

UNIVERSIDAD TÉCNICA FEDERICO SANTA MARÍA
DEPARTAMENTO DE INFORMÁTICA
SANTIAGO - CHILE



“DETECCIÓN DE TÓPICOS EN ENCUESTAS DE SATISFACCIÓN AL CLIENTE”

RODRIGO ANDRÉS PÉREZ JAMETT

MEMORIA PARA OPTAR AL TÍTULO DE
INGENIERO CIVIL EN INFORMÁTICA

Profesor Guía: Ricardo Ñanculef
Profesor Correferente: Ignacio Loayza

Diciembre - 2022

DEDICATORIA

Quiero dedicar unas palabras a todas aquellas personas (ya sean familiares o amigos) los cuales he llegado a tener el placer de conocer y cruzarme durante esta vida, si bien no soy una persona que se caracterice por expresar abiertamente el aprecio y/o estima que les tengo, he de admitir que en el interior de mi corazón les guardo un gran espacio a todos y cada uno de ellos, gracias por formar en mi memoria no solo lindos recuerdos, si no por definirme como persona... lamento profundamente ser consciente de que con el pasar de los años muchas buenas amistades quizás quedaran (o ya quedaron) en el olvido y que simplemente hay personas o mascotas que jamas podré volver a ver como para agradecerles y hacerles saber lo que significaron y aportaron para mi crecimiento personal... sin embargo, creo que a su vez no sería justo o digno dedicarles esta memoria ellos, dudo que el concepto o contenido de la misma pueda parecerles relevante, sin embargo creo que lo que realmente es importante, valioso y que con gusto me gustaría dedicarles a ellos, es el esfuerzo y trabajo que esta represento para mi, muchos fueron los días y semanas donde simplemente no pude dedicarle tiempo a realizar las cosas que realmente me gusta hacer, sacrificando así pasatiempos, pasar tiempo con la familia y amigos o simplemente descansar para poder digerir lo que sucedía en mi vida... he de admitir que muchos fueron los dolores de cabeza, tristeza, frustración y presión que en algún momento llegue a sentir durante el transcurso de estos meses... a pesar de todo esto, para mí esta fue una etapa de la vida que no representa ninguna de esas emociones negativas, si no una de total aprendizaje, donde ha pesar de haber cometido mil errores, logre de una u otra manera avanzar... les dedico todo ese esfuerzo, dedicación, tiempo, paciencia y aprendizaje; gracias a todas esas personas que estuvieron y están en mi vida (incluso a los que aportaron el más pequeño granitos de arena), que directa o indirectamente me han definido como persona, me han dado la fortaleza y el carácter necesario no solo para llegar a este punto, si no para atravesarlo y seguir adelante.

AGRADECIMIENTOS

Existe un gran número de personas a las que podría nombrar, creo que por más pequeño, grande, directo o indirecto que haya sido un aporte, este debería ser agradecido, sin embargo las primeras personas a las que debo un agradecimiento son a mi familia, agradezco profundamente tener la familia que tengo y la paciencia que me tienen, gracias a mi padre Jorge P., a mi madre Alicia J., a mi hermano Jorge P., a mi hermana Nicole P. y a mi gata Lulu por formar parte de mi vida y estar presentes en ella cada día. Por otra parte me gustaría agradecer a algunos amigos, he tenido muchísimas amistades, algunas realmente muy buenas y que me encantaría realmente mencionar, sin embargo de todas ellas quiero destacar a dos, gracias a Javier M. e Isidora U. que si bien quizás no seamos los mejores amigos o no pasemos el mayor de los tiempos juntos, han mostrado con sus acciones un tipo de amistad que no he logrado ver en otros, adicionalmente me gustaría agradecer Fernanda A. y a Clement D. ya que, ellos también fueron parte de este proceso y estuvieron presentes en los momentos más críticos, tomándose el tiempo de aconsejarme cuando tenía algunas dudas e inquietudes. Finalmente quiero dar una disculpa a todos y todas aquellas personas que se que se merecen un lugar en este escrito pero que no supe como incluir, no quiero que crean que no las tengo en consideración, realmente les aprecio y soy muy consciente de quien y como me han ayudado, si bien esta vez no fueron nombrados, quiero que sepan que si llegamos vernos y a hablar del tema les daré las gracias cara a cara y con el mayor de los afectos.

RESUMEN

Resumen—

Conocer la relación de las empresas con sus clientes es fundamental para comprender los puntos fuertes y débiles dentro de una empresa, identificar posibles mejoras y compararse con la competencia. Una métrica que mide esta relación es el Net Promoter Score (NPS). Este se calcula mediante encuestas de satisfacción que realizan las empresas a sus clientes; pidiéndoles tanto una calificación como su opinión. Es bajo este contexto que surge como propuesta organizar en distintas agrupaciones los comentarios de estas encuestas mediante técnicas de análisis de texto para facilitar la visualización y entendimiento de los mismos comentarios provistos por los clientes de la empresa.

Palabras Clave— NPS, comentarios, satisfacción, agrupaciones, procesamiento del lenguaje

ABSTRACT

Abstract— Know the relationship of companies with their customers is essential to understand the strengths and weaknesses within a company, identify possible improvements and to compare it with the competition. One metric that measures this relation is the Net Promoter Score (NPS). This is calculated through satisfaction surveys carried out by companies with their customers; asking them for both a rating and their opinion. It is under this context that the proposal arises to organize the comments of these surveys in different groups through text analysis techniques to facilitate the visualization and understanding of the same comments provided by the clients of the company.

Keywords— NPS, comments, satisfaction, agrupations, natural language processing.

GLOSARIO

VIS: Vehículos Infraestructura y Servicios.
NPS: Net Promoter Score.
NLP: Natural Language Processing.
NLU: Natural Language Understanding.
NLG: Natural Language Generation.
LDA: Latent Dirichlet Allocation.
LSA: Latent Semantic Analysis.
DTM: Dynamic Topic Models.
DTC: Dynamic Topic Correlation.
FDP: Función de probabilidad.
GSDMM: Collapsed Gibbs Sampling algorithm for the Dirichlet Multinomial Mixture.
GTM: Global term matrix.
SVD: Singular Value Decomposition.
FSLA: Fuzzy latent semantic analysis.
BoW: Bag Of Words.
AVITM: Autoencoding Variational Inference For Topic Models
SBert: Sentence Bert.
CombinedTM: Combined Topic Models.
HDBSCAN: Hierarchical Density-Based Spatial Clustering of Applications with Noise.
TF-IDF: Term frequency-inverse document frequency.
STS: Semantic Textual Similarity.
UMAP Uniform Manifold Approximation and Projection.
PCA: Principal Component Analysis.
tSNE: t-distributed stochastic neighbor embedding.
KNN: K-Nearest Neighbors.
c-TF-IDF: cluster term frequency-inverse document frequency.
ChatGPT: Chat Generative Pre-Trained Transformer.
NLTK: Natural Language Tool Kit.
StanzaNLP: Stanza Natural Language Processing.
IRBO: Inverted Rank Biased Overlap.
RBO: Rank Biased Overlap.
TU: Topic Uniqueness.
BI: Business Intelligence.
API: Application Programming Interface.
GCP: Google Cloud Platform.
IT: Information technology.
POS: Part Of Speech.
CNN: Convolutional Neural Network.
RNN: Recursive Neural Network.
LSTM: Long Short Term Memory.

ÍNDICE DE CONTENIDOS

RESUMEN	IV
ABSTRACT	IV
GLOSARIO	V
ÍNDICE DE FIGURAS	VIII
ÍNDICE DE TABLAS	IX
INTRODUCCIÓN	1
CAPÍTULO 1: DEFINICIÓN DEL PROBLEMA	2
1.1 Contexto	2
1.2 Problemática y solución actual	2
1.3 Alternativas de mejora	3
1.4 Objetivos	3
1.4.1 Objetivo general	3
1.4.2 Objetivos específicos	3
CAPÍTULO 2: MARCO CONCEPTUAL	4
2.1 El lenguaje y su expresión	4
2.1.1 Lengua	5
2.1.2 Norma	5
2.1.3 Habla	5
2.2 Procesamiento del lenguaje natural (NLP)	5
2.2.1 Limpieza y estandarización de texto	7
2.3 Análisis de sentimientos	8
2.3.1 Partes que constituyen a una opinión	8
2.4 Text Polarity	9
2.5 Modelación de tópicos	10
2.5.1 ¿Cuando es usado el análisis de tópicos?	11
2.5.2 ¿Cómo es llevada a cabo la detección de tópicos?	11
CAPÍTULO 3: ESTADO DEL ARTE	13
3.1 <i>Latent Dirichlet Allocation</i>	13
3.1.1 ¿Cómo funciona LDA?	13
3.1.2 ¿Qué es una distribución Dirichlet?	15
3.1.3 ¿Cómo se interpreta o se entiende esta definición?	16
3.2 <i>Collapsed Gibbs Sampling algorithm for the Dirichlet Multinomial Mixture (GSDMM)</i>	20
3.2.1 ¿Qué significa esto? ¿Por qué este cambio mejora al modelo?	20

3.3 Fuzzy Topic Modeling	21
3.3.1 Metodología <i>Fuzzy Topic Models</i>	22
3.4 Combined Topic Models (CTM)	23
3.5 BERTopic	25
CAPÍTULO 4: PROPUESTA DE SOLUCIÓN	30
4.1 Contexto y características de los datos	30
4.2 Deseables de la solución a proponer	31
4.3 Análisis exploratorio de la data	32
4.4 Preprocesamiento de la data	34
4.5 Hiperparametrización de los modelos	38
4.6 Selección del modelo	38
4.6.1 Puesta en producción	43
4.7 Interpretabilidad del tópico mediante ChatGPT	44
4.8 Propuesta de dashboards para <i>BI</i> mediante Tableau	46
CAPÍTULO 5: VALIDACIÓN DE LA SOLUCIÓN	48
5.1 Motivación	48
5.2 Análisis de resultados	48
5.3 Validación propuestas <i>dashboards BI</i>	50
5.3.1 Feedback dashboard estratégico 19	50
5.3.2 Feedback dashboard resumen general 20	52
5.4 Trabajo a futuro	53
CAPÍTULO 6: CONCLUSIONES	56
6.1 Metas propuestas	56
6.2 Contexto y entorno del trabajo	56
6.3 Dificultades enfrentadas	56
6.4 Propuesta de solución	57
6.5 Limitaciones enfrentadas durante la experimentación	58
6.6 Selección del modelo en base a métricas	59
6.7 Validación e interpretabilidad de los resultados	59
6.8 Puesta en producción	60
6.9 Puntos de mejora	60
ANEXOS	61
6.1 Componentes de los modelos neuronales de <i>embeddings</i> de texto	65
6.1.1 Convolutional Neural Networks, CNN	65
6.1.2 Recurrent Neural Network, RNN	68
6.1.3 Long Short Term Memory, LSTM	70
6.1.4 Attention mechanism	71
REFERENCIAS BIBLIOGRÁFICAS	74

ÍNDICE DE FIGURAS

1	Árbol del problema.	4
2	Representación gráfica de <i>LDA</i> , Fuente: Latent Dirichlet Allocation por David M. Blei, Andrew Y. Ng y Michael Jordan[Blei et al., 2001]	14
3	Ejemplificación <i>LDA</i> , Fuente: Serrano Academy	15
4	Posibles valores que tomaran 2 variables continuas	16
5	Dirichlet en 2 dimensiones	17
6	Posibles valores que tomaran 3 variables continuas	17
7	Distribución <i>Dirichlet</i> en 3 dimensiones con valores de $\alpha = (2,2,2)$, $\alpha = (.9, .9, .9)$	18
8	Distribución de tópicos, Fuente: Serrano Academy	19
9	Boceto de <i>CombinedTM</i>	24
10	Boceto de <i>Zero-Shot Cross-lingual Contextualized Topic Model</i>	25
11	BERTopic's main algorithm	26
12	Condensed Tree	28
13	Serie de tiempo de encuestas NPS respondidas	33
14	Comentarios hechos por compradores promotores en las verticales de Vehiculos e Inmuebles	37
15	Resultados hiperparametrización de tópicos sobre profesionales	41
16	Resultados hiperparametrización de tópicos sobre vendedores	41
17	Resultados hiperparametrización de tópicos sobre compradores	41
18	Pasos del <i>job</i> de automatización implementado en <i>DataSuite</i>	43
19	Dashboard estratégico	46
20	Resumen tópicos detectados	47
21	Pooling Layers	67

22	Explicación visual de las <i>Convolutional Neural Networks</i>	68
23	Types of RNN (Recurrent Neural Network)	69
24	LSTM's	71

ÍNDICE DE TABLAS

1	Encuestas 2022-2023 Mercado Libre VIS	32
2	Palabras por encuestas pre-procesamiento.	33
3	Palabras por encuestas pre-procesamiento.	33
4	Palabras por encuestas pre-procesamiento.	34
5	Palabras por encuestas post-procesamiento.	35
6	Palabras por encuestas post-procesamiento.	35
7	Palabras por encuestas post-procesamiento.	36
8	Resultados obtenidos tras hiperparametrización.	40
9	Resultados obtenidos tras hiperparametrización.	40
10	Resultados obtenidos tras hiperparametrización.	41
11	Interpretación semántica de los resultados. Fuente: Elaboración propia.	44
12	Interpretación semántica de los resultados. Fuente: Elaboración propia.	61

INTRODUCCIÓN

El análisis y entendimiento de texto es un campo amplio dentro de la minería de datos, sin embargo la solución propuesta se enfoque dentro de la rama de la modelación de tópicos, es por ello que en su mayoría el documento aborda el problema desde dos perspectivas, la primera corresponde al manejo del problema desde la modelación de tópicos para textos cortos, mientras que la segunda busca responder a las necesidades del negocio desde la analítica de los resultados, es por ello que se llega a revisar de manera transversal distintos tópicos que guardan cierto grado de relación tanto con la analítica como con la interpretación semántica que puede obtenerse a partir de los resultados de estos modelos.

