

Minería de Opinión y Comportamiento del Consumidor Digital: Un Análisis de Sentimientos Mediante Procesamiento de Lenguaje Natural (NLP) en el Sector Gastronómico de Ibagué, Tolima

*"Digital Consumer Opinion and Behavior Mining:
A Sentiment Analysis Using Natural Language Processing (NLP)
in the Gastronomic Sector of Ibagué, Tolima"*

Luis Felipe Campos Cardenas¹
Edisson Eduardo Otalora Murcia²
Juan Gilberto Giraldo Cortes³

¹ Magíster en Administración. Universidad Nacional Abierta y a Distancia - UNAD.
ORCID: <https://orcid.org/0009-0007-7092-6588>
luisf.campos@unad.edu.co

² Magíster en Analítica de Datos. Universidad Nacional Abierta y a Distancia - UNAD.
ORCID: <https://orcid.org/0000-0002-6735-6306>
edisson.otalora@unad.edu.co

³ Magíster en Administración. Universidad Nacional Abierta y a Distancia - UNAD.
ORCID: <https://orcid.org/0009-0007-7092-6588>
juan.giraldo@unad.edu.co

Resumen

La transformación digital ha reconfigurado la relación entre consumidores y establecimientos gastronómicos, convirtiendo las redes sociales en espacios donde los comensales expresan opiniones que influyen directamente en las decisiones de otros usuarios.

El presente working paper tiene como objetivo analizar el sentimiento de las opiniones expresadas por consumidores digitales en Instagram, Facebook y TikTok sobre restaurantes formales y establecimientos de comida rápida en Ibagué, Tolima, mediante técnicas de Procesamiento de Lenguaje Natural (NLP). La investigación adopta un enfoque exploratorio-descriptivo con diseño mixto, predominantemente cuantitativo, que integra métodos de recolección automatizada de datos textuales, preprocesamiento lingüístico y aplicación comparativa de modelos de análisis de sentimientos: enfoques basados en léxico (VADER, TextBlob), aprendizaje automático clásico (Naïve Bayes, SVM, Random Forest) y modelos transformer preentrenados para español (BETO, RoBERTuito). Complementariamente, se emplean técnicas de modelado de tópicos (LDA, BERTopic) para identificar las dimensiones temáticas recurrentes.

El estudio se sustenta en la brecha identificada en la literatura respecto a la escasa evidencia empírica sobre NLP aplicado al español colombiano coloquial en el sector gastronómico, y en la subutilización de redes sociales como fuentes de datos para investigación. Se espera que los resultados permitan generar recomendaciones estratégicas basadas en la voz del consumidor para el fortalecimiento del sector gastronómico local.

Palabras Clave

Análisis De Sentimientos; Procesamiento de Lenguaje Natural; Minería de Opinión; Comportamiento del Consumidor Digital; Sector Gastronómico; Redes Sociales.

Abstract

Digital transformation has reshaped the relationship between consumers and gastronomic establishments, turning social media into spaces where diners express opinions that directly influence other users' decisions. This working paper aims to analyze the sentiment of opinions expressed by digital consumers on Instagram, Facebook, and TikTok regarding formal restaurants and fast-food establishments in Ibagué, Tolima, using Natural Language Processing (NLP) techniques. The research adopts an exploratory-descriptive approach with a mixed design, integrating automated textual data collection, linguistic preprocessing, and comparative application of sentiment analysis models: lexicon-based approaches (VADER, TextBlob), classical machine learning (Naïve Bayes, SVM, Random Forest), and pre-trained transformer models for Spanish (BETO, RoBERTuito). Additionally, topic modeling techniques (LDA, BERTopic) are employed to identify recurring thematic dimensions. Results are expected to generate evidence-based strategic recommendations from the consumer's voice to strengthen the local gastronomic sector.

Keywords

Sentiment Analysis; Natural Language Processing; Opinion Mining; Digital Consumer Behavior; Gastronomic Sector; Social Media.

