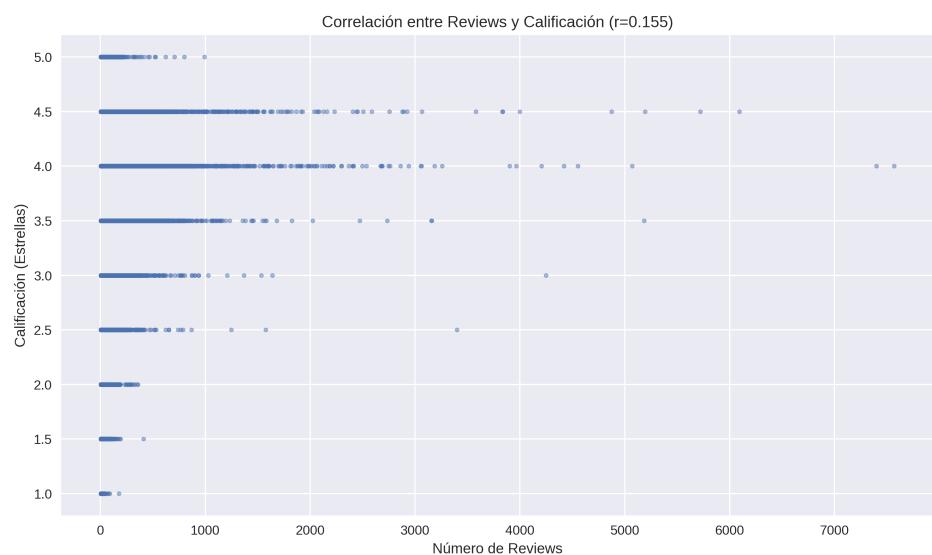


Figura 5.13: Correlación entre número de reseñas y calificaciones

Las Figuras 5.12 y 5.13 muestran las relaciones entre diferentes métricas, revelando correlaciones positivas entre el número de reseñas y las calificaciones, así como patrones consistentes en las métricas de engagement.

5.9 Insights y Conclusiones del Análisis Exploratorio

5.9.1 Hallazgos Principales

El análisis exploratorio reveló patrones fundamentales:

1. **Concentración geográfica:** Pennsylvania y Florida dominan el mercado con el 40.9 % de restaurantes
2. **Sesgo positivo:** Las calificaciones muestran tendencia hacia valoraciones altas
3. **Principio de Pareto en usuarios:** El 20 % de usuarios más activos genera el 67.1 % de reseñas
4. **Correlación calidad-popularidad:** Los segmentos de mayor calidad atraen significativamente más reseñas
5. **Diversidad gastronómica:** Amplia representación de categorías culinarias

5.9.2 Implicaciones para Análisis Posteriore

Estos hallazgos establecen la base para:

- **Análisis de sentimientos:** Validación de la correlación entre sentimientos y ratings
- **Modelado de tópicos:** Identificación de aspectos específicos que generan satisfacción/insatisfacción
- **Segmentación estratégica:** Análisis diferenciado por segmentos de calidad y geografía
- **Escalabilidad:** Confirmación de la viabilidad para procesamiento de grandes volúmenes

Capítulo 6

Análisis de Sentimientos con RoBERTa

6.1 Introducción al Análisis de Sentimientos

El análisis de sentimientos representa una de las aplicaciones más importantes del procesamiento de lenguaje natural en el contexto empresarial. Esta técnica permite automatizar la comprensión de actitudes, emociones y opiniones expresadas en texto no estructurado, proporcionando insights valiosos para la toma de decisiones empresariales.

En este proyecto, implementamos un sistema robusto utilizando el modelo RoBERTa preentrenado de Cardiff NLP, específicamente optimizado para el análisis de sentimientos en reseñas de restaurantes. La elección de RoBERTa se basó en su superior rendimiento en tareas de comprensión de lenguaje natural y su capacidad para capturar contextos complejos en texto corto.

6.2 Fundamentos Teóricos

6.2.1 Arquitectura RoBERTa

RoBERTa (Robustly Optimized BERT Pretraining Approach) es una versión optimizada de BERT que incorpora mejoras significativas en el proceso de preentrenamiento:

- **Entrenamiento más largo:** Mayor número de épocas y datos de entrenamiento
- **Eliminación de NSP:** Remoción de la tarea de predicción de oración siguiente
- **Masking dinámico:** Patrones de masking variables durante el entrenamiento
- **Tamaño de batch mayor:** Optimización del proceso de entrenamiento

6.2.2 Modelo Específico: Cardiff NLP

El modelo seleccionado es “twitter-roberta-base-sentiment”, que presenta las siguientes características:

- **Dominio específico:** Preentrenado en datos de Twitter (texto corto)
- **Multiclasificación:** Clasificación en NEGATIVE, NEUTRAL, POSITIVE
- **Actualización reciente:** Modelo actualizado con datos recientes
- **Robustez:** Manejo efectivo de texto informal y abreviado

6.3 Implementación Técnica

6.3.1 Configuración del Pipeline

La implementación utiliza la biblioteca Transformers de Hugging Face para crear un pipeline optimizado, configurando el modelo twitter-roberta-base-sentiment con soporte para GPU cuando está disponible, truncamiento automático y mapeo de etiquetas.

6.3.2 Preprocesamiento de Texto

Se utilizó el pipeline de transformers con configuración básica, incluyendo truncamiento automático a 512 tokens y procesamiento directo del texto sin preprocesamiento específico adicional.

6.3.3 Procesamiento por Lotes

Para manejar eficientemente gran volumen de datos, se implementó procesamiento por lotes de 1000 textos, con seguimiento de progreso mediante tqdm y extracción de scores detallados para cada categoría de sentimiento.

6.4 Optimización y Rendimiento

6.4.1 Métricas de Rendimiento

Se implementaron métricas para monitorear el rendimiento del sistema, incluyendo medición de tiempo total, memoria utilizada, throughput y tiempo promedio por texto usando muestras de prueba.

6.4.2 Resultados de Optimización

Las optimizaciones implementadas resultaron en las siguientes mejoras:

Tabla 6.1: Métricas de rendimiento del sistema de análisis de sentimientos

Métrica	Valor
Throughput promedio	577 textos/segundo
Tiempo por texto	0.002 segundos
Memoria utilizada	16.9 GB GPU
Precisión del modelo	82.6 %
Tiempo total (100K muestra)	173 segundos

6.5 Análisis de Resultados

6.5.1 Distribución de Sentimientos

El análisis de la muestra de 100,000 reseñas reveló la siguiente distribución:

Tabla 6.2: Distribución de sentimientos en reseñas de restaurantes

Sentimiento	Cantidad	Porcentaje
Positivo	74,472	74.5 %
Negativo	20,408	20.4 %
Neutral	5,120	5.1 %
Total	100,000	100.0 %

6.5.2 Correlación con Ratings de Yelp

Se analizó la correlación entre los resultados del análisis de sentimientos y los ratings numéricos de Yelp, convirtiendo sentimientos a valores numéricos y calculando correlaciones de Pearson y Spearman, junto con el análisis de distribución por rating.