La presente memoria se divide en 6 capítulos, entre estos es posible encontrar:

- **Definición del problema:** Donde se provee una explicación detallada respecto al problema y contexto del tema tratado, definiendo en base a ello los objetivos del escrito.
- **Marco conceptual:** Sección donde se revisan distintos tópicos relacionados al procesamiento del lenguaje natural en la informática, permitiendo de esta manera un entendimiento de otros campos que podrían ser de relevancias para dar distintos enfoques o soluciones al problema abordado, destacándose temas como la minería de texto, el análisis de sentimientos y la detección de tópicos.
- **Estado del arte:** Se estudian distintos modelos relacionados a la modelación de tópicos, detectando de esta manera sus posibles fortalezas y debilidades para el caso de uso analizado.
- **Propuesta de solución:** Durante este capítulo se analiza el problema enfrentado en mayor detalle y se explica la metodología utilizada durante el desarrollo de la solución, analizando a su vez los resultados obtenidos y su interpretabilidad.
- **Validación de la solución:** A diferencia del análisis de los resultados en base a las métricas que se realizó en la sección anterior, en esta se busca dar un análisis semántico de los resultados obtenidos mediante un entendimiento, enfoque y perspectiva del negocio, detectándose de esta manera posibles puntos de mejora que podrían ser llevados a cabo en un futuro para mejorar la solución propuesta.
- **Conclusiones:** Finalmente se procede a realizar la recapitulación de todos los elementos abordados durante la memoria.

CAPÍTULO 1

DEFINICIÓN DEL PROBLEMA

1.1. Contexto

Numerosas son las formas que tiene una empresa tanto para mejorar como para medir sus avances y logros, esto puede ser llevado a cabo mediante la analítica, la inteligencia de negocios, la mejora continua de procesos o la incorporación de nuevas tecnologías. El objetivo de esta tesis de pregrado es hacer uso de las diversas opiniones emitidas a través de las encuestas de satisfacción al cliente con el fin de generar *insights* y/o accionables para el negocio.

Durante este trabajo se manejarán datos provistos por el área de vehículos, infraestructura y servicios (VIS) de la empresa Mercado Libre, la cuál en breves palabras: corresponde a una de las tantas áreas de negocio que conforman la empresa (encontrándose en diversos países, como por ejemplo: Chile, Argentina, México, etc.), esta actúa como intermediario entre vendedores (los cuales pueden ser personas u otras empresas) y los clientes para la compra y venta de vehículos, inmuebles y servicios; es por ello que estas encuestas (más de setenta mil) abarcarán las opiniones de personas que actúan tanto de clientes o vendedores en la plataforma.

1.2. Problemática y solución actual

Actualmente para llevar a cabo el análisis del contenido provisto por estas encuestas, el equipo de *NPS* realiza una labor manual donde a partir de la definición de ciertas reglas, como por ejemplo, la aparición de una palabra estilo "mejorar" y sus múltiples derivadas actúan como un filtro para la identificación de posibles oportunidades de mejora, retroalimentación o dolores que los clientes puedan estar indicando, ya sea en forma de descontento o satisfacción.

Este esfuerzo manual representa un dolor para el área, dado que se traduce en largas horas de trabajo, las cuales terminan generando una perspectiva incorrecta o incompleta en relación a la opinión formada por los clientes respecto a sus servicios (significando esto una toma de decisiones ineficiente por parte del equipo de *NPS* al momento de levantar accionables dentro de la empresa). La dependencia definida sobre estas reglas de filtro es de gran peso y de difícil medición a la hora de calcular su efectividad, por lo tanto será necesario adoptar métodos más sofisticados y robustos para la tarea.

1.3. Alternativas de mejora

La naturaleza del problema se encuentra estrechamente relacionada al área de la ciencia de datos, esta área posee sub-ramas que permiten tanto el procesamiento del lenguaje natural y la visualización de datos (siendo esta última fundamental para presentar los resultados en reportes útiles para las diversas áreas de negocios), ante esto se guiara el análisis del contenido de estas encuestas mediante modelos de detección de tópicos para la agrupación de temas relacionados presentes en las mismas, generando a su vez breves descripciones de los mismos para dar un entendimiento sobre el contenido de los tópicos modelados.

Entre las principales dificultades y desafíos con los que se tendrá que lidiar y por ende investigar durante el desarrollo de esta memoria se encuentra el hecho de que el problema no posee etiquetas que categoricen en tópicos las respuestas de las encuestas, este hecho guía el problema al área de los métodos no supervisados, por otra parte se destaca que las opiniones emitidas por los encuestados son breves, lo cual a su vez implica una baja presencia de contexto a la hora de tomar una decisión (incluso para un ser humano) para determinar o generar conclusiones importantes que puedan guiar la toma de decisiones dentro del negocio.

1.4. Objetivos

1.4.1. Objetivo general

- Realizar una descomposición de los comentarios provenientes de las encuestas *NPS*, con el objetivo de generar agrupaciones o clusters que faciliten la identificación y generación de conocimientos clave para el desarrollo del negocio.

1.4.2. Objetivos específicos

- Comparar y analizar el desempeño de modelos *NLP* en problemas no supervisados para la identificación de tópicos, a través de métricas de coherencia y diversidad.
- Crear visualizaciones de las agrupaciones formadas por los distintos algoritmos como apoyo para la generación de *insights* del negocio.
- Dar interpretación semántica a los tópicos modelados a través de una breve descripción para canalizar la información.

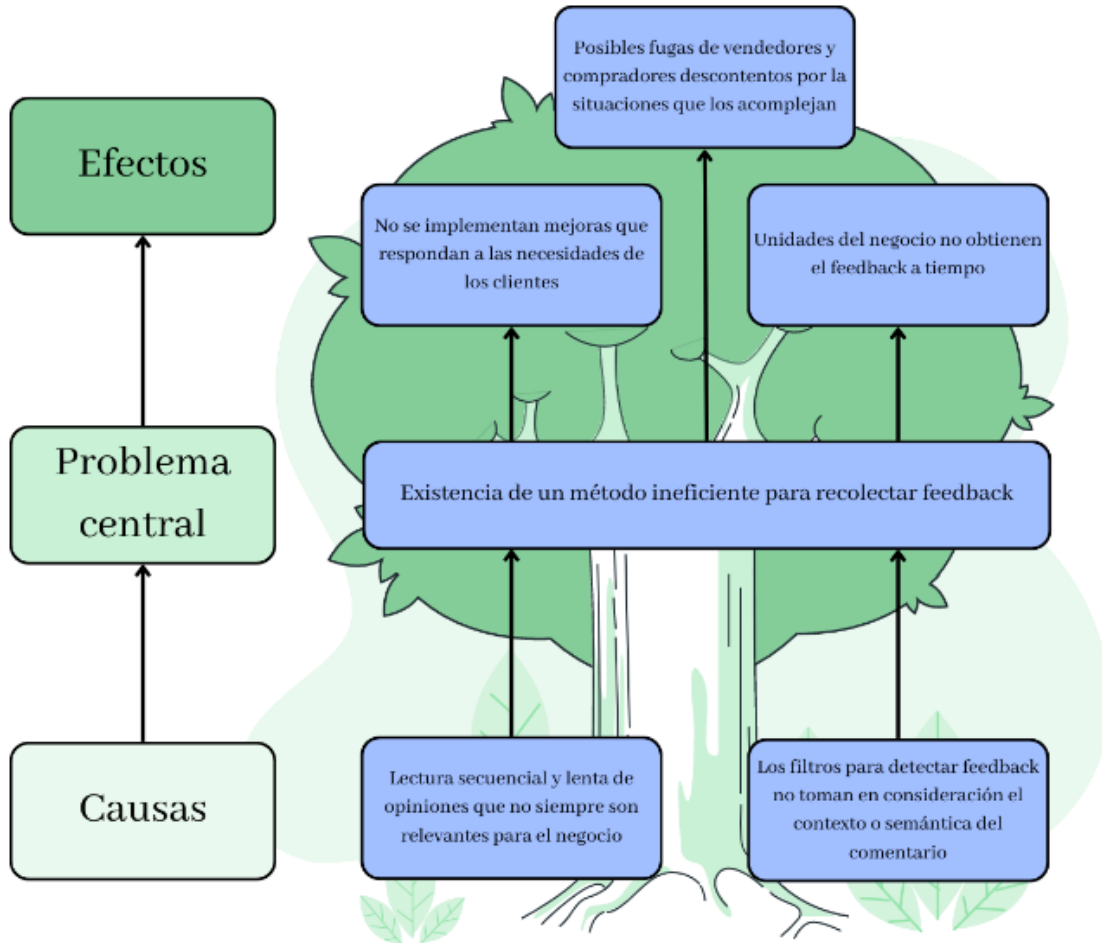


Figura 1: Árbol del problema.
Fuente: Elaboración propia.

CAPÍTULO 2

MARCO CONCEPTUAL

2.1. El lenguaje y su expresión

Tanto el habla como los escritos tienen distintas particularidades las cuales son importantes a tener en cuenta a la hora de evaluar el contexto de un problema, no es lo mismo construir un modelo capaz de procesar el Inglés o el Español, ambos idiomas tienen reglas y definiciones totalmente distintas para estructurar las oraciones e incluso poseen variantes que a pesar de provenir de una misma raíz, presentan diferencias en cuanto respecta a la forma de su

expresión (por ejemplo, aún en la misma lengua se pueden observar marcados cambios en el uso del vocabulario dependiendo del gentilicio del hablante). Si bien estas características son rasgos distintivos del lenguaje, es importante formalizar y definir sus distintos aspectos de tal manera que su comprensión facilite el entendimiento del problema de trasfondo, por ello [Monera, 2019] define:

2.1.1. Lengua

La lengua es un sistema de comunicación estructurado que esta conformado por signos verbales o escritos, este es empleado por una comunidad de hablantes del mismo. Algunos ejemplos de ellos son el Español, Inglés, Chino, etc.

2.1.2. Norma

Hace referencia al conjunto de reglas definidas por la comunidad y que rigen al uso de una lengua, estas se forman por el contexto geográfico y cultural, además estas pueden ir variando con el pasar tiempo.

2.1.3. Habla

Es el uso particular empleado por la persona al utilizar un lenguaje, si bien existen normas que lo rigen, el habla es propio y distinto para cada hablante (esto es fácil de evidenciar cuando notamos que una misma idea puede ser expresada de múltiples maneras por distintas personas).

Bajo estos conceptos, la presente memoria se enmarcará y dará foco al lenguaje Español y Portugués, donde se da uso a la lengua del Español latino y Portugués, con el uso de normas definidas por países como lo son Argentina, Chile, Colombia, México y Brasil, dado un hablar particular que se vera reflejado en la opinión de los distintos encuestados.

2.2. Procesamiento del lenguaje natural (NLP)

El análisis de texto ha dado pasos agigantados en la últimas décadas, su utilidad, importancia y usos son tantos que han beneficiado en gran medida su rápido avance, esto se ve reflejado en las increíbles aplicaciones que han logrado florecer y abrirse paso en los últimos años, claros ejemplos de esto son chatbots que imitan a la perfección el habla humano (chatgpt) [Radford y Narasimhan, 2018], herramientas de auto-completado de código que funcionan en base a la petición escrita por el programador (github copilot) [Dakhel et al., 2023] o

las mismas herramientas de traducción que han estado presente (y han ido mejorado imperceptiblemente) a lo largo de muchos años.

A grande rasgos el procesamiento del lenguaje natural forma parte de una de las ramas de la ciencia de la computación, la cual busca dar la habilidad de comprender el texto a los computadores [IBM, 2021], entre las distintas y numerosas tareas de *NLP* se encuentran:

- **Análisis de sentimientos:** Es un enfoque que busca extraer, convertir e interpretar la opinión de un texto y clasificarlo en positivo, negativo o neutral basado en el sentimiento asociado al mismo [Drus y Khalid, 2019].
- **Categorización de contenido:** Disciplina que busca asignar automáticamente a documentos (ya sean textos planos o páginas webs) una categoría predefinida basada en su contexto [Shen, 2009].
- **Modelación de tópicos:** Esta rama provee una forma conveniente de analizar grandes cantidades de texto no clasificados, la modelación de tópicos permite conectar palabras con significados similares y distinguir entre los usos de las palabras con múltiples significados para asociar documentos y su contenido a tópicos [Xia *et al.*, 2019].
- **Part-of-speech tagging:** También llamado *grammatical tagging*, corresponde a la asignación automática de etiquetas a las partes de un discurso (palabras en una oración), esta es una clasificación gramática que comunmente incluye la identificación de verbos, adjetivos, adverbios, sustantivos entre otros [Chiche y Yitagesu, 2022].
- **Resumen de documentos:** Se define como el proceso de identificar la información más importante y relevante en un documento o conjunto de palabras relacionadas, comprimiendolas en una versión más corta que preserva sus significados generales [Babar, 2013].
- **Traducción de textos:** Tarea que automatiza los procesos de traducción de un idioma a otro.

Con el pasar de los años el campo del *NLP* a ido moviéndose y adoptando nuevos enfoques, como lo son la creación de la subrama del *NLU* o *Natural language understanding* [Kavlakoglu,], el cual va mas allá de entender e interpretar los textos en base a la estructura del lenguaje, contexto y ambigüedades, este aborda directamente el problema de comprender la interpretación semántica de los textos, comprendiendo así lo que se habla y escribe de igual manera a como nosotros los humanos lo interpretamos.

Otra subrama importante a destacar es el *NLG* o *Natural language generation*, la cual busca producir (generar) textos escritos, esta subrama toma importancia a la hora de automatizar la creación de contenido y simplificar información difícil de comprender incluso para un ser humano.

Para la realización de métodos y modelos computacionales capaces procesar texto, se encuentran numerosas técnicas pertenecientes al área de la inteligencia artificial, distinguiéndose las subramas conocidas como machine learning o deep learning. A pesar de ello todas comparten un flujo de trabajo preliminar muy similar, destacándose así las tareas de preprocesamiento y estandarización del texto como uno de los primeros pasos llevados a cabo durante estos trabajos.