Introducción

El análisis de sentimientos, también denominado minería de opinión, constituye un campo de estudio que se ocupa del tratamiento computacional de opiniones, sentimientos y subjetividad en textos (Pang y Lee, 2008; Liu, 2012). En la última década, este campo ha experimentado un crecimiento exponencial impulsado por la proliferación de plataformas digitales donde los consumidores generan volúmenes masivos de datos no estructurados en forma de comentarios, reseñas y publicaciones (Medhat et al., 2014). El análisis de sentimientos se ha consolidado como herramienta estratégica en sectores como el comercio electrónico, el turismo, la salud y las finanzas, donde permite clasificar automáticamente las opiniones en categorías de polaridad —positiva, negativa o neutra— proporcionando información valiosa sobre preferencias, niveles de satisfacción y tendencias emergentes del mercado (Cambria et al., 2017; Jim et al., 2024).

En el sector gastronómico, las redes sociales han transformado la forma en que los consumidores descubren, evalúan y recomiendan establecimientos de alimentos. El boca a boca electrónico (eWOM) generado en plataformas como Instagram, Facebook y TikTok influye de manera determinante en las decisiones de consumo gastronómico (Hennig-Thurau et al., 2004; Filieri, 2015). En este contexto, el uso de influencers en estrategias de mercadeo ha demostrado impactos significativos en la percepción del consumidor y en las decisiones de compra (Pulido Sánchez y Ortégón Cortázar, 2023), fenómeno particularmente relevante en el sector gastronómico donde la presentación visual de los platos constituye un factor diferenciador. Asimismo, las estrategias de gestión de marca ciudad pueden potenciar la visibilidad del sector gastronómico local como atributo competitivo del territorio (Franco Ávila et al., 2024).

Desde la perspectiva técnica, el procesamiento de lenguaje natural ha avanzado significativamente con la arquitectura transformer. Devlin et al. (2019) presentaron BERT, un modelo que revolucionó el NLP al capturar relaciones contextuales bidireccionales. Para el español, Cañete et al. (2020) desarrollaron BETO, mientras que Pérez et al. (2022) presentaron RoBERTuito, preentrenado en más de 500 millones de tweets en español, que

superó a otros modelos en tareas de análisis de sentimientos y detección de emociones. La minería de datos, entendida como estrategia organizacional para la toma de decisiones basada en evidencia, ha demostrado su eficacia en contextos universitarios y empresariales (Ovelar y Riveros, 2024), principio que este estudio traslada al sector gastronómico mediante técnicas de NLP.

La inteligencia artificial como estrategia para la gestión de procesos ha sido documentada en el contexto colombiano por De La Hoz Suárez et al. (2024), quienes evidenciaron su capacidad para automatizar la extracción de información relevante de grandes volúmenes de datos no estructurados. En la misma línea, los negocios electrónicos y la integración regional han generado nuevas dinámicas de consumo digital que impactan directamente la percepción y evaluación de servicios gastronómicos en plataformas digitales (Martínez, 2025).

No obstante, persisten brechas significativas: una brecha lingüística, dado que la mayoría de sistemas han sido diseñados para inglés; una brecha geográfica, con ciudades intermedias como Ibagué invisibilizadas; una brecha de plataforma, pues la integración de Instagram, Facebook y TikTok es poco explorada; y una brecha sectorial, ya que el análisis de sentimientos en gastronomía colombiana carece de evidencia empírica (Salgado Reyes y Trujillo Moreno, 2024). El presente working paper busca abordar estas brechas mediante un estudio exploratorio que aplique técnicas de NLP al sector gastronómico de Ibagué, Tolima.

Metodología

La investigación adopta un enfoque exploratorio-descriptivo con diseño mixto, predominantemente cuantitativo, complementado con componentes cualitativos para la interpretación de tópicos. Se emplea un diseño no experimental de corte transversal (Hernández-Sampieri y Mendoza, 2018). El enfoque exploratorio se justifica por la ausencia de estudios previos que apliquen análisis de sentimientos mediante NLP al sector gastronómico de Ibagué.

Población, muestra y unidad de análisis

La unidad de análisis está constituida por comentarios, reseñas y publicaciones textuales sobre establecimientos gastronómicos de Ibagué en Instagram, Facebook y TikTok. Los criterios de inclusión comprenden: textos en español, geolocalizados o referidos explícitamente a Ibagué, publicados dentro del rango temporal definido, y con mínimo de 10 caracteres. Se excluyen: spam, publicaciones promocionales de los establecimientos, contenido sin componente textual y publicaciones en otros idiomas. El tamaño muestral objetivo se establece en mínimo 5.000 comentarios, consistente con estudios similares (Li et al., 2023; Bi et al., 2024).