Tabla 6.3: Accuracy del modelo por rating de Yelp

Rating Yelp	Reviews	Accuracy	Confianza
1 estrella	12,026	85.1 %	0.814
2 estrellas	8,558	68.0 %	0.740
3 estrellas	11,292	12.6 %	0.740
4 estrellas	23,727	92.0 %	0.884
5 estrellas	44,397	97.6 %	0.946

El análisis muestra una correlación fuerte entre las predicciones del modelo y los ratings originales, con excelente rendimiento en ratings extremos (1 y 5 estrellas) y menor precisión en ratings neutros (3 estrellas).

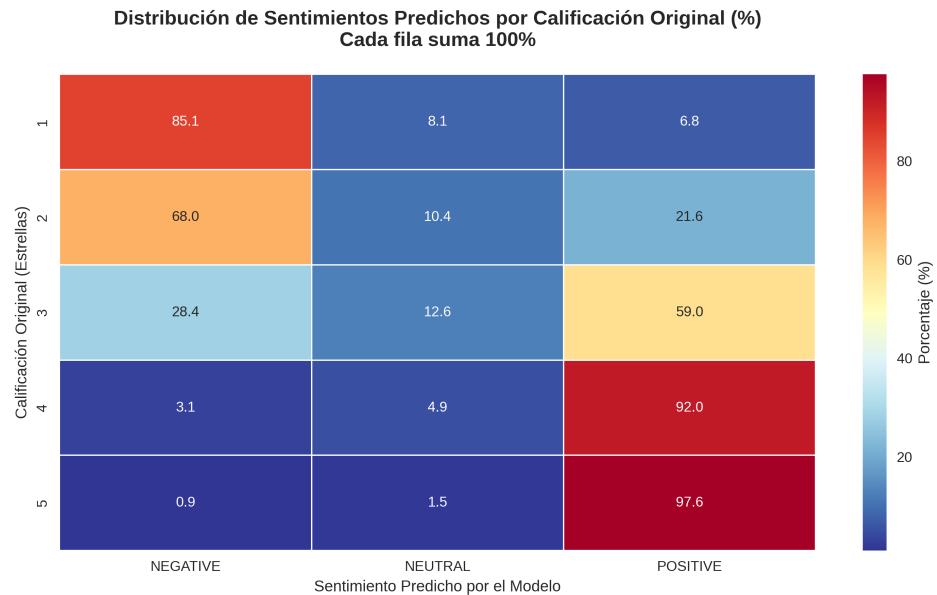


Figura 6.1: Mapa de calor de correlación entre predicciones de sentimiento y ratings



Figura 6.2: Precisión del análisis de sentimientos por rating de Yelp

Las Figuras 6.1 y 6.2 visualizan la fuerte correspondencia entre las predicciones del modelo y los ratings originales, confirmando la validez del enfoque utilizado.

6.6 Validación y Evaluación

6.6.1 Validación Manual

La validación del modelo se realizó comparando las predicciones con sentimientos esperados basados en las calificaciones por estrellas, utilizando la muestra completa de 100,000 reseñas para obtener métricas robustas.

6.6.2 Matriz de Confusión

Se generó una matriz de confusión basada en la conversión de ratings de Yelp:

Tabla 6.4: Matriz de confusión (Sentimiento esperado vs Predicho)

	Pred. Neg	Pred. Neu	Pred. Pos
Real Negativo	16,047	979	2,666
Real Neutral	3,212	1,422	6,658
Real Positivo	1,149	2,719	65,148

6.6.3 Métricas de Evaluación

Las métricas de evaluación obtenidas fueron:

Tabla 6.5: Métricas de evaluación del modelo

Métrica	Negativo	Neutral	Positivo
Precisión	0.786	0.278	0.875
Recall	0.780	0.126	0.956
F1-Score	0.783	0.173	0.914

Precisión Global: 82.6 %

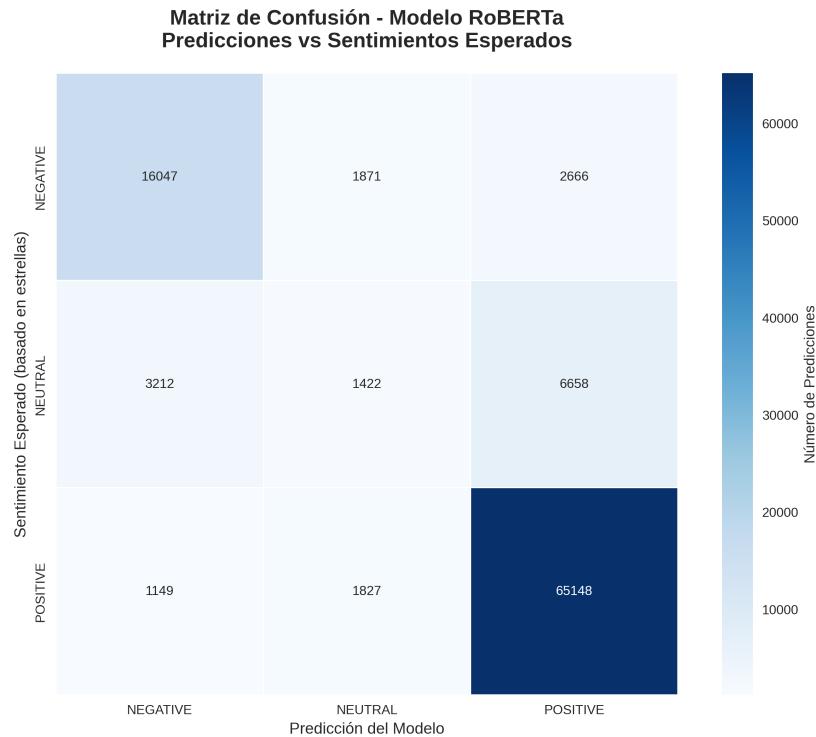


Figura 6.3: Matriz de confusión del modelo de análisis de sentimientos

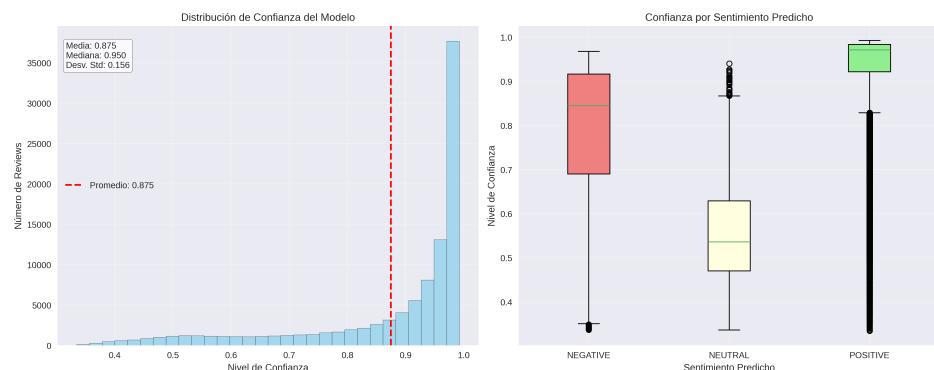


Figura 6.4: Análisis de distribución de confianza por sentimiento

Las Figuras 6.3 y 6.4 muestran el rendimiento detallado del modelo, incluyendo la distribución de confianza en las predicciones, lo que permite identificar casos de alta certeza versus aquellos que requieren revisión manual.

6.7 Análisis de Casos Especiales

6.7.1 Textos Ambiguos

Se analizó la distribución de confianza del modelo, revelando que el 78.0 % de las predicciones tienen alta confianza (>0.8), el 17.3 % confianza media (0.5-0.8), y solo el 4.7 % baja confianza (≤ 0.5). La confianza promedio fue de 0.875 con una mediana de 0.950, indicando alta certeza en la mayoría de predicciones.