2.2.1. Limpieza y estandarización de texto

Esta es una de las primera etapas a manejar y quizá uno de los puntos más críticos y claves para un buen funcionamiento de los algoritmos a utilizar posteriormente [Camacho-Collados y Pilehvar, 2018].

Como ha sido mencionado anteriormente 2.1, el uso del lenguaje varía entre los hablantes producto de un contexto cultural, geográfico o temporal que los condiciona, esto puede ser evidenciado en un uso excesivo de determinadas palabras como los modismos, o en la incursión de errores ortográfico provocando tanto ruido como un cambio semántico en las palabras que se utilizan. Es por ello que estandarizar los escritos se ha vuelto una de las etapas claves para los modelos de la inteligencia artificial.

Entre las distintas técnicas existentes y utilizadas para la tarea, se encuentran:

- **Lematización:** En [Kanerva *et al.*, 2020] se explica que lematizar hace referencia al proceso de llevar a una palabra hacia su forma canónica, la cual usualmente corresponde a la forma de esta que aparecería en el diccionario.
Ejemplo: Jugando, jugamos, jugaron y jugadas serían llevados a la forma de jugar.
- **Stemming:** En [Ismailov *et al.*, 2016] se explica que este es un proceso el cual tiene por objetivo reducir las diversas formas gramaticales de una palabra, como su sustantivo, adjetivo, verbo, adverbio, etc. a su forma origen morfológica o raíz.
Ejemplo: estudiado y estudios serían reducidos a estud.
- **Eliminación de caracteres especiales:** Técnica usualmente performada mediante expresiones regulares con el fin de mantener solamente letras alfabéticas y eliminar así números, *emojis* u otros caracteres especiales.
- **Correcciones ortográficas:** Aplicado para arreglar errores de tipografía que cometen las personas, usualmente se aplican algoritmos que midan la distancia entre palabras, eligiendo así la más cercana. Ejemplo: *passar* sería reemplazado por *pasar*.

El uso de estas y otras técnicas han dado grandes mejoras en los resultados de diversos algoritmos, sin embargo es posible evidenciar en estudios como [Mahendra *et al.*, 2021] que

la evaluación y uso de cada una de ellas queda al criterio del investigador, dependiendo de la efectividad y beneficios de la incorporación que estas traen consigo, cabe destacar que hacer uso de estas etapas de preprocesamiento es opcional y no todos los autores optan por su uso [Cataldi *et al.*, 2010].

2.3. Análisis de sentimientos

Si bien el enfoque de esta memoria es la modelación de tópicos, no es posible dejar de lado el análisis de sentimiento, esta sub-rama del procesamiento del lenguaje natural se encuentra relacionada a la minería de la opinión, lo cual es un tema de trasfondo que se encuentra estrechamente ligado al proceso de análisis de las encuestas NPS.

El análisis de sentimiento es definido como: *"la tarea de detectar, extraer y resumir opiniones, polaridades y/o emociones, basadas en la presencia o ausencia de rasgos de sentimiento"* [Moens *et al.*, 2014], esta tarea puede ser realizada a distintos niveles [Zhang y Liu, 2014], ya sea mediante el análisis de documentos completos, por oraciones o por aspectos (donde en este último se considera que las opiniones y sentimientos están relacionadas a una entidad, siendo así un nivel donde se complementa la detección de tópicos y el análisis de sentimientos).

Es importante destacar que el análisis de sentimiento es realizado en su mayoría sobre opiniones, ante esto es importante definir como se compone una opinión.

2.3.1. Partes que constituyen a una opinión

En [Hernandez, 2019] se explica que: *"De forma general, una opinión se representa por la combinación de cinco componentes: El nombre de una entidad, un aspecto de la entidad, la orientación de la opinión sobre el aspecto de la entidad, el propietario de la opinión, y el tiempo donde la opinión es expresada por el propietario de la opinión"*.

- **Nombre de la entidad:** También conocida como la tarea *Named Entity Recognition*, es una rama perteneciente al procesamiento del lenguaje natural la cual es definida por [Roy, 2021] como la tarea de reconocer e identificar objetos importantes en un texto (personas, compañías o locaciones por ejemplo).
- **Aspecto de la entidad:** Hace referencia a los atributos de una entidad de los cuales se esta dando una opinión positiva o negativa, este punto también puede ser aludido como el tópico, dado que se interpreta como el tema del cual se esta hablando.
- **Orientación de la opinión:** En [Tomar y Sharma, 2016] se explica como la tarea que busca indicar la polaridad de una opinión, asociando de esta manera un sentimiento

positivo, neutro o negativo y con distintos niveles de intensidad a la opinión analizada.

- **Propietario de la opinión:** Esta persona, dentro del contexto de la presente tesis puede ser clasificado como: comprador, vendedor o vendedor profesional. Dada las diferencias intrínsecas de cada perfil, es probable que los tópicos detectados estén gravemente condicionados a ser distintos para estos perfiles.

Este campo es de suma importancia en estudios como [Cataldi *et al.*, 2010] donde no solo se realiza la detección de tópicos, si no que también se da peso a la fuente de la información, realizándose así un trabajo en paralelo para el *ranking* de los distintos perfiles de las personas analizadas.

- **Tiempo de la opinión:** Es importante evidenciar que el tiempo es un factor que tiene efectos y condiciona los diversos temas sobre los que hablan los encuestados, es poco probable que un tema siga siendo discutido con el pasar de los meses o años, identificar la evolución temporal de estos tópicos es un análisis que puede ser útil en ciertos casos. Autores como [Garcia y Berton, 2021] o [Cataldi *et al.*, 2010] realizan un análisis post exploratorio sobre los resultados de los tópicos detectados mediante visualizaciones para enfatizar *trends* o *hot-topics* y relacionar la proliferación de noticias con *tweets* relacionados al tema. Se destaca la relevancia de hacer estos análisis de *trends*, detección de *hot-topics*, detección del surgimiento de un tema, periodicidades, entre otros, cosas que sin duda agregan un valor añadido al análisis para la generación de *insights* del negocio.

2.4. Text Polarity

Como ha sido mencionado anteriormente, las encuestas *NPS* se encuentran ligadas a una opinión que puede ser expresada en manera de descontento o satisfacción por parte de un cliente, a raíz de esto se vuelve interesante también realizar análisis exploratorios que permitan evidenciar la posible presencia o ausencia de estos sentimientos, ver sus correlaciones con las notas que dejan los mismos clientes o incluso segmentar en grupos distintos que se evalúan y analizan por separados en distintos modelos especializados para la modelación de tópicos.

En [Tomar y Sharma, 2016] se explica que la polaridad permite expresar la positividad, neutralidad o negatividad asociada a un texto, esta puede ser entendida como una técnica que permite dar un mayor entendimiento respecto al sentimiento asociado a un texto, por lo general la polaridad de texto se entiende como un valor el cual puntúa entre -1 y 1, con el fin de indicar si una opinión es positiva, neutral o negativa.

Es importante destacar que por ejemplo: la palabra "*mal*" puede fácilmente etiquetarse como negativa dada su usual relación con situaciones de negación o disconformidad, sin embargo este sentimiento puede verse afectado por las demás palabras utilizadas al momento de formar una oración, destacándose así la importancia del contexto para las

mismas, un ejemplo para la palabra "mal" sería la oración: "*no esta para nada mal*", donde el sentimiento evocado por la misma deja de ser negativo. Es fácilmente evidenciable en ejemplos como este que las palabras dependen de un contexto, es por ello que autores como por ejemplo [Li et al., 2010], explican que técnicas como la extracción de la polaridad de un texto toman en cuenta las palabras tanto de manera individual como en conjunto a la hora de entregar una clasificación para las opiniones.

[Bhadane et al., 2015] explica dos posibles maneras de llevar a cabo a esta tarea, la primera es mediante modelos de aprendizaje supervisados, donde se construye un clasificador el cual posee un conjunto de datos de entrenamiento etiquetados donde se expresa si la opinión es positiva o negativa, por otro lado se cuenta con diccionarios de orientación semántica, donde cada palabra es etiquetada dependiendo del sentimiento asociado a su semántica, logrando de esta manera medir el grado en que una palabra es considerada positiva o negativa, a su vez para generar estos diccionarios existen dos formas: creación manual o semiautomática, esta última es llevada a cabo mediante la selección de palabras semillas, desde las cuales haciendo uso de procesos iterativos se logra construir un diccionario mediante las relaciones de proximidad física entre las palabras. Ambas formas de realizar el cálculo de polaridad tienen sus pros y contras, la primera alternativa si bien obtiene buenos resultados, es dependiente del contexto en que el modelo es entrenado, sufriendo así una degradación en su efectividad a la hora de detectar polaridades cuando el texto que se tiene que clasificar se aleja de los tópicos o *corpus* usados a la hora de entrenar tal clasificador, por otro lado, la segunda alternativa es más robusta dada su independencia frente al cambio de distribuciones o temas en los textos, sin embargo su desventaja radica en que estos diccionarios deben ser reconstruidos cada vez que se requiera llevar a cabo un nuevo análisis.

La relevancia y aplicación de técnicas de análisis de sentimientos no se descarta como un trabajo a futuro, su incorporación y el uso de esta técnica en conjunto con la modelación de tópicos abre la posibilidad de llevar a cabo análisis más detallados, por lo que no es menor su mención y entendimiento. Ante esto, resulta relevante ampliar la comprensión que se tiene en el ámbito de la minería de datos, si bien el análisis de sentimientos es capaz de proporcionar una mirada profunda en cuanto respecta a las emociones expresadas en un texto, la modelación de tópicos va más allá, posibilitando descubrir automáticamente los temas subyacentes en un conjunto de textos, es por ello que su estudio será fundamental para llevar a cabo los objetivos definidos en la presente memoria.

2.5. Modelación de tópicos

La modelación, detección o extracción de tópicos es una rama perteneciente al área del procesamiento del lenguaje natural, este es un enfoque el cual puede ser llevado a cabo a través de diversas técnicas de *machine learning* [Snyder, 2015] o mediante redes neuronales [Zhao et al., 2021]. El autor [Xia et al., 2019] explica que esta rama de la minería de datos

tiene por objetivo analizar grandes volúmenes de textos (usualmente llamados documentos en la literatura) buscando y determinando mediante el análisis de aquellas palabras y/o frases que conforman estas serie de documentos, un conjunto de *clusters* que agrupen las ideas o conceptos similares sobre los cuales se esta hablando. Esto es realizado con el fin de detectar patrones subyacentes en la data que muchas veces pueden ser imperceptibles para el ser humano, posibilitando de esta manera la detección de conocimiento nuevo que permita, favorezca y facilite la toma de decisiones inteligentes basadas en los datos.

2.5.1. ¿Cuándo es usado el análisis de tópicos?

La detección de tópicos es una tarea que se vuelve necesaria cuando se tienen grandes volúmenes de documentos desordenados, ya que posibilita organizarlos de manera automática en *clústeres* temáticos, facilitando de tal manera mediante la definición de categorías coherentes la navegación, entendimiento y recuperación de la información perteneciente a los mismos. Para el caso específico de esta tesis se destaca la oportunidad de poder detectar los aspectos o características del negocio (tópicos) sobre los cuales se encuentran hablando los clientes de Mercado Libre (área VIS), sin embargo esta es una tarea extrapolable a otros contextos, como por ejemplo: la detección de tópicos en *reviews* de usuarios a tiendas, películas, análisis de *tweets*, noticias, entre otros.

2.5.2. ¿Cómo es llevada a cabo la detección de tópicos?

La detección de tópicos puede ser categorizada en 3 maneras distintas dependiendo de que enfoque se opte por ocupar al llevarla a cabo [Alkhamees y Fasli, 2016]:

- **Document-pivot:** Técnica que consiste en agrupar documentos relacionados en base al contenido semántico de estos mismos.
- **Feature-pivot:** Se hace uso de aquellas características que respondan a como la gente esta hablando respecto a un tema (por ejemplo: ¿Cuántas personas están *tweteando* sobre un tema?)
- **Probabilistic-pivot:** Se busca realizar la detección de tópicos mediante la frecuencia de las palabras por co-ocurrencias (enfoque sintáctico).

Por lo general en la actualidad, los modelos basados en redes neuronales toman en cuenta de manera directa o indirecta (dependiendo de como sean entrenados) estos tres enfoques.

A pesar de ello, es posible apreciar que autores como [Alghamdi Rubayyi, 2015] dividen la forma en que puede ser llevada a cabo la detección de tópicos en dos partes: La primera corresponde a métodos de modelación de tópicos que consideran el factor del

tiempo irrelevante para el análisis (como por ejemplo *LDA* o *LSA*), mientras que la segunda categoría si incorpora este elemento durante el análisis, denominando así esta familia de algoritmos como *Topic Over Time* (donde se encuentran modelos como por ejemplo *DTM* o *DTC*), durante esta tesis se opta por estudiar modelos pertenecientes a la rama de la modelación de tópicos no dependientes del tiempo, debido a que el problema de negocio analizado en la presente memoria corresponde a un proceso iterativo que se repite tras cada cuatrimestre, donde el factor del tiempo no es tomado en consideración para la toma de decisiones al momento de levantar accionables en el negocio.

Si bien a priori puede parecer que la modelación de tópicos es independiente del largo de los textos analizados, los autores [Sbalchiero y Eder, 2020] y [Albalawi *et al.*, 2020] explican que los modelos existentes en la literatura pueden verse afectados por esta condición, entre más corta sea una oración menos contexto esta maneja, a su vez entre mayor sea el largo de esta, mayor cantidad y diversidad de tópicos pueden ser detectados. Dado que esta tesis se enmarca dentro del campo del análisis de textos cortos, es importante aclarar que a pesar de la existencia de una falta de investigación en la problemática de la detección de tópicos para textos cortos en la literatura, es posible encontrar trabajos como los de [Murakami y Chakraborty, 2022] o [Liu y Xu, 2020], los cuales ponen a prueba diversos modelos en la presencia de este tipo de textos, se denota que en común ambos autores llegan a utilizar modelos basados en atención o *transformers*, sin embargo, es importante también transparentar otras dificultades como la fuente u origen de los textos usados, no es lo mismo analizar textos escritos en un lenguaje coloquial a uno formal, por lo que será importante tener en cuenta que el/los modelos a utilizar sean robustos ante este tipo de ruido, también autores como [Hu *et al.*, 2016], destacan la problemática del cambio de contexto, donde los modelos de *NLP* han mostrado sufrir un degradamiento en sus rendimientos a la hora de ser utilizados en otros contextos bajo los cuales no fueron entrenados (se destaca como solución a esto la técnica *fine-tuning* de la cual se hablara luego).