Recolección de datos

La extracción de datos se realizará mediante técnicas de web scraping y acceso a APIs, utilizando Python como lenguaje principal. Para Instagram, se emplearán herramientas como Instaloader y Selenium para extraer captions, comentarios, hashtags y geotags de publicaciones relacionadas con restaurantes de Ibagué. Para Facebook, se utilizará la librería facebook-scraper complementada con Selenium para recopilar reseñas, comentarios en publicaciones y reacciones. Para TikTok, se empleará la librería TikTokApi junto con Playwright para extraer comentarios en videos, descripciones y hashtags. La recolección de datos a través de múltiples plataformas responde a la necesidad de diversificar las fuentes, dado que cada red social presenta características discursivas y demográficas diferenciadas (Trezza, 2023; Brown et al., 2025). El periodo de recolección abarcará los últimos 12 meses, con el fin de capturar un volumen representativo de opiniones y minimizar efectos estacionales.

Preprocesamiento de datos textuales

Se implementará un pipeline secuencial: (1) normalización (minúsculas, eliminación de URLs, menciones, hashtags y caracteres especiales); (2) tratamiento de emojis mediante conversión a texto descriptivo; (3) tokenización; (4) eliminación de stopwords en español complementada con lista personalizada; (5) lematización con spaCy (modelo es_core_news_lg), preferida sobre stemming

por su precisión semántica en español (Tellez et al., 2017); y (6) manejo de negaciones para preservar contexto de polaridad invertida.

Modelos de análisis de sentimientos

Se implementará una estrategia comparativa progresiva que abarca tres familias de modelos, siguiendo las recomendaciones metodológicas de la literatura reciente (Jim et al., 2024; Medhat et al., 2014). El primer nivel corresponde a los enfoques basados en léxico, que no requieren datos de entrenamiento: se aplicarán VADER (Valence Aware Dictionary and sEntiment Reasoner), adaptado para español, y TextBlob como línea base. El segundo nivel comprende los modelos de aprendizaje automático clásico supervisado: Naïve Bayes, Support Vector Machines (SVM) y Random Forest, implementados con scikit-learn y utilizando representaciones TF-IDF como vectores de características. Para estos modelos, se requiere un subconjunto de datos etiquetados manualmente; se seguirá un protocolo de anotación con al menos dos anotadores independientes, evaluando la concordancia inter-anotadores mediante el coeficiente Kappa de Cohen ($\kappa \geq 0.70$ como umbral de aceptabilidad). El tercer nivel corresponde a los modelos de aprendizaje profundo basados en transformers preentrenados para español: BETO (Cañete et al., 2020), RoBERTuito (Pérez et al., 2022) y el toolkit pysentimiento (Pérez et al., 2021), que representan el estado del arte en NLP para textos en español de redes sociales.

Modelado de tópicos y evaluación

Se aplicarán LDA (gensim) y BERTopic para descubrir dimensiones temáticas subyacentes, determinando el número óptimo de tópicos mediante coherencia C_v (Blei et al., 2003). El desempeño de los modelos se evaluará con Accuracy, Precision, Recall, F1-Score (macro y weighted), matriz de confusión y ROC-AUC, con validación cruzada estratificada ($k=10$). Las comparaciones entre tipos de establecimientos y plataformas se realizarán con pruebas no paramétricas (Kruskal-Wallis, Mann-Whitney U). El desarrollo se realizará en Python 3.10+ con NLTK, spaCy, scikit-learn, Hugging Face Transformers, gensim y BERTopic en Google Colab Pro con GPU.