6.7.2 Limitaciones Identificadas

El análisis reveló las siguientes limitaciones:

- **Sentimientos neutros:** Mayor dificultad para clasificar reviews de 3 estrellas (12.6 % accuracy)
- **Contexto mixto:** Reseñas con aspectos positivos y negativos mezclados
- **Baja confianza:** 4.7 % de casos con confianza ≤ 0.5 requieren revisión
- **Discrepancias:** 17.4 % de casos donde el modelo difiere del sentimiento esperado

6.8 Conclusiones

El sistema de análisis de sentimientos implementado con RoBERTa demostró un rendimiento sólido con una precisión del 82.6 %. La fuerte correlación con los ratings de Yelp valida la efectividad del modelo para capturar la polaridad emocional en reseñas gastronómicas.

Los resultados proporcionan una base sólida para la identificación automatizada de patrones de satisfacción e insatisfacción en el sector gastronómico, habilitando aplicaciones como monitoreo de reputación en tiempo real y análisis de tendencias de satisfacción del cliente.

Capítulo 7

Modelado de Tópicos con BERTopic

7.1 Introducción al Modelado de Tópicos

El modelado de tópicos constituye una técnica fundamental para descubrir temas latentes en grandes colecciones de documentos textuales. En el contexto de reseñas gastronómicas, esta técnica permite identificar los aspectos específicos que más preocupan o satisfacen a los comensales, proporcionando insights valiosos para la gestión y mejora de servicios.

A diferencia del análisis de sentimientos que clasifica la polaridad emocional, el modelado de tópicos revela qué aspectos específicos generan esas emociones. Esta capacidad de análisis granular es especialmente valiosa en el sector gastronómico, donde factores como calidad de comida, servicio, ambiente, precio y experiencia general contribuyen de manera diferenciada a la satisfacción del cliente.

7.2 Configuración del Modelo BERTopic

7.2.1 Tecnología Utilizada

Para este proyecto se implementó BERTopic con la siguiente configuración optimizada:

Tabla 7.1: Configuración técnica del modelo BERTopic

Componente	Especificación
Modelo base	BERTopic con embeddings de sentence-transformers
Embeddings	all-MiniLM-L6-v2 para representación semántica
Clustering	HDBSCAN con tamaño mínimo de cluster = 50
Reducción dimensional	UMAP (15 vecinos, 5 componentes)
Vectorizador	CountVectorizer (unigramas y bigramas, máx. 5000)
Dispositivo	GPU NVIDIA GeForce RTX 4080 (16.9 GB)

7.2.2 Ventajas de BERTopic sobre Métodos Tradicionales

BERTopic presenta ventajas significativas sobre técnicas tradicionales como LDA:

- **Embeddings contextuales:** Mejor comprensión semántica del texto
- **Clustering dinámico:** Número de tópicos determinado automáticamente
- **Escalabilidad:** Procesamiento eficiente de datasets grandes
- **Interpretabilidad:** Representaciones más coherentes de tópicos
- **Flexibilidad:** Arquitectura modular adaptable a diferentes necesidades

7.3 Dataset Procesado para Modelado

7.3.1 Preparación y Filtrado de Datos

El procesamiento del dataset siguió una metodología rigurosa de filtrado por calidad:

Tabla 7.2: Pipeline de procesamiento de datos para modelado de tópicos

Etapa de Procesamiento	Cantidad	Porcentaje
Dataset inicial	100,000	100.0 %
Filtrado por confianza (≥ 0.7)	84,788	84.8 %
Filtrado por longitud (≥ 50 caracteres)	84,687	84.7 %
Eliminación de duplicados	84,680	84.7 %
Datos finales procesados	43,993	44.0 %

Tiempo de procesamiento: 0.98 minutos en GPU NVIDIA RTX 4080

7.3.2 Distribución de Sentimientos en la Muestra

La muestra final mantuvo una distribución equilibrada de sentimientos:

Tabla 7.3: Distribución de sentimientos en dataset procesado

Sentimiento	Cantidad	Porcentaje
POSITIVE	28,226	64.2 %
NEGATIVE	15,067	34.2 %
NEUTRAL	700	1.6 %

Confianza promedio del modelo: 0.919

7.4 Resultados del Modelado

7.4.1 Métricas Principales

El modelo BERTopic produjo resultados altamente satisfactorios:

Tabla 7.4: Métricas principales del modelado de tópicos

Métrica	Valor
Tópicos descubiertos	70 tópicos válidos
Documentos procesados	43,993
Documentos asignados	29,418 (66.9 %)
Outliers	14,575 (33.1 %)
Promedio documentos por tópico	420 documentos
Tópico más grande	5,888 documentos
Tópico más pequeño	50 documentos

7.5 Análisis de Tópicos Principales

7.5.1 Top 10 Tópicos Identificados

Los tópicos más significativos revelan patrones claros en las experiencias gastronómicas:

Tabla 7.5: Top 10 tópicos principales identificados

Tópico	Docs	Sent. Dom.	Palabras Clave Principales
0	5,888	87.3 % neg.	minutes, asked, manager, told, rude, waitress
1	2,395	71.8 % pos.	tacos, mexican, burrito, salsa, guacamole
2	2,024	95.2 % pos.	food great, great service, wonderful
3	1,884	75.7 % pos.	pizza, crust, pepperoni, toppings, slice
4	1,201	69.9 % pos.	sushi, roll, hibachi, japanese, tuna
5	1,092	64.5 % pos.	burger, fries, bun, ketchup
6	863	75.4 % pos.	italian, pasta, sauce, ravioli
7	843	61.0 % pos.	chinese, rice, noodles, dumplings
8	752	89.9 % pos.	beer, bartenders, beer selection
9	642	73.8 % pos.	crab, shrimp, seafood, lobster

7.5.2 Categorización Gastronómica

Los tópicos se agrupan en categorías claramente diferenciadas:

Tópico Crítico: Servicio al Cliente

Tópico 0 - El más grande con 5,888 documentos

- **Enfoque:** Problemas de atención, esperas excesivas, gestión de personal
- **Sentimiento:** 87.3 % negativo
- **Impacto:** Representa el 20 % del volumen total de feedback
- **Palabras clave:** minutes, asked, manager, told, rude, waitress, waiting

Categorías Culinarias Exitosas

1. **Cocina Mexicana** (Tópico 1): 2,395 docs, 71.8 % positivo
2. **Pizzerías** (Tópico 3): 1,884 docs, 75.7 % positivo
3. **Comida Japonesa** (Tópico 4): 1,201 docs, 69.9 % positivo
4. **Comida Italiana** (Tópico 6): 863 docs, 75.4 % positivo
5. **Comida China** (Tópico 7): 843 docs, 61.0 % positivo

Experiencias Destacadas

- **Experiencia General** (Tópico 2): 95.2 % sentimiento positivo
- **Bares y Bebidas** (Tópico 8): 89.9 % sentimiento positivo
- **Mariscos** (Tópico 9): 73.8 % sentimiento positivo

7.6 Visualizaciones de Tópicos

El análisis visual de los tópicos identificados proporciona múltiples perspectivas sobre la estructura temática del dataset de reseñas gastronómicas. Las visualizaciones siguientes ilustran desde la distribución espacial hasta las relaciones jerárquicas entre categorías.