Conocer y comprender el trabajo de otros autores y como han ido evolucionando las distintas técnicas y modelos pertenecientes a esta área es vital para construir soluciones que permitan mejorar los resultados de modelos previos, es por ello que se revisarán las técnicas más populares que se han desarrollado en el área del procesamiento del lenguaje natural y/o arquitecturas que han sido desarrolladas para otros usos pero que pueden ser usadas en esta área.

CAPÍTULO 3

ESTADO DEL ARTE

3.1. *Latent Dirichlet Allocation*

Esta técnica también conocida por sus siglas en inglés como *LDA* es uno de los algoritmos más usados y conocidos en cuanto a la modelación de tópicos, este fue propuesto por [Blei *et al.*, 2001], los autores explican que *LDA* es un modelo generativo probabilístico que permite modelar una colección discreta de textos, definiéndose así tres niveles jerárquicos de modelos Bayesianos, en el cual cada elemento de la colección es modelado como una combinación finita sobre un conjunto de probabilidades de tópicos subyacentes, esto puede entenderse como que *LDA* es un modelo capaz de aprender a generar documentos y que a su vez es capaz de aprender a modelar tópicos para los mismos.

Es posible observar por el nombre de esta técnica que se buscarán tópicos escondidos o latentes en los documentos, a su vez *Dirichlet* hace alusión al hecho de que se asumirá que la distribución que modela a las palabras presentes en estos tópicos y, valga la redundancia los propios tópicos siguen una distribución *Dirichlet*.

LDA asume dos cosas importantes:

- Los documentos están conformados por una variedad de temas, mientras que los tópicos están formados por una diversidad de palabras.
- Los documentos están condicionados por la distribución de probabilidad de los temas, mientras que los tópicos son la distribución de densidad de probabilidad de las palabras.

3.1.1. ¿Cómo funciona *LDA*?

El modelo *LDA* tiene por objetivo principal aprender una distribución que sea capaz de modelar y asignar tópicos coherentes a un conjunto de textos, a grandes rasgos esto es realizado sobre un conjunto de documentos no etiquetados, donde a partir de una serie de iteraciones se logra modelar una distribución que permite asignar tópicos a este conjunto de datos. Para lograr esto *LDA* entrena una serie de modelos generativos con distintas configuraciones, preservando así aquella configuración que acerque más la distribución aprendida por el modelo generador a la distribución real de los datos que se posea.

Este modelo generador suele ser representado con el siguiente diagrama en la literatura:

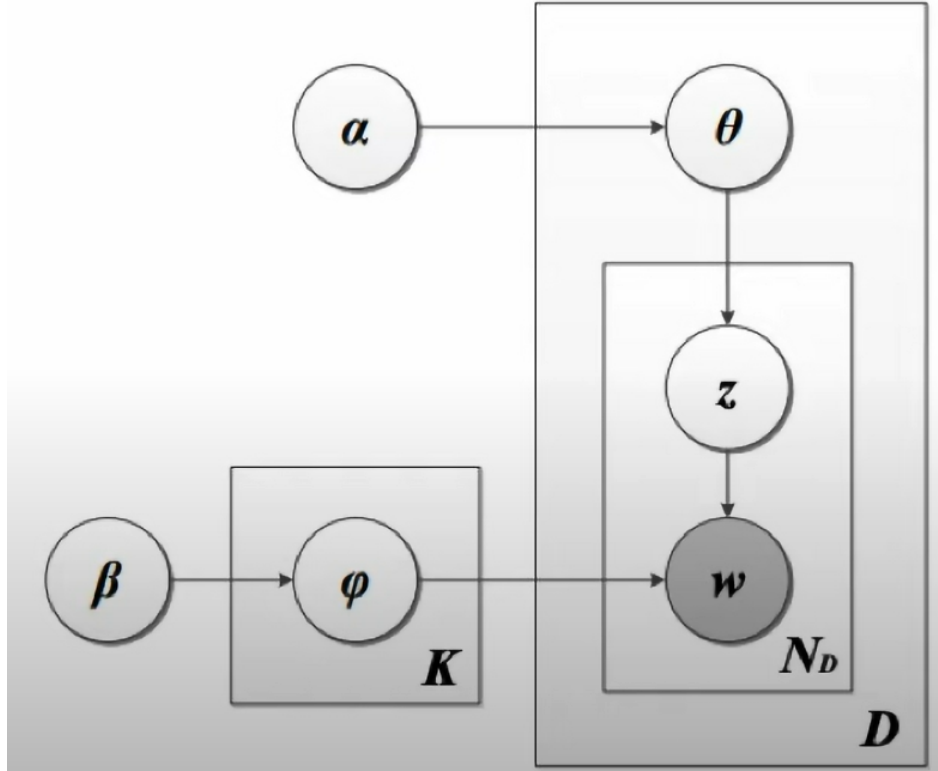


Figura 2: Representación gráfica de LDA, Fuente: Latent Dirichlet Allocation por David M. Blei, Andrew Y. Ng y Michael Jordan[Blei et al., 2001]

A su vez, los autores formalizan el modelo mediante la siguiente ecuación:

$$P(W, Z, \theta, \varphi, \alpha, \beta) = \prod_{j=1}^M P(\theta_j; \alpha) \prod_{i=1}^K P(\varphi_i; \beta) \prod_{t=1}^N P(Z_{j,t} | \theta_k) P(W_{j,t} | \varphi Z_{j,t}) \quad (1)$$

Donde se definen las variables:

W = Conjunto de palabras

Z = Conjunto de tópicos

θ = Distribución multinomial para la obtención de tópicos

φ = Conjunto de distribuciones multinomiales para la obtención de palabras

α = Parámetro que controla la FDP de la distribución Dirichlet que actúa sobre el vector θ

β = Parámetro que controla la FDP de la distribución Dirichlet que actúa sobre W

(2)

A partir de lo observado en esta ecuación es posible identificar tres multiplicatorias, estas pueden ser interpretadas como:

- $\prod_{j=1}^M P(\theta_j; \alpha)$ = Distribución Dirichlet encargada de definir los tópicos asociado a un documento generado j, parte encargada de definir la distribución de probabilidades θ para un documento j.
- $\prod_{i=1}^K P(\varphi_i; \beta)$ = Distribución Dirichlet encargada de definir el vector de probabilidades φ de las palabras asociadas de los tópicos a modelar.
- $\prod_{t=1}^N P(Z_{j,t}|\theta_k)P(W_{j,t}|\varphi Z_{j,t})$ se divide en dos partes:
 - $P(Z_{j,t}|\theta_k)$ = Distribución multinomial de tópicos la cual sirve como función de muestreo de los mismos.
 - $P(W_{j,t}|\varphi Z_{j,t})$ = Conjunto de distribuciones multinomiales que permite muestrear las palabras asociadas a los tópicos muestreados en el punto anterior.

A continuación se disponibiliza de una representación visual para su mejor entendimiento:

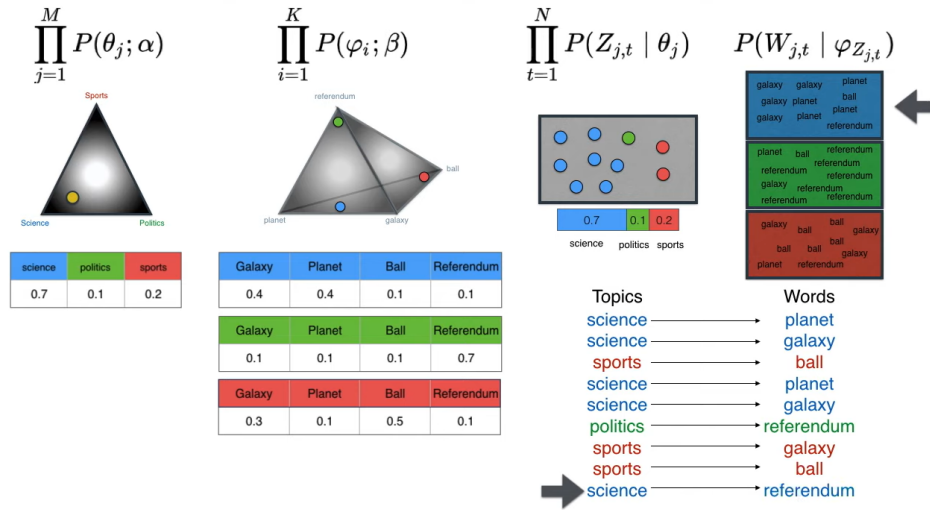


Figura 3: Ejemplificación LDA, Fuente: Serrano Academy

Ante esta definición surge de manera natural la siguiente pregunta:

3.1.2. ¿Qué es una distribución Dirichlet?

Esta es una familia de distribuciones multivariadas continuas que generalizan la distribución beta, estas se parametrizan por un valor generalmente denominado α .

En [Sinharay, 2010] se define de tal manera que, considerando una variable aleatoria

k-dimensional $X = (X_1, X_2, \dots, X_k)$ de modo que los valores X_i sean positivos y su suma sea igual a 1. Si X sigue una distribución *Dirichlet*, entonces su FDP esta dada por:

$$f(x_1, \dots, x_K; \alpha_1, \dots, \alpha_K) = \frac{1}{B(\alpha)} \prod_{i=1}^K x_i^{\alpha_i-1} \quad (3)$$

Donde se define:

$$B(\alpha) = \frac{\prod_{i=1}^K \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^K \alpha_i)}$$

$$\text{Teniendo } \alpha = (\alpha_1, \dots, \alpha_K)$$

3.1.3. ¿Cómo se interpreta o se entiende esta definición?

Para un mejor entendimiento de esto, se define el siguiente ejemplo: Dos variables continuas x_1 y x_2 denotan la probabilidad de ocurrencia de ciertos sucesos, una analogía de esto puede ser que x_1 denote la probabilidad de que el documento a generar sea 90 % conformado por el tópico denominado fútbol y un 10 % conformado por la política. Ante esto, es natural definir que la probabilidad de ocurrencia de cada evento debe tomar valores entre 0 y 1, por lo tanto, el espacio de posibles valores se restringe a:

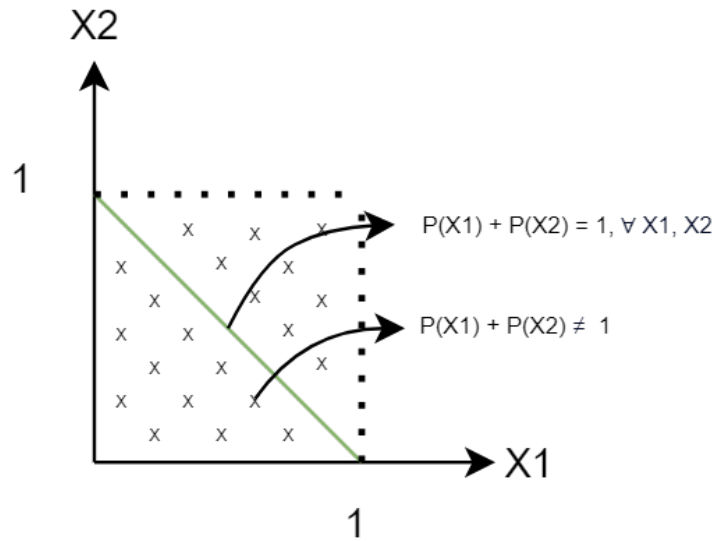


Figura 4: Posibles valores que tomaran 2 variables continuas
Fuente: Elaboración propia.

Sobre este espacio de posibles probabilidades de tópicos a obtener por cada documento a generar (denotado por la recta verde), la función de densidad de probabilidad asociada a

la distribución *Dirichlet* podrá modelar distintas funciones de densidad dependiendo de los valores α_1 y α_2 que se asignen:

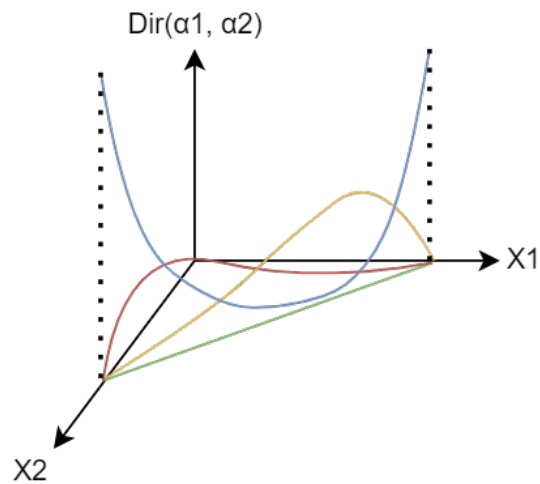


Figura 5: Dirichlet en 2 dimensiones
Fuente: Elaboración propia.