Consideraciones éticas

La investigación se enmarca en el cumplimiento de la Ley Estatutaria 1581 de 2012 de protección de datos personales en Colombia y su Decreto Reglamentario 1377 de 2013. Se garantizará la anonimización completa de los usuarios mediante la eliminación de nombres de usuario, fotografías de perfil y cualquier información de identificación personal previo al análisis. Solo se recolectarán datos de publicaciones de acceso público, excluyendo cuentas privadas o de acceso restringido. Se respetarán los términos de servicio de cada plataforma en la medida en que las herramientas de extracción lo permitan, y se seguirán las directrices éticas para la investigación con datos de redes sociales establecidas por la Association of Internet Researchers (Markham y Buchanan, 2012; Brown et al., 2025). La base de datos resultante no será compartida públicamente con datos identificables; únicamente se publicarán resultados agregados y anónimos.

Resultados

Dado que la presente investigación se encuentra en fase de desarrollo, esta sección presenta los resultados esperados a modo de hipótesis fundamentadas en la revisión de literatura y en la evidencia empírica de estudios previos con metodologías análogas.

Caracterización esperada del corpus

Se anticipa la conformación de un corpus de entre 5.000 y 10.000 comentarios textuales, con distribución asimétrica entre plataformas: Instagram como mayor aportante dado su carácter visual-textual y amplia adopción gastronómica, Facebook con reseñas más extensas, y TikTok con comentarios breves pero con mayor carga emocional y uso intensivo de emojis.

Distribución de sentimientos esperada

Se hipotetiza una distribución global con sesgo positivo (55-65 % positivo, 15-25 % negativo, 15-25 % neutro), consistente con la asimetría positiva documentada en reseñas gastronómicas en línea (Li et al., 2023; Filieri, 2015). Se esperan variaciones entre tipos de establecimientos: los restaurantes formales concentrarían más comentarios críticos sobre calidad-precio y servicio, mientras que la comida rápida generaría opiniones más polarizadas —positivas en sabor y accesibilidad, negativas en higiene y presentación—. A nivel de plataformas, TikTok exhibiría el sentimiento más polarizado por su naturaleza espontánea, Instagram un sesgo positivo marcado por su cultura visual aspiracional, y Facebook el contenido más equilibrado y detallado.

Rendimiento comparativo de modelos

Se hipotetiza que los modelos basados en transformers (RoBERTuito y BETO) superarán consistentemente a los enfoques basados en léxico y a los modelos de aprendizaje automático clásico, en concordancia con los hallazgos de Pérez et al. (2022), quienes demostraron que RoBERTuito alcanza puntuaciones Macro F1 superiores en tareas de análisis de sentimientos en español. Específicamente, se espera que RoBERTuito obtenga valores de F1-Score entre 0.78 y 0.85, dado su preentrenamiento especializado en más de 500 millones de tweets en español, lo que le confiere mayor capacidad para comprender modismos, abreviaciones y expresiones coloquiales características de las redes sociales colombianas. Los modelos de ML clásico (SVM, Naïve Bayes, Random Forest) probablemente alcancen valores de F1-Score entre 0.68 y 0.76, mientras que los enfoques basados en léxico, al no estar diseñados para las particularidades del español colombiano, presentarían el rendimiento más bajo, con F1-Score estimado entre 0.55 y 0.65.

Dimensiones temáticas anticipadas

Se anticipa la emergencia de al menos seis dimensiones recurrentes: (a) calidad y sabor; (b) atención al cliente; (c) ambiente y espacio físico; (d) relación calidad-precio; (e) tiempos de espera; y (f) experiencia de domicilio. Podrían surgir

tópicos específicos del contexto ibaguereño como referencias a la gastronomía tolimente tradicional. Este tipo de inteligencia aplicada es consistente con las tendencias de la economía naranja y creativa que valorizan las expresiones culturales locales como activos competitivos (Montoya Rendón y Uribe, 2025).

Limitaciones anticipadas

Se reconocen limitaciones potenciales: sesgo de selección por analizar solo usuarios activos en redes; restricciones técnicas de scraping en plataformas con políticas cada vez más restrictivas (Brown et al., 2025; Trezza, 2023); posible sesgo de modelos pre entrenados predominantemente en datos de Twitter; e imposibilidad de establecer relaciones causales dado el diseño no experimental.

Conclusiones

Las conclusiones que se presentan tienen carácter provisional e hipotético, en tanto el estudio se encuentra en fase de desarrollo.