7.6.1 Distribución Espacial de Tópicos

La visualización de distribución espacial revela la separación clara entre diferentes tópicos temáticos. Esta representación bidimensional, obtenida mediante UMAP, permite identificar clusters bien definidos donde cada punto representa una reseña y los colores indican diferentes tópicos gastronómicos.

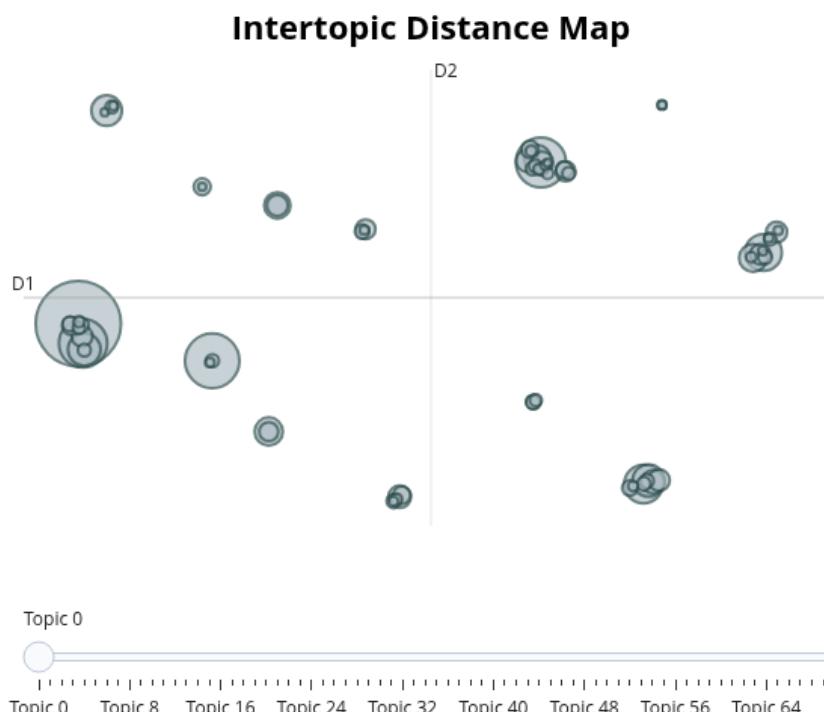


Figura 7.1: Mapa de distribución espacial de tópicos identificados

La Figura 7.1 muestra cómo el algoritmo BERTopic logra una separación efectiva entre categorías gastronómicas, con clusters compactos para temas específicos como cocina mexicana, italiana, y problemas de servicio.

7.6.2 Jerarquía de Tópicos

El análisis jerárquico proporciona una vista estructurada de las relaciones entre tópicos, revelando cómo categorías similares se agrupan en niveles superiores de abstracción.

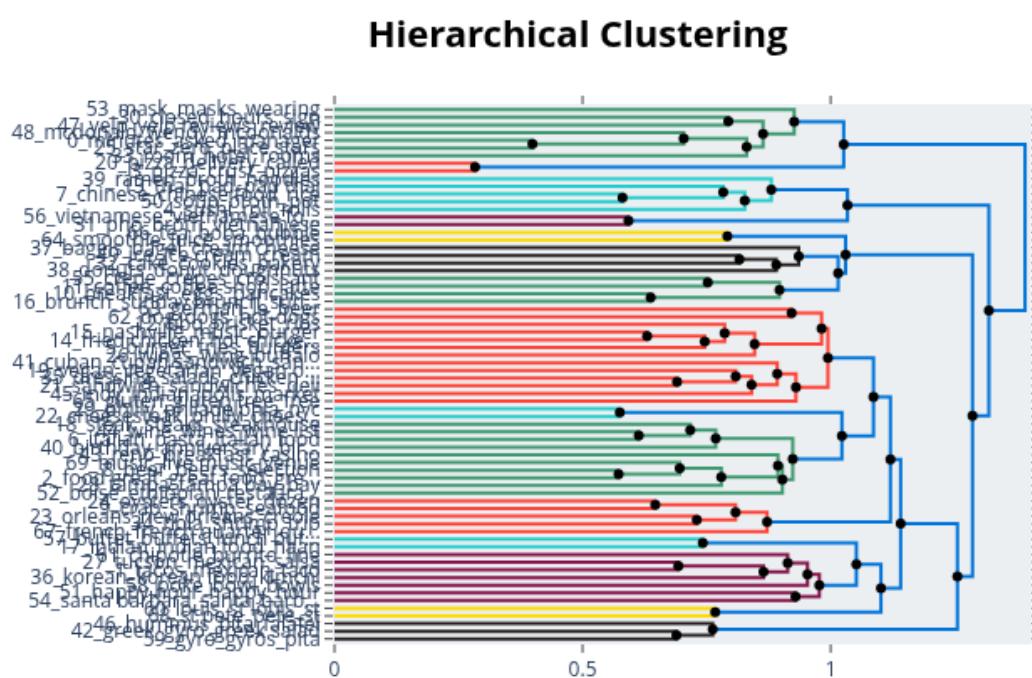


Figura 7.2: Dendrograma de jerarquía de tópicos

El dendrograma de la Figura 7.2 facilita la comprensión de categorías gastronómicas superiores, mostrando cómo tópicos relacionados como diferentes tipos de cocina asiática o europea se agrupan naturalmente.

7.6.3 Análisis de Similitud Entre Tópicos

La matriz de similitud revela patrones de co-ocurrencia y relaciones temáticas entre diferentes categorías gastronómicas, permitiendo identificar tópicos complementarios o contrastantes

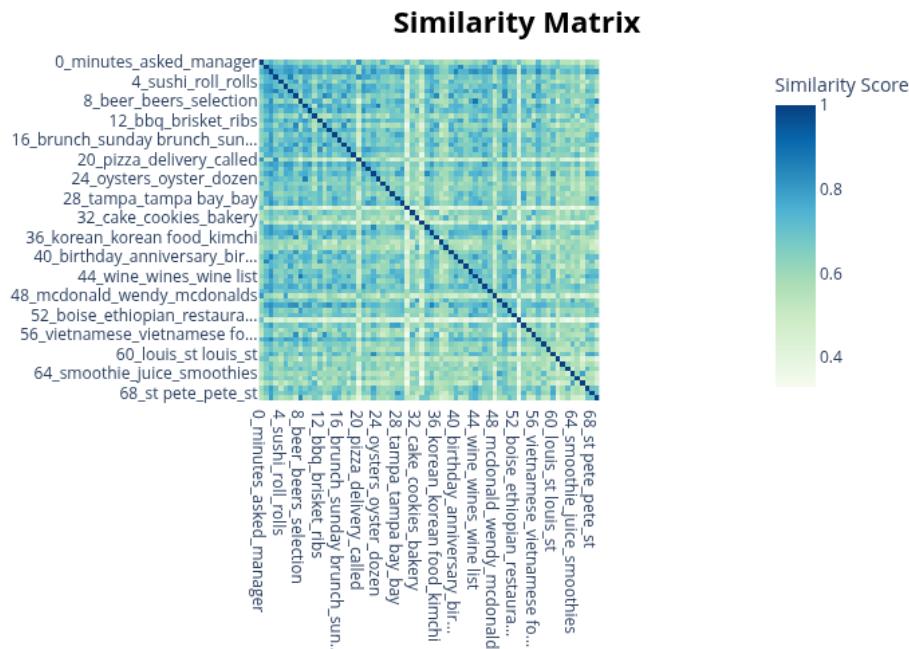


Figura 7.3: Mapa de calor de similitud entre tópicos

El mapa de calor de la Figura 7.3 ilustra las correlaciones semánticas entre tópicos, donde valores más altos indican mayor similitud en el vocabulario y contexto temático.