Esta idea es aplicable a más dimensiones (en consiguiendo más tópicos), en el caso de 3 variables conjuntas se tendría:

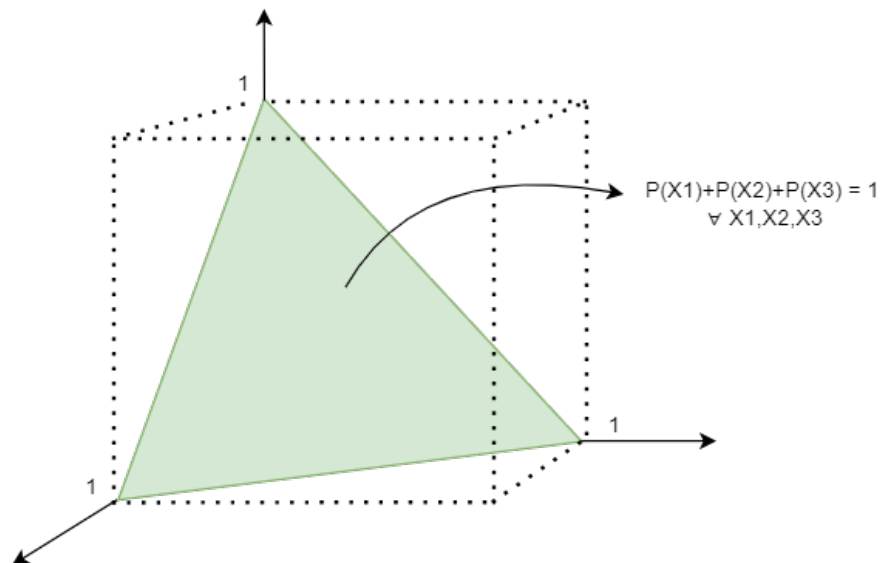


Figura 6: Posibles valores que tomaran 3 variables continuas
Fuente: Elaboración propia.

Donde como ha sido mencionado anteriormente, sobre este espacio de posibles valores, la función de densidad de probabilidad asociada a la distribución *Dirichlet* podrá tomar distintas funciones de densidades dependiendo de los valores α_1 , α_2 y α_3 que se asignen.

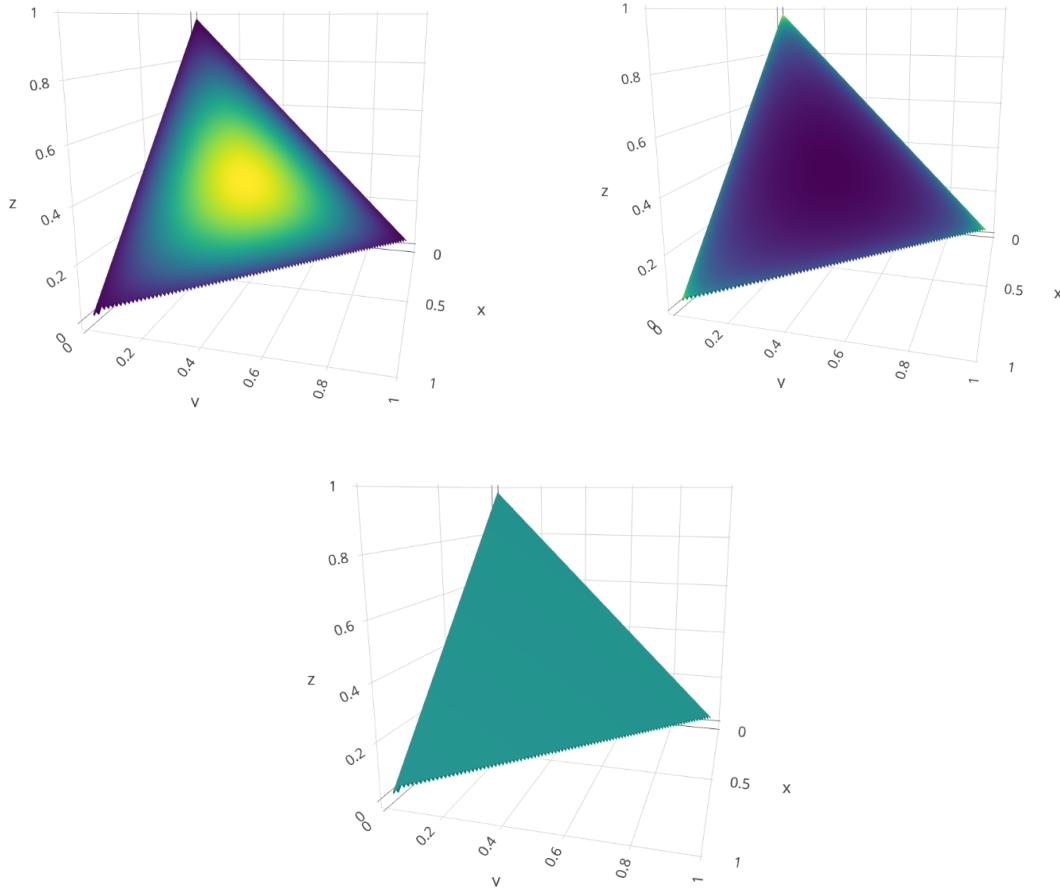


Figura 7: Distribución *Dirichlet* en 3 dimensiones con valores de $\alpha = (2, 2, 2)$, $\alpha = (.9, .9, .9)$ y $\alpha = (1, 1, 1)$ respectivamente Fuente: Dirichlet Plotly.

A partir de estas imágenes es importante señalar que las regiones más oscuras poseen una baja densidad, mientras que aquellas más claras poseen una alta densidad, la cual esta sujeta a los valores de α , esto puede interpretarse como una función generadora de funciones de densidad, donde al ajustar los valores de α se obtendrán distintas funciones de densidad, además es importante señalar que esta función no esta sujeta a 2 o 3 dimensiones, estas figuras (también conocidas como *simplex*) no están sujetas a un número determinado de dimensiones, es aquí donde los modelos *LDA* aprovechan esto para definir distintas funciones de densidad que sean capaces de modelar las distribuciones tanto de tópicos como palabras mediante el uso de la función de probabilidad de la distribución *Dirichlet* 3.1.1.

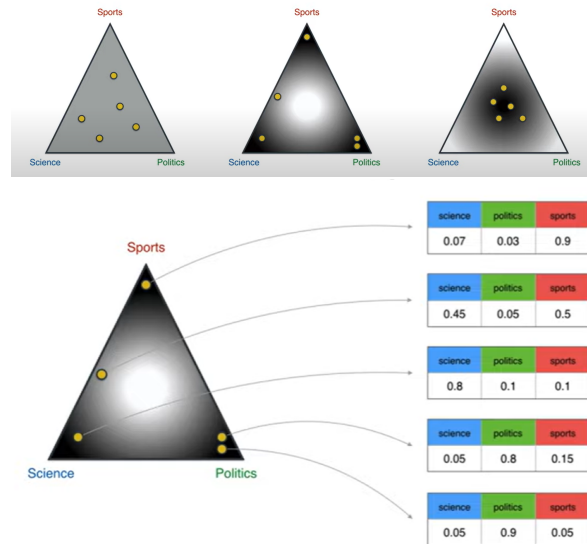


Figura 8: Distribución de tópicos, Fuente: Serrano Academy

Es posible evidenciar en las figuras, que hacer uso de valores de α menores a 1 favorece la obtención de probabilidades de coocurrencia de eventos menos "intersectables" entre sí, dentro del contexto de la generación de tópicos esto se entiende como que una menor probabilidad de que una palabra sea asignada a múltiples tópicos que compartan la misma probabilidad.

A pesar de que *LDA* a revolucionado el campo de la modelación de tópicos en un su momento, este presenta varias limitaciones que ya han sido detectadas por diversos autores [Stoy, 2021] entre las más importantes se destacan:

- El número de tópicos es estático, por lo que se requiere conocer de antemano.
- Para su uso se ha de emplear *Bag of Words*, la cual es una técnica que asume que el intercambio de palabras no afecta el significado de las oraciones.
- Descuidan comúnmente la relaciones de co-ocurrencia, dando como resultado detecciones incompletas y la incapacidad de descubrir relaciones latentes a través del contexto.
- No son situables para textos cortos generados por personas. [Hajjem y Latiri, 2017].

A pesar de la evidencia en la literatura sobre las limitantes que puede suponer el uso de este modelo para el desarrollo de la presente memoria, se rescata su uso dado que este servirá como punto de partida para llevar a cabo una comparación que permita contrastar los beneficios de otros modelos por sobre este.

3.2. ***Collapsed Gibbs Sampling algorithm for the Dirichlet Multinomial Mixture (GSDMM)***

Cómo ya se ha explicado con anterioridad *Latent Dirichlet Allocation* presenta grandes desventajas a la hora de manejar textos cortos y dispersos, dado que depende fuertemente de la co-ocurrencia de las palabras y además ignora el orden inherentemente presente en las oraciones, perdiendo así la habilidad de capturar o comprender la semántica presente en los mismos, sin embargo con el pasar de los años han surgido diversas variaciones al mismo, dentro de las cuales destaca *Collapsed Gibbs Sampling algorithm for the Dirichlet Multinomial Mixture (GSDMM)*, esta es una modificación que se aplica al modelo LDA, donde se reemplaza el enfoque aplicado a la hora de llevar a cabo la inferencia de los tópicos, pasando de un enfoque de inferencia variacional a uno de muestreo de Gibbs, sin embargo:

3.2.1. ¿Qué significa esto? ¿Por qué este cambio mejora al modelo?

Bueno, la inferencia variacional busca aproximar la distribución posterior de los tópicos y los distintos parámetros del modelo mediante un proceso iterativo llamado algoritmo de *expectation maximization* (esto debido a que trabajar la distribución posterior es computacionalmente costoso, denominándose así un problema intratable), durante este proceso se busca aproximar la distribución real con otra distribución más simple y que sea por ende más fácil de manejar computacionalmente, bajo esta idea se busca minimizar la divergencia *Kullback-Leibler* entre la distribución real y la aproximada, al realizar todo este proceso se cae en supuestos, lo cual lleva consigo sesgos en los resultados y estimaciones finales, por otro lado el muestreo de Gibbs, es un algoritmo perteneciente a la familia de algoritmos estadísticos llamada *Markov chain Monte Carlo*; este algoritmo selecciona aleatoriamente una palabra y re asigna el tópico (aleatorio en un principio) de esta basándose en las asignaciones de las demás palabras tanto a nivel oración como corpus, convergiendo eventualmente hacia una distribución mas estable con el pasar de las iteraciones; es esta aleatoriedad lo que se conoce como muestreo estocástico, el cual permite probar distintas configuraciones de asignaciones de tópicos, logrando así una mayor diversidad de distribuciones posterior conjuntas a probar, obteniendo a su vez muestras más representativas de la muestra posterior conjunta que buscamos aproximar, sin la necesidad de hacer supuestos que simplifiquen el problema a tratar y sesguen los valores de las variables y parámetros a seleccionar. Ante estos motivos se opta por probar esta variación del modelo LDA, que segun autores como [Yin y Wang, 2014] se ha demostrado mejoras considerables respecto al modelo LDA convencional, incluso ante la presencia de textos cortos.

3.3. Fuzzy Topic Modeling

Esta familia de modelos deben su nombre dado el uso de algoritmos *soft-clustering* los cuales son también conocidos como modelos difusos. Los autores [Karami *et al.*, 2017] propusieron el estudio de modelos *Fuzzy* para el manejo de textos médicos como una solución alternativa a *LDA* y sus variantes. Dentro esta familia de algoritmos se destacan tanto *FLSA* como su variante *FLSA-W*.

Los autores hacen uso del modelo *Fuzzy c-means*, el cual es un modelo que surge como una alternativa a *K-Means*, donde a diferencia de este último que asigna los elementos a un único *cluster* se opta por asignar un grado de pertenencia a los mismos.

Algorithm 1 Fuzzy c-means

```

1:  $c \leftarrow$  número de clusters
2:  $X \leftarrow \{x_1, x_2, x_3, \dots, x_n\}$ , puntos de datos
3:  $V \leftarrow \{v_1, v_2, v_3, \dots, v_c\}$  ▷ centroides de los clusters
4: while  $\varepsilon > \max_{ij} \{|u_{ij}^{(k+1)} - u_{ij}^k| \}$  do ▷ Criterio de termino
5:    $u_{ij} = \frac{1}{\sum_{k=1}^c (\frac{d_{ij}}{d_{ik}})^{\frac{2}{m-1}}}$  ▷ Pertenencia del dato i al cluster j
6:    $v_j = \frac{(\sum_{i=1}^n (u_{ij})^m x_i)}{(\sum_{i=1}^n (u_{ij})^m)}, \forall j = 1, 2, \dots, c$  ▷ Re ajuste centroides de los clusters
7: end while

```

Este modelo al igual que *KMeans* es un algoritmo iterativo que busca reajustar los centroides tras cada asignación de los puntos a un *cluster* (en este caso un grado de pertenencia a los múltiples tópicos), usando como criterio de termino un valor ϵ como *threshold* para denotar la mínima distancia de cambio necesaria para no volver a iterar.

Es posible evidenciar en el punto 5, que entre mayor sea la distancia entre el punto x_i con el centroide k y menor sea la distancia entre el punto x_i al centroide j , mayor grado de pertenencia tendrá el punto i al cluster k dado que menor grado de pertenencia aportaran los centroides de clusters lejanos, a su vez centroides más cercanos disminuirán el grado de pertenencia del punto i al cluster k .

Es esta idea de grado de pertenencia la que motiva a los autores [Karami *et al.*, 2017] a utilizar un enfoque similar pero con un acercamiento distinto al usado por los modelos probabilísticos como *LDA*, asignando de esta manera a cada documento un grado de pertenencia a los distintos tópicos.

Al igual que los modelos de análisis semántico latente, el objetivo principal de la familia de algoritmos *Fuzzy Topic Modeling* es encontrar dos matrices de probabilidades: $P(T|D)$ (matriz de probabilidad del tópico dado el documento) y $P(W|T)$ (matriz de probabilidad de la palabra dado un tópico).