En primer lugar, se espera que la investigación demuestre la viabilidad de aplicar técnicas de NLP para el análisis sistemático de la percepción del consumidor gastronómico en ciudades intermedias colombianas. La integración de tres plataformas de redes sociales como fuentes complementarias constituiría un aporte metodológico significativo, dado que la mayoría de estudios se limitan a una sola fuente (Bi et al., 2024). Se anticipa que cada plataforma capture dimensiones diferenciadas: Instagram privilegiaría la estética y presentación visual, Facebook la reflexión detallada, y TikTok la reacción emocional espontánea.

En segundo lugar, se prevé que los modelos transformer para español — particularmente RoBERTuito— confirmen su superioridad frente a enfoques tradicionales para clasificación de sentimientos en texto coloquial colombiano (Pérez et al., 2022), reforzando la recomendación de adoptar modelos especializados por idioma y dominio discursivo.

En tercer lugar, desde la perspectiva práctica, los hallazgos proporcionarían insumos estratégicos directamente aplicables por los actores del ecosistema gastronómico de Ibagué. La identificación de dimensiones de insatisfacción permitiría focalizar esfuerzos de mejora en la gestión organizacional de los establecimientos. Adicionalmente, los resultados podrían alimentar estrategias de gestión de reputación digital y políticas públicas de promoción del sector gastronómico, un activo estratégico para el desarrollo territorial y la competitividad local, en línea con los pilares del desarrollo sostenible en las organizaciones contemporáneas (Avendaño Avendaño y Mondragón, 2025).

Como líneas futuras de investigación, se proyectan: (a) ampliación a otras ciudades intermedias colombianas para comparaciones regionales; (b) incorporación de análisis multimodal (imágenes y audio); (c) análisis longitudinal para capturar dinámicas temporales; y (d) desarrollo de un dashboard interactivo para monitoreo en tiempo real de la percepción del consumidor, convirtiendo hallazgos académicos en inteligencia de negocio aplicada, en consonancia con el fortalecimiento de las herramientas digitales enfocadas al conocimiento en las organizaciones (Figueroa Infante y Orjuela, 2025).

Bibliografía

- Avendaño Avendaño, M., y Mondragón, S. R. (Coords.). (2025). Pilares del desarrollo sostenible en las organizaciones 5.0. Sello Editorial UNAD. <https://libros.unad.edu.co/index.php/selloeditorial/catalog/book/310>
- Bi, J.-W., Zhu, X.-E., y Han, T.-Y. (2024). Text analysis in tourism and hospitality: A comprehensive review. *Journal of Travel Research*, 63(5), 1-28. <https://doi.org/10.1177/00472875241247318>
- Blei, D. M., Ng, A. Y., y Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3, 993-1022.

Brown, M. A., Gruen, A., Maldoff, G., Messing, S., Sanderson, Z., y Zimmer, M. (2025). Web scraping for research: Legal, ethical, institutional, and scientific considerations. *Big Data & Society*, 12(1). <https://doi.org/10.1177/20539517251381686>

Cambria, E., Poria, S., Gelbukh, A., y Thelwall, M. (2017). Sentiment analysis is a big suitcase. *IEEE Intelligent Systems*, 32(6), 74-80. <https://doi.org/10.1109/MIS.2017.4531228>

Cañete, J., Chaperon, G., Fuentes, R., Ho, J. H., Kang, H., y Pérez, J. (2020). Spanish pre-trained BERT model and evaluation data. PML4DC at ICLR 2020.

De La Hoz Suárez, B. A., Luna Moran, I. F., Manjarres Tete, A. E., y De La Hoz Suárez, A. I. (2024). Inteligencia artificial como estrategia para gestionar los procesos de auditoría financiera. *Revista Estrategia Organizacional*, 13(1), 57-72. <https://doi.org/10.22490/issn.2539-2786>

Devlin, J., Chang, M. W., Lee, K., y Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of NAACL-HLT*, 4171-4186.