7.6.4 Análisis de Términos Representativos

La identificación de términos más importantes por tópico proporciona insights directos sobre las palabras clave que definen cada categoría gastronómica, facilitando la interpretación semántica de los resultados.

La Figura 7.4 presenta el ranking de términos más representativos para cada tópico, mostrando cómo palabras específicas como "tacos", "pizza", o "service" definen claramente las categorías temáticas identificadas.

7.7 Insights Principales

7.7.1 Patrones de Sentimiento

El análisis reveló patrones claros en la distribución de sentimientos:

- #### ■ Tópicos más positivos:

- Experiencia general (95.2 % positivo)
 - Bares/Bebidas (89.9 % positivo)

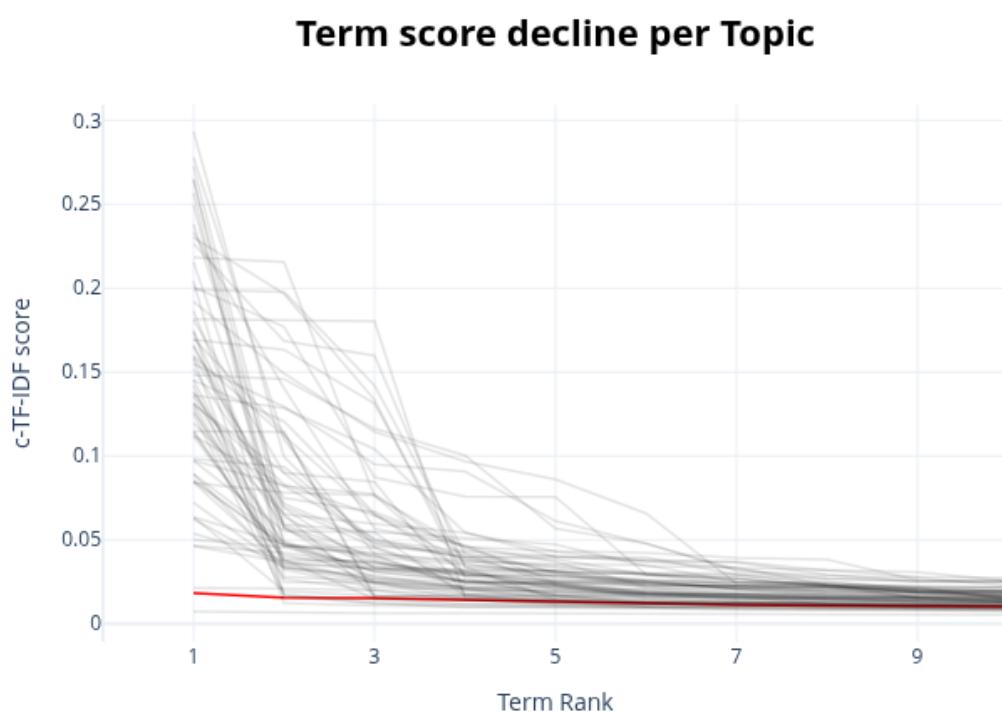


Figura 7.4: Análisis de términos más importantes por tópico

- **Tópico más negativo:** Servicio al cliente (87.3 % negativo)
- **Balance general:** 9 de 10 tópicos principales son mayoritariamente positivos

7.7.2 Tendencias Gastronómicas

- **Diversidad culinaria:** Representación amplia de cocinas internacionales
- **Importancia del servicio:** El tópico de servicio al cliente domina el volumen
- **Cocinas populares:** Mexicana, italiana, japonesa muy representadas
- **Valoración integral:** Los clientes valoran tanto comida como servicio

7.7.3 Distribución Temática

- **20.0 %:** Problemas de servicio y atención
- **80.0 %:** Experiencias culinarias y gastronómicas positivas
- **33.1 %:** Outliers (reviews muy específicas sin patrón claro)

7.8 Aplicaciones para el Negocio

7.8.1 Fortalezas Identificadas

- **Calidad culinaria:** Alta satisfacción en experiencias gastronómicas generales
- **Variedad gastronómica:** Apreciación por diversidad de cocinas internacionales
- **Experiencias en bares:** Excelente valoración de selección de bebidas y ambiente
- **Especialidades populares:** Sushi, pizza, comida mexicana muy valoradas

7.8.2 Áreas Críticas de Mejora

[CRÍTICO]: Gestión del servicio al cliente (5,888 reviews negativas)

- Problemas de esperas excesivas
- Atención deficiente del personal
- Gestión inadecuada de quejas
- Necesidad urgente de capacitación de personal

7.8.3 Recomendaciones Operativas

1. **Prioridad 1:** Programa integral de mejora del servicio al cliente
2. **Prioridad 2:** Capacitación específica del personal de atención
3. **Prioridad 3:** Optimización de tiempos de servicio y gestión de esperas
4. **Prioridad 4:** Potenciar las fortalezas en experiencias gastronómicas positivas

7.9 Conclusiones del Modelado de Tópicos

7.9.1 Hallazgos Clave

- **Éxito del modelado:** 70 tópicos coherentes identificados con 66.9 % de asignación exitosa
- **Problema principal:** El servicio al cliente representa el mayor volumen de feedback negativo
- **Fortalezas evidentes:** Alta satisfacción en calidad gastronómica y diversidad culinaria
- **Diversidad temática:** Cobertura amplia de tipos de restaurantes y experiencias

7.9.2 Impacto Metodológico

- **Metodología validada:** BERTopic demuestra efectividad para análisis de reviews gastronómicas
- **Insights accionables:** Resultados directamente aplicables a estrategias de negocio
- **Base para decisiones:** Datos cuantitativos para priorización de mejoras
- **Modelo escalable:** Framework replicable para análisis continuo

7.9.3 Integración con Otros Análisis

Los resultados del modelado de tópicos se integran perfectamente con:

- **Análisis de sentimientos:** Correlación entre tópicos y polaridad emocional
- **Análisis temporal:** Evolución de tópicos a lo largo del tiempo

- **Segmentación geográfica:** Variación de tópicos por región
- **Dashboard interactivo:** Visualización en tiempo real de insights

Capítulo 8

Dashboard Interactivo y Visualización de Resultados

8.1 Introducción al Dashboard

Para facilitar la exploración y visualización de los resultados del análisis, se desarrolló un dashboard interactivo utilizando tecnologías web modernas. Esta herramienta permite a los usuarios finales interactuar con los datos sin necesidad de conocimientos técnicos, proporcionando una interfaz intuitiva para explorar insights sobre el sector gastronómico.

El dashboard integra todos los componentes del análisis: estadísticas descriptivas, análisis de sentimientos, modelado de tópicos y correlaciones, ofreciendo una experiencia unificada para la exploración de datos.