3.3.1. Metodología *Fuzzy Topic Models*

Estos modelos parten desde la construcción de dos matrices para la obtención de una representación vectorial de los documentos en función de las palabras que los componen:

- **Matriz de pesos locales (LTM):** La cual denota la ocurrencia de las palabras en cada documento.
- **Matriz de pesos globales (GTM):** Esta se conforma por la multiplicación de la matriz de pesos locales con respecto a la matriz definida por uno de los siguientes métodos.

$$\text{Entropy} = 1 + \frac{\sum_j p_{ij} \log_2 p_{ij}}{\log_2 n} \quad (4)$$

$$\text{IDF} = \log_2 \frac{n}{\sum_j b(t_{ij})} \quad (5)$$

$$\text{Normal} = \frac{1}{\sqrt{\sum_j b(t_{ij}^2)}} \quad (6)$$

$$\text{ProbIDF} = \log_2 \frac{n - \sum_j b(t_{ij})}{\sum_j b(t_{ij})} \quad (7)$$

Es importante destacar que el uso de estas u otras técnicas tendrá un gran impacto en los resultados obtenidos, dado que estas matrices de pesos globales conforman la representación vectorial que sientan las bases del análisis, describiendo el contenido de los documentos.

Tras el calculo de estas representaciones vectoriales, al igual que en el análisis semántico latente se procede a realizar reducción de dimensionalidad mediante la técnica *Singular Value Decomposition (SVD)* en las matrices $U\Sigma V^T$, obteniendo de esta manera a partir de la matriz de términos de los documentos *GTM*, la matriz de documentos (U), la matriz diagonal (Σ) y la matriz de términos (V^T), esto se realiza con el fin de aplicar *fuzzy clustering* sobre la matrices U^T o V^T (en caso de llevar a cabo la primera opción se estaría realizando *clustering* sobre documentos, mientras que en la segunda opción sobre las palabras), para así comenzar a calcular las matrices de probabilidades $P(T|D)$ y $P(W|T)$.

FLSA opta por aplicar *Fuzzy clustering* sobre la matriz de documentos U , obteniéndose de esta manera el grado de pertenencia de los tópicos dado un documento que se denota como $P(T|D)^T$. Tras obtener este resultado se procede a calcular $P(W|T)$ para la obtención de la matriz de probabilidades de la palabras dado un tópico, la cual se define como:

$$P(W_i|T_k) = \sum_{j=1}^N P(W_i|D_j)^T P(D_j|T_k) \quad (8)$$

Donde el primer subtermino $P(W_i|D_j)$ corresponde a la probabilidad de la palabra i dado el documento j , la cual se calcula como:

$$P(W_i|D_j) = \frac{P(W_i, D_j)}{\sum_{i=1}^M P(W_i, D_j)} \quad (9)$$

Mientras que el segundo subtermino, denota la probabilidad del documento j dado el t pico k , el cual se define como:

$$\begin{aligned} P(D_j|T_k) &= \frac{P(D_j, T_k)}{\sum_{j=1}^N P(D_j, T_k)} \\ P(D_j, T_k) &= (P(T_k|D_j) \cdot P(D_j))^T \\ P(D_j) &= \frac{\sum_{i=1}^M P(W_i, D_j)}{\sum_{i=1}^M \sum_{j=1}^N P(W_i, D_j)} \end{aligned} \quad (10)$$

Donde M es el n mero de palabras  nicas y N es el n mero de documentos.

Los autores [Rijcken *et al.*, 2022] destacan avances en el estado del arte a partir de los resultados del modelo *FLSA-W*, donde reportan que el modelo es capaz de obtener mejores resultados tanto en coherencia como diversidad en comparaci n a modelos como LDA y sus variaciones. Esto se atribuye al hecho de que los modelos fuzzy son m s flexibles a la hora de asignar t picos a los documentos, el simple hecho de tener grados de pertenencia permite manejar de mejor manera documentos ambiguos o que abarquen m ltiples t picos.

3.4. Combined Topic Models (CTM)

Los autores de este modelo parten desde la base que la mayor a de los modelos de t picos de la  poca siguen utilizando *BoW* para representar el *input* de datos [Bianchi *et al.*, 2021a], lo cual, como ha sido explicado anteriormente, elimina la representaci n sint ctica y sem ntica que esta presente en el mismo, es por ello que los autores proponen extender *Neural ProgLDA* [Srivastava y Sutton, 2017], incluyendo las representaciones vectoriales producidas por un modelo de *sentence embedding*.

ProgLDA es un modelo de t picos basado en LDA, que reemplaza la Multinomial que permite realizar el muestreo de las palabras a usar en la generaci n de los documentos por un producto de expertos (el cual corresponde a una promedio geom trico de los par metros que describen a estas matrices multinomiales de palabras), esta decisi n se basa en el hecho de que las predicciones hechas por las distintas multinomiales utilizadas, no ser n m s precisas que las componentes que son combinadas [Hinton y Salakhutdinov, 2009]. Por otro lado *progLDA* propone un nuevo m todo de inferencia *AVITM*, el cual mediante el uso de *Auto Encoders* busca llevar la bolsa de palabras a una representaci n que siga una

distribución *Gaussiana* para que mediante el uso de la inferencia variacional se aproximen las variables latentes que representen de mejor manera los tópicos a modelar; en este punto los autores [Bianchi *et al.*, 2021a] proponen una mejora; añadir los *embeddings* de modelos como *SBERT* en el proceso de reconstrucción de las *BoW* de *AVITM* para incorporar la información semántica de los textos en la estimación de los parámetros que describirán la distribución de tópicos a modelar.

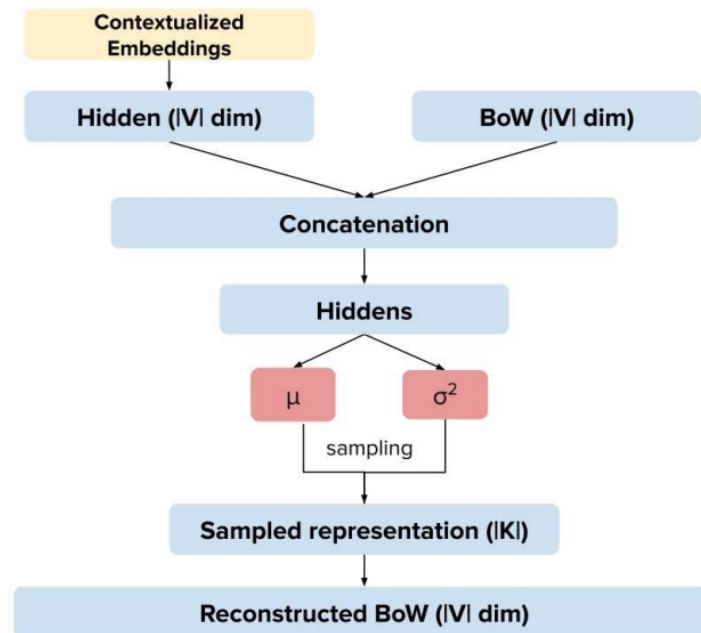


Figura 9: Boceto de *CombinedTM*
Fuente: [Bianchi *et al.*, 2021a]

Una variación importante de este modelo a mencionar es *Zero-Shot Topic Model* el cual fue presentado por los mismos autores en [Bianchi *et al.*, 2021b], estos destacan como principal ventaja de este modelo, la posibilidad de realizar modelación de tópicos en idiomas en los cuales no fue entrenado inicialmente el modelo (denominándose por ende *Zero-Shot*), dada la necesidad del modelo de predecir tópicos en palabras fuera del vocabulario original, se elimina la necesidad del *BoW* como *input* del modelo, sin embargo se mantiene como *output* auxiliar para la obtención de las palabras más representativas de cada tópicos.

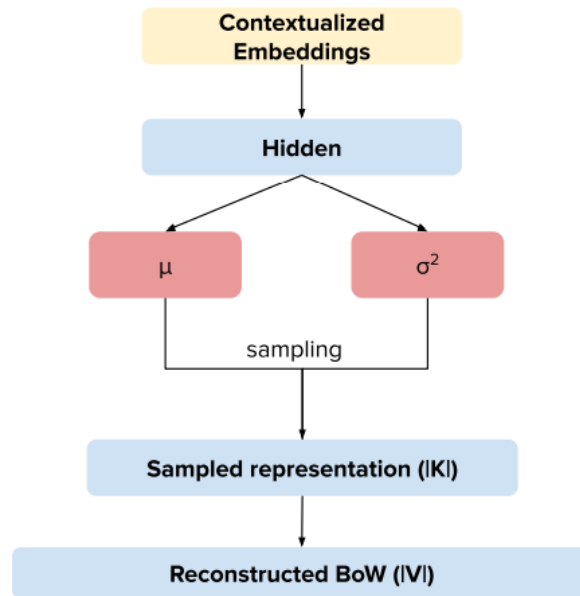


Figura 10: Boceto de Zero-Shot Cross-lingual Contextualized Topic Model
Fuente: [Bianchi *et al.*, 2021b]

3.5. BERTopic

Este modelo permite descubrir tópicos latentes en una colección de documentos extendiendo la idea de que la modelación de tópicos es una tarea de *clusterización* de documentos.

BERTopic surge como una variación del modelo propuesto por [Sia *et al.*, 2020], al igual que este, *BERTopic* emplea representaciones vectoriales provistas por modelos pre-entrenados basados en *transformers* (como lo son la familia de modelos *SBERT* [Reimers y Gurevych, 2019a]) con el fin de realizar *clusterización* sobre los mismos utilizando modelos no supervisados como *HDBSCAN*, sin embargo a diferencia de este, los autores [Grootendorst, 2022] proponen un método distinto para dar interpretabilidad y entendimiento a los tópicos modelados, estos destacan que el método empleado por [Sia *et al.*, 2020] para dar la interpretabilidad a los tópicos modelados puede ser un método engañoso en ciertos casos dado que la formación de los tópicos es llevada a cabo mediante técnicas de *clusterización* por densidad, mientras que las palabras elegidas para representar a los mismos son seleccionadas por cercanía a los centroides de los mismos, esto supone un problema según [Grootendorst, 2022] dado que no se estaría tomando en cuenta la forma del *cluster* modelado. Como solución a este problema, los autores proponen una variación de *TF-IDF* dado que esta técnica es agnóstica a la forma de los mismos, dependiendo únicamente de la frecuencia de las palabras.

Una de las principales características de este modelo es la flexibilidad, dado que el funcionamiento de sus distintas componentes es independiente, esto implica que sus distintas piezas pueden ser fácilmente reemplazadas. A continuación se muestra una representación gráfica del modelo hecha por los autores de *BERTopic*:

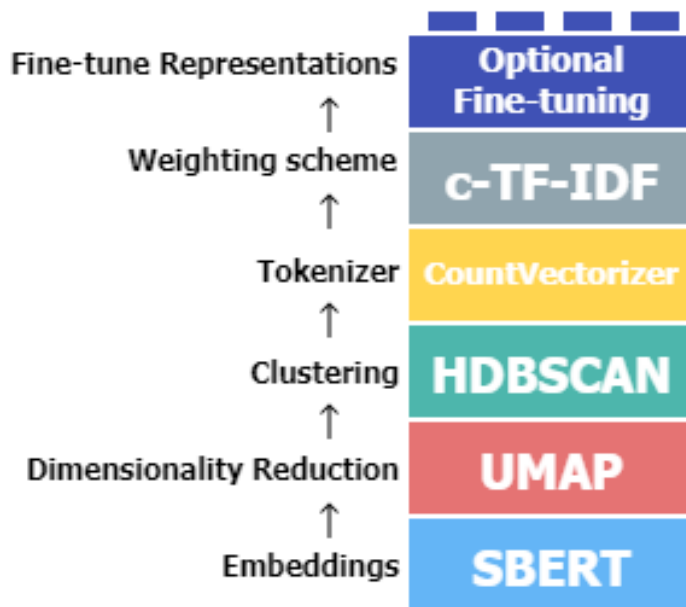


Figura 11: BERTopic's main algorithm
Fuente: Maarten Grootendorst

Los autores refuerzan la idea de independencia e intercambiabilidad entre los distintos componentes que conforman el modelo mediante el uso de estas figuras de construcción, a partir de la figura es posible identificar los siguientes bloques:

- **SBERT**: Pieza inicial que permite transformar oraciones completas en *embeddings*, esta técnica está basada en *transformers* y es una variación de *BERT* propuesta por [Reimers y Gurevych, 2019b], esta hace uso de arquitecturas basadas en redes siamesas que toman como base los modelos *BERT* tradicionales, combinando de esta manera los *embeddings* por palabra mediante técnicas de *Pooling*, finalmente el modelo es capaz de aprender *embeddings* similares a nivel de oración (en cuanto respecta a la semántica y contenido de las mismas oraciones) pasando por una capa que calcula la distancia entre las distintas oraciones procesadas. Este tipo de modelos ha demostrado grandes avances en el estado del arte en múltiples tareas, los autores [Reimers y Gurevych, 2019b] destacan su alta eficiencia para la tarea *semantic textual similarity* (STS), donde para la obtención de pares de oraciones similares dentro de un conjunto de 10.000 oraciones, *BERT* tarda alrededor de 65 horas, mientras que con *SBERT* la tarea se reduce a solamente 5 segundos.

- **UMAP:** *Uniform Manifold Approximation and Projection* es una técnica que al igual que *PCA* o *tSNE* permite realizar reducción de dimensionalidad sobre los vectores de características que describen a los datos. En [Coenen y Pearce,] se explica que este algoritmo puede ser abstraído en 2 grandes pasos, el primero corresponde a la construcción de un grafo que representa la similaridad en función de las distancias entre los distintos puntos en el espacio original, donde a partir de esto se da inicio a una segunda etapa donde a través de un proceso de optimización iterativo se construye la representación vectorial de menores dimensiones, durante esta etapa se seleccionan puntos al azar donde de manera similar al algoritmo *tSNE* se procede a modificar la posición de los puntos en el plano de menores dimensiones, de tal manera que la distancia entre puntos cercanos en el grafo construido en la primera etapa se minimice y la distancia entre puntos lejanos (usando una *t-distribution* como función de optimización) se maximice.

BERTopic hace uso de esta técnica para reducir el impacto del problema de la maldición de la dimensionalidad sobre los algoritmos de *clustering*, los autores eligen esta técnica por sobre *PCA* o *t-SNE* dado que, *UMAP* ha mostrado ser más efectivo al momento de preservar las características tanto a nivel local como global tras hacer la proyección de datos a una menor dimensión [McInnes et al., 2020], además de escalar mejor ante un alto número de dimensiones en comparación a *tSNE*.