Figuroa Infante, S. R., y Orjuela, S. G. (2025). Estado del arte sobre las herramientas digitales enfocadas al conocimiento financiero. *Documentos de Trabajo ECACEN*. <https://publicaciones.unad.edu.co/index.php/working/article/view/10121>

Filieri, R. (2015). What makes online reviews helpful? A diagnosticity-adoption framework. *Journal of Business Research*, 68(6), 1261-1270. <https://doi.org/10.1016/j.jbusres.2014.11.006>

Franco Ávila, J. A., Paredes Castellanos, E., Díaz Casallas, E., y Rodríguez Martínez, A. N. (2024). Gestionar el uso de la “marca ciudad” para contribuir al desarrollo sostenible. *Revista Estrategia Organizacional*, 13(2), 117-141. <https://doi.org/10.22490/issn.2539-2786>

- Hennig-Thurau, T., Gwinner, K. P., Walsh, G., y Gremler, D. D. (2004). Electronic word-of-mouth via consumer-opinion platforms. *Journal of Interactive Marketing*, 18(1), 38-52. <https://doi.org/10.1002/dir.10073>
- Hernández-Sampieri, R., y Mendoza, C. P. (2018). *Metodología de la investigación: las rutas cuantitativa, cualitativa y mixta*. McGraw-Hill Interamericana.
- Jim, J. R., Talukder, M. A. R., y Malakar, P. (2024). Recent advancements and challenges of NLP-based sentiment analysis: A state-of-the-art review. *Natural Language Processing Journal*, 6, 100059. <https://doi.org/10.1016/j.nlp.2024.100059>
- Li, H., Bruce, X., Li, G., y Gao, H. (2023). Restaurant survival prediction using customer-generated content: An aspect-based sentiment analysis. *Tourism Management*, 96, 104707. <https://doi.org/10.1016/j.tourman.2022.104707>
- Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, 5(1), 1-167. <https://doi.org/10.2200/S00416ED1V01Y201204HLT016>
- Markham, A., y Buchanan, E. (2012). *Ethical decision-making and Internet research: Recommendations from the AoIR Ethics Working Committee (Version 2.0)*. Association of Internet Researchers.
- Martínez, D. (2025). *Negocios electrónicos e integración regional*. Documentos de Trabajo ECACEN. <https://publicaciones.unad.edu.co/index.php/working/article/view/10112>
- Medhat, W., Hassan, A., y Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, 5(4), 1093-1113. <https://doi.org/10.1016/j.asej.2014.04.011>
- Montoya Rendón, L. E., y Uribe, P. F. (2025). *Conceptualización de la economía naranja o economía creativa*. Documentos de Trabajo ECACEN. <https://publicaciones.unad.edu.co/index.php/working/article/view/10119>

- Ovelar, W., y Riveros, R. (2024). La minería de datos para desarrollar una estrategia organizacional eficaz en las universidades. *Revista Estrategia Organizacional*, 13(2), 25-48. <https://doi.org/10.22490/issn.2539-2786>
- Pang, B., y Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2), 1-135. <https://doi.org/10.1561/15000000011>
- Pérez, J. M., Giudici, J. C., y Luque, F. (2021). pysentimiento: A Python toolkit for sentiment analysis and social NLP tasks. *arXiv preprint arXiv:2106.09462*.
- Pérez, J. M., Furman, D. A., Alonso Alemany, L., y Luque, F. M. (2022). RoBERTuito: A pre-trained language model for social media text in Spanish. *Proceedings of LREC 2022*, 7235-7243.
- Pulido Sánchez, L. S., y Ortégón Cortázar, L. (2023). Análisis del uso de influencers en mercadeo. *Revista Estrategia Organizacional*, 12(2), 7-27. <https://doi.org/10.22490/issn.2539-2786>
- Salgado Reyes, N., y Trujillo Moreno, G. E. (2024). Análisis de sentimientos en datos de redes sociales. *Dominio de las Ciencias*, 10(1), 314-327. <https://doi.org/10.23857/dc.v10i1.3714>
- Tellez, E. S., Miranda-Jiménez, S., Graff, M., et al. (2017). A case study of Spanish text transformations for Twitter sentiment analysis. *Expert Systems with Applications*, 81, 457-471. <https://doi.org/10.1016/j.eswa.2017.03.071>
- Trezza, D. (2023). To scrape or not to scrape, this is dilemma. *Frontiers in Sociology*, 8, 1145038. <https://doi.org/10.3389/fsoc.2023.1145038>