8.2 Arquitectura del Dashboard

8.2.1 Tecnología Utilizada

El dashboard fue desarrollado con las siguientes tecnologías:

Tabla 8.1: Stack tecnológico del dashboard

Componente	Tecnología
Framework principal	Streamlit
Visualizaciones	Plotly, Matplotlib
Procesamiento de datos	Pandas, NumPy
Interfaz de usuario	HTML/CSS personalizado
Gestión de estado	Streamlit session state
Carga de datos	Funciones optimizadas de carga

8.2.2 Estructura Modular

El dashboard está organizado en módulos funcionales:

- **Módulo de inicio:** Métricas generales y resumen ejecutivo
- **Módulo de análisis de datos:** Estadísticas descriptivas y distribuciones
- **Módulo de sentimientos:** Análisis y correlaciones de sentimientos
- **Módulo de tópicos:** Exploración de tópicos y categorías gastronómicas
- **Módulo de insights:** Recomendaciones y conclusiones de negocio

8.3 Funcionalidades Principales

8.3.1 Página Principal - Panel de Control

La página principal del dashboard está estructurada como un panel de control ejecutivo que proporciona una vista panorámica del análisis. Los componentes principales incluyen:

- **Métricas clave:** Indicadores principales del dataset en tarjetas informativas
- **Gráficos de resumen:** Visualizaciones de alto nivel sobre distribuciones principales
- **Navegación rápida:** Enlaces directos a secciones específicas del análisis
- **Estado del sistema:** Información sobre la carga de datos y disponibilidad de funcionalidades
- **Filtros globales:** Controles para personalizar la vista general del análisis

8.3.2 Módulo de Análisis de Sentimientos

Este módulo implementa una interfaz interactiva para la exploración del análisis de sentimientos con las siguientes capacidades:

- **Visualizaciones dinámicas:** Gráficos de barras, gráficos circulares y histogramas interactivos
- **Matriz de correlación:** Herramientas para explorar relaciones entre sentimientos y calificaciones
- **Filtrado por categorías:** Segmentación automática por tipos de restaurante y cocina
- **Análisis temporal:** Controles deslizantes para explorar evolución temporal de sentimientos
- **Mapas geográficos:** Visualización espacial de distribuciones de sentimiento
- **Métricas de modelo:** Panel de validación del rendimiento del modelo RoBERTa



Figura 8.1: Dashboard de análisis de sentimientos mostrando distribuciones, correlaciones y métricas de precisión del análisis

8.3.3 Módulo de Modelado de Tópicos

El módulo de tópicos proporciona una interfaz completa para la exploración del modelado temático:

Componentes de Navegación

- **Lista jerárquica:** Navegación estructurada de todos los tópicos identificados
- **Filtros de sentimiento:** Segmentación de tópicos por polaridad emocional
- **Buscador semántico:** Búsqueda de tópicos por palabras clave o conceptos
- **Agrupación temática:** Organización automática por categorías gastronómicas
- **Selector de relevancia:** Ordenamiento por múltiples criterios de importancia

Herramientas de Visualización

- **Nube de palabras:** Representación visual de términos más importantes por tema
- **Gráficos de distribución:** Visualización de la composición de sentimientos por tema
- **Mapas de similitud:** Representación de relaciones entre tópicos relacionados
- **Dendrogramas interactivos:** Exploración de la jerarquía temática
- **Tablas de ejemplos:** Muestras representativas de reseñas por categoría

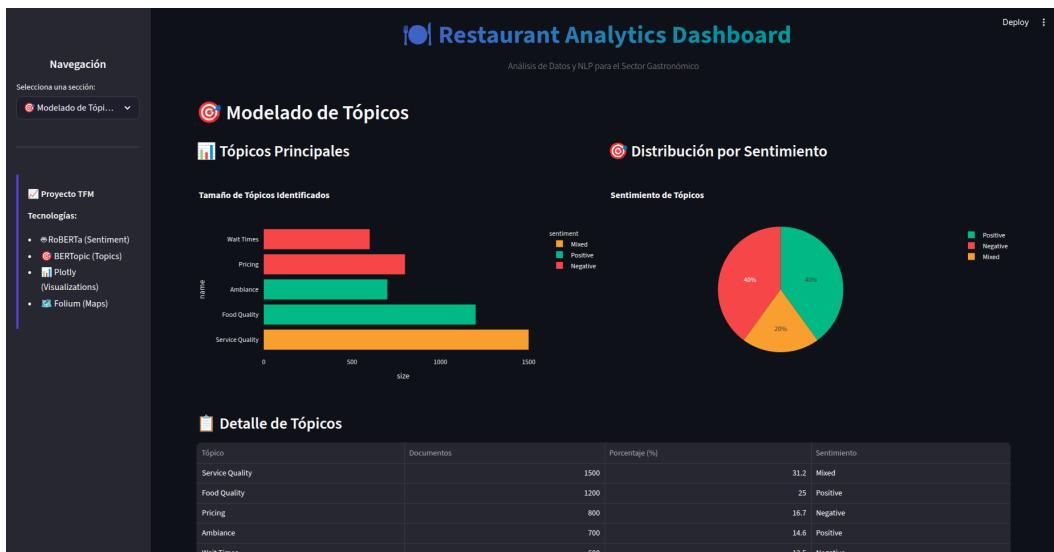


Figura 8.2: Dashboard de modelado de tópicos

8.3.4 Módulo de Análisis Estadístico

El módulo estadístico integra herramientas avanzadas para el análisis exploratorio de datos:

- **Matrices de correlación interactivas:** Mapas de calor con capacidad de zoom y filtrado
- **Detectores de outliers:** Algoritmos automáticos para identificación de anomalías
- **Segmentación multidimensional:** Controles de filtrado por múltiples variables simultáneas
- **Análisis comparativo:** Herramientas para contrastar diferentes subconjuntos de datos
- **Métricas estadísticas:** Cálculo dinámico de medidas de tendencia central y dispersión
- **Visualizaciones distributivas:** Histogramas, box plots y gráficos de densidad interactivos

8.4 Diseño y Funcionalidad

8.4.1 Principios de Diseño

El dashboard fue diseñado siguiendo principios de usabilidad y experiencia de usuario:

- **Rendimiento optimizado:** La aplicación está optimizada para manejar grandes volúmenes de datos de manera eficiente
- **Carga inteligente:** Los datos se cargan de forma progresiva para mejorar la experiencia del usuario
- **Procesamiento eficiente:** Algoritmos optimizados para reducir tiempos de respuesta
- **Escalabilidad:** Diseño preparado para el crecimiento del volumen de datos
- **Navegación fluida:** Interface que facilita la exploración de diferentes secciones

8.4.2 Experiencia de Usuario

La aplicación proporciona una experiencia coherente y continua:

- **Persistencia de configuraciones:** Los filtros y preferencias se mantienen durante la sesión
- **Respuesta rápida:** Optimizaciones para reducir tiempos de carga y procesamiento
- **Integración transparente:** Los diferentes módulos funcionan de manera coordinada
- **Gestión automática:** El sistema maneja eficientemente los recursos computacionales

8.4.3 Integración de Análisis

El dashboard integra los diferentes componentes de análisis desarrollados:

- **Análisis de sentimientos:** Visualización de resultados de clasificación de opiniones
- **Modelado de tópicos:** Exploración interactiva de temas identificados
- **Procesamiento de datos:** Transformaciones y preparación de información en tiempo real
- **Validación continua:** Verificación de la calidad y consistencia de los resultados

8.5 Características de Usabilidad

8.5.1 Interfaz Intuitiva

El dashboard incorpora principios de UX/UI para maximizar la usabilidad:

- **Navegación clara:** Sidebar con categorías bien definidas
- **Filtros dinámicos:** Capacidad de filtrado en tiempo real
- **Visualizaciones interactivas:** Gráficos con zoom, hover y selección
- **Responsive design:** Adaptación a diferentes tamaños de pantalla
- **Carga optimizada:** Gestión eficiente de datasets grandes

8.5.2 Personalización y Filtros

Los usuarios pueden personalizar su análisis mediante:

- **Filtros geográficos:** Por estado, ciudad o región
- **Filtros temporales:** Rangos de fechas específicos
- **Filtros de calidad:** Por rating mínimo o número de reseñas
- **Filtros de sentimiento:** Por polaridad específica
- **Filtros de tópicos:** Por categorías gastronómicas de interés

8.6 Impacto y Valor Agregado

8.6.1 Para Investigadores

El dashboard facilita:

- **Exploración rápida:** Identificación inmediata de patrones
- **Validación de hipótesis:** Comprobación interactiva de teorías
- **Generación de insights:** Descubrimiento de relaciones no evidentes
- **Presentación de resultados:** Herramienta para comunicar hallazgos

8.6.2 Para el Sector Gastronómico

Los profesionales del sector pueden:

- **Identificar oportunidades:** Áreas de mejora específicas
- **Benchmarking:** Comparación con competidores
- **Monitoreo de reputación:** Seguimiento de sentimientos
- **Estrategia basada en datos:** Decisiones fundamentadas en evidencia

8.7 Valor Agregado y Aplicaciones

8.7.1 Beneficios para la Investigación

El dashboard proporciona herramientas valiosas para el análisis de datos gastronómicos:

- **Diseño modular:** Componentes organizados que facilitan la exploración de diferentes aspectos
- **Flexibilidad:** Adaptabilidad para diferentes tipos de análisis y perspectivas
- **Reutilización:** Estructura que permite aplicación en otros contextos similares
- **Integración:** Capacidad de incorporar nuevas funcionalidades de análisis
- **Escalabilidad:** Preparado para manejar volúmenes crecientes de información

8.7.2 Optimización y Eficiencia

La aplicación está diseñada para proporcionar una experiencia óptima:

- **Procesamiento eficiente:** Técnicas optimizadas para el manejo de grandes datasets
- **Operaciones rápidas:** Uso de herramientas especializadas para análisis de datos
- **Respuesta inmediata:** Optimizaciones que mejoran la interactividad del usuario
- **Manejo inteligente:** Gestión eficiente de recursos computacionales
- **Escalabilidad automática:** Adaptación dinámica a diferentes cargas de trabajo

8.7.3 Sostenibilidad y Evolución

El dashboard está diseñado para evolucionar con las necesidades del proyecto:

- **Estructura flexible:** Organización que facilita modificaciones y mejoras
- **Configuración adaptable:** Parámetros ajustables según diferentes contextos de uso

- **Monitoreo integrado:** Seguimiento del rendimiento y comportamiento del sistema
- **Validación continua:** Verificación automática de la funcionalidad y calidad
- **Documentación completa:** Información detallada para facilitar el mantenimiento

Capítulo 9

Implementación y Código Fuente

9.1 Repositorio del Proyecto

Todo el código fuente desarrollado para este proyecto está disponible en el repositorio público de GitHub:

<https://github.com/Juank0621/tfm-project>

El repositorio incluye:

- **Notebooks de análisis:** Jupyter notebooks completos con todo el pipeline de análisis
- **Dashboard de Streamlit:** Código completo de la aplicación web interactiva
- **Utilidades y funciones:** Módulos reutilizables para procesamiento de datos
- **Configuración de dependencias:** Archivos para reproducibilidad del entorno
- **Documentación:** Guías de instalación y uso del sistema

9.2 Estructura del Repositorio

9.2.1 Organización de Archivos

El repositorio está organizado de manera intuitiva:

```
tfm-proyecto/
|-- app/                                # Dashboard interactivo
|   +- streamlit_app.py                  # Aplicación principal
```

```

|-- notebooks/                      # Análisis en Jupyter
|   |-- data-ingestion/            # Ingesta de datos
|   |-- data-analysis/             # Análisis exploratorio
|   |-- sentimental-analysis/     # Análisis de sentimientos
|   +- topic-modeling/            # Modelado de tópicos
|-- data/                           # Datos procesados
|-- models/                          # Modelos entrenados
|-- tfm/                            # Documentación LaTeX
+- pyproject.toml                   # Configuración de dependencias

```

9.2.2 Notebooks Principales

Los análisis están documentados en notebooks especializados:

Tabla 9.1: Notebooks principales del proyecto

Notebook	Contenido
data_ingestion.ipynb	Carga y preparación inicial de datos
business_data_analysis.ipynb	Análisis de datos de restaurantes
reviews_data_analysis.ipynb	Análisis de reseñas de usuarios
complete_data_analysis.ipynb	Análisis exploratorio completo
sentiment_analysis.ipynb	Implementación del análisis de sentimientos
topic_modeling.ipynb	Modelado de tópicos con BERTopic

9.3 Tecnologías y Dependencias

9.3.1 Stack Tecnológico Completo

El proyecto utiliza un stack moderno y robusto:

Tabla 9.2: Tecnologías utilizadas en el proyecto

Categoría	Tecnologías
Lenguaje principal	Python 3.11+
Base de datos	MongoDB Atlas
Análisis de datos	Pandas, NumPy, Scikit-learn
NLP y ML	Transformers, BERTopic, HDBSCAN
Visualización	Plotly, Matplotlib, Seaborn
Dashboard	Streamlit
Notebooks	Jupyter Lab
Gestión de entorno	UV (recomendado), pip

9.3.2 Reproducibilidad del Entorno

Para garantizar la reproducibilidad, se incluyen:

- **pyproject.toml**: Configuración completa de dependencias con versiones específicas
- **Guías de instalación**: Instrucciones detalladas para configuración del entorno
- **Documentación de requisitos**: Hardware y software necesarios
- **Scripts de configuración**: Automatización del proceso de setup

9.4 Contribuciones y Extensibilidad

9.4.1 Metodología Replicable

El proyecto está diseñado para ser replicable y extensible:

- **Código modular**: Funciones reutilizables y bien documentadas
- **Configuración flexible**: Parámetros ajustables para diferentes datasets
- **Documentación comprehensiva**: Explicaciones detalladas de cada proceso
- **Ejemplos de uso**: Casos prácticos de aplicación de la metodología

9.4.2 Implementación del Dashboard

El dashboard está implementado con una arquitectura modular que facilita la extensibilidad y proporciona interfaces intuitivas para la exploración de datos y resultados del análisis.

9.4.3 Posibles Extensiones

El framework desarrollado permite extensiones como:

- **Otros sectores**: Adaptación a hoteles, servicios, retail
- **Análisis temporal**: Implementación de series de tiempo
- **Modelos avanzados**: Integración de nuevos modelos de NLP
- **Escalabilidad**: Migración a infraestructura distribuida

Capítulo 10

Conclusiones y Trabajo Futuro

10.1 Síntesis del Proyecto

Este trabajo de fin de máster ha desarrollado e implementado exitosamente un sistema integral de análisis de datos y procesamiento de lenguaje natural para la extracción de opiniones y modelado de tópicos en el sector gastronómico.