- **HDBSCAN:** *Hierarchical Density-Based Spatial Clustering of Applications with Noise* es un modelo jerárquico basado en agrupación por densidad que es robusto ante el ruido, este bloque de *BERTopic* puede ser reemplazado por cualquier modelo no supervisado de *clustering* como por ejemplo *KMeans*, *KNN*, *Fuzzy c-Means*, entre otros.

En [Berba, 2020] se explica que en *HDBSCAN* es posible identificar tres secciones:

- **Estimación de densidad:** *HDBSCAN* esta basado en la idea de que aquellas zonas de alta densidad conforman un *clusters*, es por ello que mediante los k vecinos más cercanos se calcula la mínima distancia ε (también conocida como *core-distance* o $core_k(x)$) que permite contener los k puntos más cercanos, esto posibilita modelar una función de densidad de probabilidad donde aquellas regiones más dispersas (conformadas por *outliers*) poseerán una baja *PDF*, mientras que aquellas regiones más densas (candidatas a un *cluster*) poseerán una alta *PDF*, esta idea se formaliza mediante la matriz de alcance mutuo:

$$d_{mreach-k}(a, b) = \max(core_k(a), core_k(b), d(a, b)) \quad (11)$$

- **Selección jerárquica de regiones de alta densidad:** A partir de la matriz de alcanzabilidad mutua se construye un *minimum spanning tree* a través del algoritmo *Prim's*, este retorna un grafo (tipo árbol) donde los vértices se encuentran conectados de tal manera que se minimice (mediante un enfoque *greedy*) la suma total

de la distancia entre sus arcos (la cual se mide mediante la distancia de alcanzabilidad mutua). Este árbol representa una jerarquía en si dado que los arcos del mismo permiten observar una transición desde zonas menos densas hasta las zonas más densas.

- **Obtención de clusters:** Dado que el objetivo final es encontrar *clusters* significativos, se podan aquellas hojas del árbol jerárquico que no posean un tamaño mínimo como para ser considerados *clusters*, tras esto se procede a comparar la pérdida que representa la separación de un *cluster* para evitar seleccionar *sub-clusters* que no sean lo suficientemente significativos, de esta manera se preservan aquellos *clusters* que posean una alta estabilidad a través de los distintos niveles de la jerarquía. Se define la estabilidad del *cluster* como:

$$\sum_{p \in \text{cluster}} \lambda_p - \lambda_{\text{birth}} \quad (12)$$

Donde λ_{birth} corresponde al momento donde el punto comienza a pertenecer al *cluster* mientras que λ_p corresponde al momento en que el punto p deja de pertenecer al *cluster* y λ como $1/\text{core}_k(x)$. Los autores explican que este proceso puede ser entendido como la preservación de aquellos *clusters* con mayor área de densidad de probabilidad a distintos niveles de distancia de alcance mutua en el árbol.

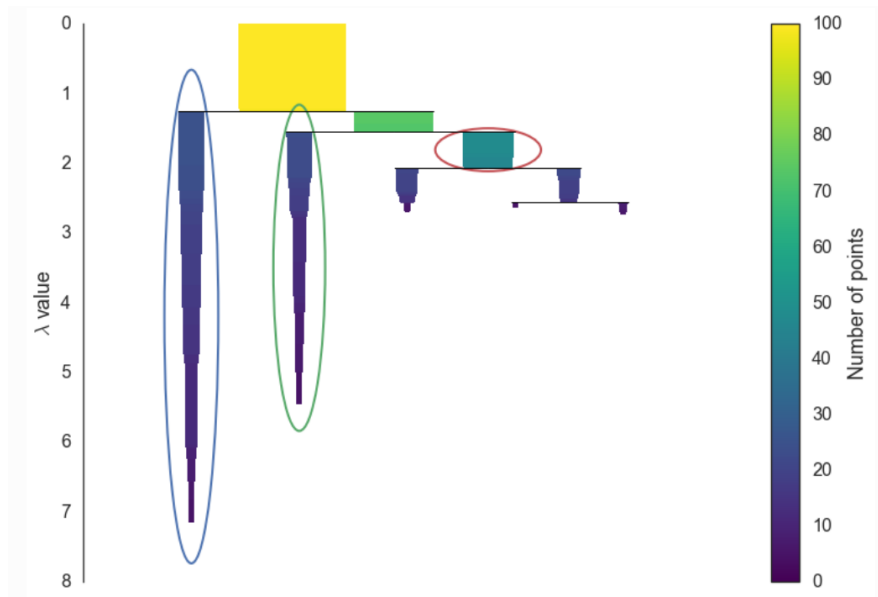


Figura 12: Condensed Tree

Fuente: https://hdbscan.readthedocs.io/en/latest/how_hdbscan_works.html.

Los autores de *BERTopic* optan por el uso de este modelo dado que la naturaleza jerárquica de este enfoque, posibilita realizar detección de tópicos que engloben otros subtópicos, facilitando el análisis de los resultados tanto desde una vista macro como

detallada, por otra parte este algoritmo es robusto ante el ruido, lo cual previene la asignación errónea de documentos a tópicos poco relacionados (preservando la coherencia de los tópicos modelos).

- **CountVectorizer:** Tanto este bloque como el siguiente son piezas fundamentales para dar interpretabilidad a los resultados. Dado que *HDBSCAN* es una técnica que forma *clusters* con mayores o menores grados de densidad y con distintas formas geométricas, los autores indican que asumir una perspectiva basada en centroides no es lo más apropiado, es por ello que proponen utilizar *BoW* a nivel de *clusters* para eliminar cualquier suposición previa respecto a la estructura de los *clusters*, cabe destacar que esta bolsa de palabras es normalizada con la norma *L1* dado que los *clusters* pueden tener distintos tamaños.
- **c-TF-IDF:** Este paso utiliza la representación vectorial del paso anterior para obtener las palabras más representativas a nivel de tópico, esto es realizado con el fin de obtener un puntaje asociado a cada palabra dentro de un tópico, de esta manera aquellas palabras con mayor puntaje serán candidatas para dar una interpretación al tópico modelado, los autores definen *c-TF-IDF* como:

$$W_{x,c} = ||tf_{x,c}|| \log(1 + \frac{A}{f_x}) \quad (13)$$

$tf_{x,c}$ = Frecuencia de la palabra x en cluster c

f_x = Frecuencia de la palabra x en todos los clusters

A = Promedio de palabras por cluster

x = termino x

c = cluster c

CAPÍTULO 4

PROPUESTA DE SOLUCIÓN

La fase de propuesta de solución es un proceso el cual busca responder ¿Cómo se llevarán a cabo los objetivos propuestos en un inicio 1.4?, para ello es importante detectar las principales dificultades a abordar en conjunto con las principales dolencias o necesidades del negocio, de esta manera será posible tanto proponer como elegir una solución a partir de múltiples soluciones candidatas, de tal manera que esta sea la que mejor se adecue a los problemas a enfrentar. Ante esto se establecen diversas características y/o principales problemas en función al caso estudiado.

4.1. Contexto y características de los datos

1. Existe un alto volumen de encuestas contestadas, sin embargo el recibimiento de estas por cada día es baja, es por ello que el análisis y entendimiento de las encuestas es realizado tras cada cuatrimestre por el equipo de *NPS*.
2. Existe una presencia de comentarios tanto en Español como Portugués, con distinta lengua, norma y habla sujetos a cada encuestado y su contexto.
3. Los encuestados no suelen explyar sus ideas en textos largos, por lo que es importante dejar fuera del análisis encuestas de pocas palabras.
4. El uso del lenguaje empleado por los encuestados no es de carácter formal, lo cual se traduce en ruido para los modelos a usar.
5. Se dispone de tres tipos de clientes distintos los cuales estan condicionados por el uso e interacción con la plataforma y sus servicios. Los comentarios recibidos en las encuestas provienen desde distintos países, verticales de negocio y categorías distintas de clientes (compradores, vendedores y vendedores profesionales), ante esto es importante definir y estandarizar un método que sea capaz de abordar o extrapolarse a todos los involucrados.

En función del contexto planteado, es importante a su vez responder ¿Cuáles son los deseables que se esperaría conseguir?, ¿Qué puntos pueden ser dejados de lado y cuáles no?, ante esto se define una lista de deseables en conjunto con el equipo de *NPS* para guiar la selección de un modelo como candidato a solución.

4.2. Deseables de la solución a proponer

1. Se permiten soluciones que no sean las más rápidas posibles, priorizando así la calidad por sobre la eficiencia, es por ello que el tiempo de ejecución de la solución a proponer no será relevante para el análisis.
2. La existencia de distintos lenguajes debe ser abordada, pero no es considerado necesario que la solución a utilizar sea capaz de manejar más de un idioma a la vez. Si bien este deseable por parte del negocio da espacio y holgura para explorar diversas soluciones, la implicancia de tener un modelo por cada lenguaje se traduce en mayores esfuerzos tanto a la hora de entrenar, validar y mantener a largo plazo el funcionamiento correcto de las distintas soluciones propuestas.
3. Se permite no tomar en consideración durante el análisis aquellas respuestas de los encuestados que sean demasiado cortas, esta consideración es empleada actualmente por el equipo NPS, su motivo se fundamenta ante el hecho de que la asignación de un tema para los comentarios extremadamente cortos representan un problema incluso para un ser humano, entre menor sea el contexto u elaboración de ideas que posea una opinión se vuelve menos probable que esta posea algo útil para el negocio, ante esto la tarea de preprocesamiento debe ser lo suficientemente robusta como para deshacerse de estos elementos que representan más ruido que un verdadero aporte al análisis de las encuestas, sin embargo se considera como un deseable que el modelo a utilizar sea también a su vez robusto ante esta problemática, dado que no es posible delimitar una regla que sea capaz de eliminar a priori todas las encuestas que ante esta definición son consideradas como ruido.
4. Se considera como un deseable de alto valor, que la solución a utilizar sea lo suficientemente robustas ante la presencia de errores ortográficos. El método empleado por el equipo de NPS no toma en consideración este factor, por lo que abordar este problema ya sea mediante técnicas de preprocesamiento o a través de modelos que sean inherentemente robustos ante esto, representa un alto valor para el negocio.
5. Se da la holgura de poseer modelos distintos en base a los distintos grupos de clientes que se poseen, ante esto se opta por evaluar la calidad de los modelos a proponer frente a tres conjuntos de datos por separado (vendedores, profesionales y compradores), se deja como trabajo a futuro evaluar la necesidad de especializar los modelos de cada grupo de clientes en base a las distintas verticales (vehículos e inmobiliarias) o por países, se opta por esta decisión dado que un análisis más detallado puede ser manejado durante esta etapa con filtros simples inherentes a la data que permiten limitar la muestra de datos post detección de tópicos, a su vez se suma el hecho de que los temas de trasfondo no son necesariamente excluyentes o únicamente propios por verticales o países.

Ante los distintos objetivos, deseables y problemas a afrontar que han sido explicados, se decide evaluar la calidad de diversos modelos en la tarea de detección de tópicos, a su vez

se propone el uso de *ChatGPT* como generador de resúmenes para otorgar una interpretación semántica de manera automática a los tópicos detectados, finalmente se presentan los resultados al negocio mediante *dashboards* generados en Tableau. En función de todo lo explicado, se establecen los siguientes pasos:

1. Análisis exploratorio de los datos.
2. Preprocesamiento de la data.
3. Hiperparametrización de modelos candidatos.
4. Selección del modelo a utilizar en base a métricas.
5. Puesta en producción del modelo.
6. Análisis post exploratorio por tópico mediante ChatGPT.
7. Disponibilización de los resultados a través de *dashboards* hechos en Tableau.

4.3. Análisis exploratorio de la data

Durante esta etapa se busca analizar y comprender los datos disponibles, evidenciando así el nivel de ruido y el largo de las respuestas que proveen los encuestados, a su vez se realizará un preprocesamiento de los datos con el fin de disminuir la presencia de ruido en los mismos, por último se realizara una exploración visual de los datos mediante el uso de *word-clouds*, permitiendo de esta manera evidenciar aquellas palabras más frecuentes en los textos, lo cual en futuras iteraciones puede ser utilizado para la extensión del diccionario de *stopwords* o como un elemento que refuerza la idea de que para la obtención de información útil en base al contenido de estas encuestas es necesario llevar a cabo análisis más sofisticados como lo son el uso de los modelos de *machine learning*.

Los datos recopilados abarcan un periodo entre 2019-07-01 y 2023-03-17, sin embargo no todos los encuestados optan por responder estas encuestas, a continuación se muestran datos preliminares sobre estas encuestas para corroborar lo mencionado:

Tabla 1: Encuestas 2022-2023 Mercado Libre VIS
Fuente: Elaboración Propia.

Encuestas NPS			
	Vendedores	Compradores	Profesionales
Total	63199	58392	24900
Con respuesta	24939	44723	16119
Porcentaje	39.5 %	76.6 %	64.7 %

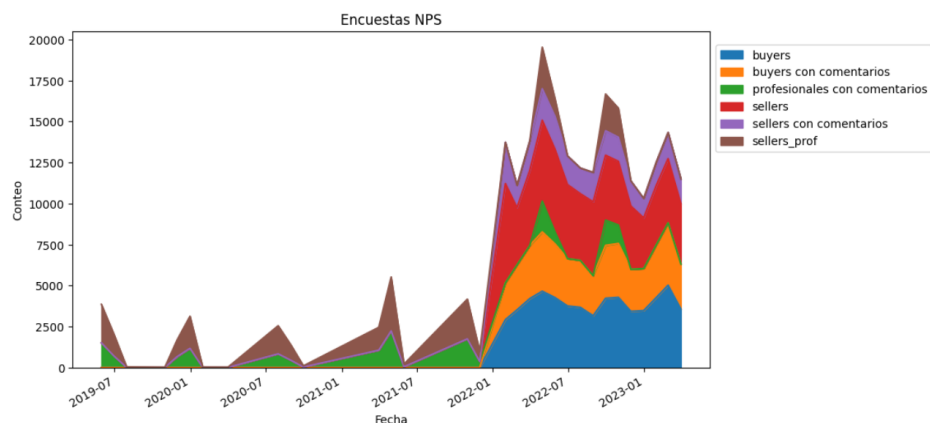


Figura 13: Serie de tiempo de encuestas NPS respondidas
Fuente: Elaboración propia.