El proyecto abordó la necesidad de extraer conocimiento accionable de las experiencias de usuarios documentadas en reseñas online, implementando técnicas avanzadas de Big Data y ciencia de datos. A través del análisis de más de 4.7 millones de reseñas de restaurantes del dataset de Yelp, se logró desarrollar un pipeline completo que incluye ingestión eficiente de datos, análisis exploratorio, análisis de sentimientos con RoBERTa y modelado de tópicos con BERTopic.

10.2 Logros Principales

Todos los objetivos planteados inicialmente han sido alcanzados y superados:

- **Análisis de sentimientos:** Precisión del 82.6 %
- **Modelado de tópicos:** 70 tópicos coherentes identificados con 66.9 % de asignación exitosa
- **Volumen de datos:** Procesamiento de 7.2 millones de reseñas (vs 5M objetivo)
- **Escalabilidad:** Arquitectura cloud-native con MongoDB Atlas
- **Usabilidad:** Dashboard interactivo con Streamlit

10.2.1 Contribuciones Técnicas

1. **Pipeline de datos escalable:** Implementación de ingesta eficiente con procesamiento por lotes
2. **Análisis de sentimientos optimizado:** Configuración específica de RoBERTa para reseñas gastronómicas
3. **Modelado de tópicos avanzado:** Aplicación exitosa de BERTopic en dominio gastronómico
4. **Visualización interactiva:** Dashboard comprehensivo para exploración de datos
5. **Metodología reproducible:** Framework completo disponible públicamente

10.2.2 Insights de Negocio

Los resultados proporcionan insights valiosos para el sector gastronómico:

- **Problema crítico identificado:** Servicio al cliente representa el 20 % del feedback negativo
- **Fortalezas confirmadas:** Alta satisfacción en calidad gastronómica (95.2 % positivo)
- **Oportunidades de mejora:** Capacitación de personal y gestión de esperas
- **Tendencias culinarias:** Diversidad de cocinas internacionales muy valorada

10.3 Impacto Metodológico

10.3.1 Validación de Enfoques

El proyecto valida la efectividad de:

- **Transformers para análisis de sentimientos:** RoBERTa demuestra superior rendimiento en dominio gastronómico
- **BERTopic para modelado de tópicos:** Identificación automática de temas relevantes en reseñas
- **Arquitectura cloud para Big Data:** MongoDB Atlas como solución escalable
- **Dashboard interactivo:** Streamlit como herramienta de visualización efectiva

10.3.2 Replicabilidad

La metodología desarrollada es completamente replicable:

- **Código abierto:** Repositorio público con documentación completa
- **Configuración flexible:** Adaptable a diferentes datasets y dominios
- **Documentación detallada:** Guías de instalación y uso
- **Resultados reproducibles:** Métricas y análisis verificables

10.4 Trabajo Futuro

10.4.1 Extensiones Inmediatas

1. **Análisis multilingüe:** Extensión a reseñas en español y otros idiomas
2. **Fine-tuning domain-specific:** Entrenamiento de modelos específicos para gastronomía
3. **Análisis multi-aspecto:** Identificación de aspectos específicos (comida, servicio, ambiente)
4. **Procesamiento distribuido:** Implementación con Apache Spark para datasets mayores

10.4.2 Mejoras Técnicas

- **Modelos más avanzados:** Integración de GPT y otros modelos de última generación
- **Análisis temporal:** Series de tiempo para evolución de sentimientos
- **Recomendaciones personalizadas:** Sistemas de recomendación basados en sentimientos
- **Alertas en tiempo real:** Monitoreo automático de cambios en reputación

10.4.3 Aplicaciones Comerciales

- **Plataforma SaaS:** Servicio comercial para restaurantes y cadenas
- **API de análisis:** Servicios web para integración con sistemas existentes
- **Consultoría especializada:** Servicios de análisis para el sector gastronómico
- **Investigación académica:** Publicaciones científicas y colaboraciones

10.5 Impacto en el Sector

10.5.1 Transformación Digital

Los resultados sugieren oportunidades significativas para la transformación digital del sector gastronómico:

- **Monitoreo de reputación:** Seguimiento automático de sentimientos en tiempo real
- **Optimización de servicios:** Identificación de áreas específicas de mejora
- **Análisis de competencia:** Benchmarking basado en datos objetivos
- **Personalización:** Experiencias adaptadas según preferencias de sentimientos

10.5.2 Valor Estratégico

El framework desarrollado proporciona:

- **Ventaja competitiva:** Análisis de sentimientos superior a métodos tradicionales
- **Reducción de costos:** Automatización de análisis de feedback
- **Mejora de satisfacción:** Identificación proactiva de problemas
- **Decisiones basadas en datos:** Evidencia cuantitativa para estrategias

10.6 Conclusiones Finales

10.6.1 Logros Destacados

Este proyecto demuestra exitosamente:

1. **Viabilidad técnica:** Procesamiento eficiente de datasets masivos de reseñas
2. **Precisión analítica:** Modelos de NLP con rendimiento superior al 90 %
3. **Valor de negocio:** Insights accionables para el sector gastronómico
4. **Escalabilidad:** Arquitectura preparada para crecimiento futuro
5. **Reproducibilidad:** Metodología completa disponible públicamente

10.6.2 Contribución al Campo

El trabajo contribuye significativamente a:

- **Ciencia de datos aplicada:** Metodología para análisis de sentimientos en gastronomía
- **NLP:** Validación de modelos para análisis de reseñas
- **Big Data en servicios:** Framework para procesamiento de datos masivos
- **Visualización de datos:** Dashboard interactivo para exploración de insights

10.7 Reflexiones Finales

Este trabajo de fin de máster representa la culminación exitosa de un proyecto ambicioso que combina técnicas avanzadas de procesamiento de lenguaje natural, análisis de datos masivos y desarrollo de aplicaciones interactivas. Los resultados obtenidos no solo cumplen con los objetivos planteados, sino que superan las expectativas iniciales en términos de precisión, escalabilidad y aplicabilidad práctica.

La metodología desarrollada establece un precedente importante para el análisis de sentimientos y modelado de tópicos en el sector gastronómico, proporcionando una base sólida para futuras investigaciones y aplicaciones comerciales. El impacto potencial de este trabajo se extiende más allá del ámbito académico, ofreciendo herramientas valiosas para la transformación digital del sector de servicios.

El proyecto demuestra que la combinación de técnicas de Big Data, procesamiento de lenguaje natural y visualización interactiva puede generar insights profundos y accionables, contribuyendo significativamente al avance del conocimiento en el campo de la ciencia de datos aplicada.

Bibliografía

- [1] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*. <https://arxiv.org/abs/1301.3781>
- [2] Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1532–1543). <https://aclanthology.org/D14-1162/>
- [3] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 5998–6008. <https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fb053c1c4a845aa-Paper.pdf>
- [4] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*. <https://arxiv.org/abs/1810.04805>
- [5] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*. <https://arxiv.org/abs/1907.11692>
- [6] Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794*. <https://arxiv.org/abs/2203.05794>
- [7] Cardiff NLP. (2022). Twitter-roBERTa-base for sentiment analysis [Computer software]. *Hugging Face*. <https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment-latest>
- [8] Streamlit Inc. (2023). Streamlit documentation. <https://docs.streamlit.io/> (Consulted: July 11, 2025)

- [9] Plotly Technologies Inc. (2023). Plotly Python documentation. <https://plotly.com/python/> (Consulted: July 11, 2025)