Tabla 2: Palabras por encuestas pre-procesamiento.

Fuente: Elaboración propia.

Profesionales								
País	Palabras	Prom.	D.E	Min	25 %	50 %	75 %	Max
Argentina	4318	26.60	35.63	1	5	16	36	630
Brasil	1945	23.49	29.70	0	5	14	31	377
Chile	4325	33.93	35.36	0	11	23	45	566
Colombia	1025	23.26	24.04	1	7	16	30	185
Mexico	2413	26.47	31.04	0	6	17	36	391
Uruguay	793	35.58	39.85	0	12	24	46	441
Venezuela	1273	36.28	39.74	1	11	24	49	506

Tabla 3: Palabras por encuestas pre-procesamiento.

Fuente: Elaboración propia.

Vendedores								
País	Palabras	Prom.	D.E	Min	25 %	50 %	75 %	Max
Argentina	10996	14.28	13.27	0	3	10	21	55
Brasil	4180	14.46	12.69	0	4	10	21	55
Chile	1876	17.95	13.62	1	6	14	28	51
Colombia	3716	12.02	12.37	0	2	7	18	53
Mexico	4171	12.24	12.35	0	2	8	18	53

Tabla 4: Palabras por encuestas pre-procesamiento.

Fuente: Elaboración propia.

Compradores								
País	Palabras	Prom.	D.E	Min	25 %	50 %	75 %	Max
Argentina	13248	15.43	12.45	0	6	12	22	58
Brasil	8542	12.76	11.27	0	4	9	17	51
Chile	7285	14.75	11.79	0	6	11	20	90
Colombia	5209	13.62	11.77	0	4	10	19	52
Mexico	9259	14.38	11.62	0	5	11	20	55
Uruguay	915	15.30	12.22	1	6	12	22	52
Venezuela	265	13.79	10.89	1	5	11	19	46

Mediante estos datos es posible evidenciar que por lo general el largo de las respuestas recibidas a través de las encuestas realizadas tanto a vendedores, compradores como vendedores profesionales corresponden a respuestas cortas, esta falta en la elaboración de las respuestas de los encuestados abre la posibilidad de que los modelos basados en *embeddings* no logren reflejar los resultados esperados en comparación a sus contrapartes basados únicamente en *Bag Of Words*, esto se justificaría dado que los modelos basados en *embeddings* dependen en cierta medida del contexto para un entendimiento e interpretación de la semántica de las palabras, sin embargo es importante remarcar que esto es solo una hipótesis que ha de ser corroborada durante el análisis de los resultados de la experimentación.

Con el fin de eliminar aquellas respuestas poco elaboradas y eliminar el ruido presente en las mismas, se utilizan diversas técnicas para la limpieza de los datos.

4.4. Preprocesamiento de la data

Durante esta fase se establece y aplica el siguiente flujo de trabajo:

1. **Normalización del texto:** Se transforma a minúsculas todo el texto para reducir la variabilidad de este, facilitando de esta manera la co-ocurrencia de las palabras.
2. **Eliminación de números:** Dado que lo que se busca es modelar tópicos, la presencia de números es considerada como ruido e información irrelevante para la modelación de los mismos.
3. **Remoción de palabras de longitud 1:** Estas se eliminan, ya que suelen ser conectores o palabras mal escritas que no suponen más que ruido para los modelos.
4. **Eliminación encuestas muy cortas:** Se decide colocar un *threshold* de 5 palabras como mínimo; dado que lo que se busca es generar *insights* para el negocio, hacer uso de

comentarios muy poco elaborados no es una buena práctica para la toma de decisiones.

5. **Eliminación de stopwords:** Se remueven las palabras consideradas frecuentes y poco relevantes en los textos mediante el uso de la librería *NLTK*, dado que esta permite manejar tanto español como portugués.
6. **Lemmatización:** Se llevan las palabras a su forma morfológica mediante la librería *StanzaNLP*, dado que permite trabajar tanto el español como el por portugués.

Durante esta etapa se evaluó la posibilidad de arreglar errores ortográficos y la eliminación de tildes, sin embargo se optó por no llevar a cabo la corrección de fallas ortográficas, debido a la lentitud de su aplicación, se suma a su vez el hecho de que se espera que un buen modelo de detección de tópicos sea lo suficientemente robusto como para manejar este problema, por su parte no se realizó la eliminación de tildes, ya que, hacer esto modifica el significado de algunas palabras, tanto en el español como en el portugués.

Tras aplicar este preprocesamiento, los datos presentan esta nueva distribución:

Tabla 5: Palabras por encuestas post-procesamiento.

Fuente: Elaboración propia.

Profesionales								
País	Palabras	Prom.	D.E	Min	25 %	50 %	75 %	Max
Argentina	3295	34.36	37.55	5	13	24	43	630
Brasil	1424	31.34	31.21	5	12	22	40	377
Chile	3967	36.73	35.61	5	14	26	48	566
Colombia	853	27.38	24.35	5	11	19	35	185
Mexico	1909	32.95	31.89	5	12	22	43	391
Uruguay	728	38.51	40.31	5	14	27	48	441
Venezuela	1166	39.34	40.15	5	14	27	53	506

Tabla 6: Palabras por encuestas post-procesamiento.

Fuente: Elaboración propia.

Vendedores								
País	Palabras	Prom.	D.E	Min	25 %	50 %	75 %	Max
Argentina	7495	19.93	12.55	5	10	16	28	55
Brasil	3022	19.07	12.06	5	9	15	26	55
Chile	1547	21.26	12.73	5	10	18	32	51
Colombia	2274	18.42	12.00	5	9	14	25	53
Mexico	2588	18.39	12.00	5	9	15	25	53

Tabla 7: Palabras por encuestas post-procesamiento.

Fuente: Elaboración propia.

Compradores								
País	Palabras	Prom.	D.E	Min	25 %	50 %	75 %	Max
Argentina	10542	18.71	11.92	5	9	15	26	58
Brasil	6087	16.77	11.03	5	8	13	22	51
Chile	5928	17.47	11.45	5	9	14	23	90
Colombia	3787	17.85	11.16	5	9	15	24	52
Mexico	7307	17.47	11.20	5	9	14	23	55
Uruguay	725	18.59	11.65	5	9	15	25	52
Venezuela	207	16.90	11.35	5	9	14	22	46

Es posible observar que si bien el número total de respuestas recibidas en las encuestas se ha reducido considerablemente (~10 mil respuestas menos aproximadamente por cada conjunto de datos), el largo de las respuestas aumento a 15-20 palabras por respuesta aproximadamente, las cuales a su vez no son *stopwords*, por lo que es esperable que estas (al menos un gran porcentaje) expresen alguna idea o contenido interesante para el negocio.

Tras haber realizado este preprocesamiento de los datos, se realizaron nubes de palabras por cada país y vertical del negocio a analizar (separando los datos entre promotores y detractores para facilitar la visualización de patrones presentes en la data).



Figura 14: Comentarios hechos por compradores promotores en las verticales de Vehículos e Inmuebles

Fuente: Elaboración propia.

Adicionalmente se llevó a cabo la realización de esta visual con el resto de las muestras disponibles, a partir de estas no es posible detectar patrones subyacentes evidentes, sin embargo se destaca su utilidad ante la posibilidad de evidenciar la existencia de palabras que a priori, pueden ser candidatas a tópicos de interés para el negocio, destacando así palabras como: precio, variedad, filtro, plataforma, búsqueda, seguridad y confianza, por otro lado se descarta la relevancia de palabras muy frecuentes como lo son vehículo e inmueble, dado que los datos de estudio recopilados forman parte de un contexto que se encuentra condicionado al área de venta de vehículos e inmuebles en Mercado Libre,

convirtiéndolas de tal manera en palabras inherentes al contexto que no son relevantes para un mejor entendimiento del negocio.

Tras la realización del preprocesamiento y estandarización de los datos se procede a la selección y entrenamiento de los distintos modelos.

4.5. Hiperparametrización de los modelos

Dado que tanto la selección de parámetros por cada modelo y la elección del número de tópicos a modelar corresponden a variables no conocidas de antemano, se vuelve necesario probar distintas configuraciones de parámetros que permitan obtener los mejores resultados posibles, sin embargo es importante destacar que durante la presente tesis de pregrado se optó por no llevar a cabo una búsqueda rigurosa y detallada de aquellos hiper parámetros que conforman a cada modelo, la justificación de esta decisión radica ante el hecho de que el espacio de búsqueda es demasiado amplio (se debe entrenar 7 modelos en 3 muestras de datos distintas), lo cual se traduce en un consumo de tiempo demasiado elevado. Con el fin de reducir los tiempos de computo se limita el reentrenamiento de cada modelo modificando únicamente el número de tópicos, por consiguiente se define el siguiente espacio de búsqueda: $K \in (40, 50, 60, 70, 80, 90)$ con K como el número de tópicos, este rango de valores fue elegido dado que tras conversaciones con los *skateholders* se determinó que era deseable tener un número de tópicos que estuviera entre 50 y 100 con el fin de favorecer y apuntar a análisis detallados que a su vez permitan levantar accionables de manera directa.

Es importante destacar que el motivo de esta etapa es permitir realizar una comparación que guíe la selección de un modelo entre los distintos candidatos, es por ello que la especialización del modelo candidato a solución puede ser un trabajo realizado tras su selección en caso de que los resultados obtenidos por este no reflejen un buen rendimiento en la práctica.

4.6. Selección del modelo

Con el fin de guiar la selección del mejor modelo entre los distintos candidatos se calcularon las métricas *UMass*, *IRBO* y *TU* dado que permiten evaluar tanto la coherencia como la diversidad de los tópicos generados por cada modelo. Se opta por el uso de estas métricas y no otras que si toman en consideración la semántica (como por ejemplo pair-wise word embedding topic coherence o centroid word embedding topic coherence propuestas por el autor [Ding et al., 2018]) dado que no todos los modelos probados hacen uso de esta característica.

- **Topic Uniqueness:** Los autores [Nan et al., 2019] proponen esta métrica con el fin de

medir el número de palabras únicas que conforman a los tópicos modelados, cuantificando de esta manera la diversidad de los tópicos. Este valor varia entre 0 y 1, donde 0 indica una mala diversidad, mientras que a su vez 1 representa una gran diversidad entre los tópicos modelados.

$$TU = PU / (topk * K) \quad (14)$$

Donde se define:

- PU: Número de palabras únicas dentro del top k de palabras que definen a cada tópico.
 - topk: Número de palabras con mayor importancia para describir un tópico.
 - K: Número de tópicos.
- **Inverted Rank Biased Overlap:** Los autores [Bianchi *et al.*, 2021a] proponen el uso de esta métrica para medir la diversidad de los tópicos modelados dado que permite comparar la similaridad entre el contenido de dos conjuntos, los cuales para el caso de estudio corresponderían a las palabras que describen un tópico.

$$InvertedRBO = 1 - RBO$$

$$RBO(S, T, p) = (1 - p) \sum_{d=1}^{\infty} p^{d-1} \cdot A_d \quad (15)$$

Donde se define:

- S: Lista/ranking de elementos infinitos.
 - T: Lista/ranking de elementos infinitos.
 - p: Parámetro que controla la ponderación de la métrica. Cuando este es cero, solo el elemento mejor rankeado es considerado (por lo que el RBO tomara valor 0 o 1 siempre), por otro lado entre más se acerca de 1 se encuentre, los pesos de los elementos se vuelven más planos y por ende la evaluación considera cada vez mas elementos.
 - A_d : Se define como el tamaño de la intersección entre S y T dividido por la profundidad de elementos considerados.
- **UMass Score:** Métrica propuesta por [Mimno *et al.*, 2011], los autores explican que esta métrica permite medir que tanto, dentro de las palabras usadas para describir un tópico, una palabra común es en promedio un buen predictor para una palabra menos común, permitiendo de esta manera medir la coherencia de los tópicos modelados.

$$UMass(T) = \sum_{m=2}^M \sum_{l=1}^{m-1} \log \left(\frac{p(w_m, w_l) + \frac{1}{D}}{p(w_l)} \right) \quad (16)$$

Donde se define:

- $T = \langle w_1, \dots, w_n \rangle$: Lista ordenada de palabras
- $p(w_m, w_l)$: Ratio de documentos que contienen ambas palabras w_m, w_l
- D : Largo del corpus.

Mediante el uso de estas métricas se entrenan 7 modelos pertenecientes a 4 familias de modelos distintos; dos modelos BERTopic que se diferencian únicamente en el método utilizado para la clusterización de los tópicos a modelar, dos modelos pertenecientes a la familia de los modelos *FLSA*, los modelos *Combined Topic Model* y *ZeroShoot Cross Lingual* pertenecientes a la familia de modelos *Contextualized Topic Models* y *LDA* como punto de referencia a superar por los otros modelos.

Tabla 8: Resultados obtenidos tras hiperparametrización.

Fuente: Elaboración propia.

Resultados sobre muestra de vendedores profesionales					
Modelo	Tópicos	Umass	iRBO	TU	Puntaje final
BERTopic HDBSCAN	40	-0.53	0.99	0.76	0.94
BERTopic KMeans	40	-0.20	0.94	0.51	0.86
FLSA W	40	-8.67	0.99	0.99	0.77
FLSA	40	-6.52	0.89	0.62	0.71
CTM	40	-7.99	0.97	0.66	0.70
ZS-CL	40	-5.67	0.93	0.40	0.69
LDA	40	-17.02	0.99	0.99	0.53

Tabla 9: Resultados obtenidos tras hiperparametrización.

Fuente: Elaboración propia.

Resultados sobre muestra de vendedores					
Modelo	Tópicos	Umass	iRBO	TU	Puntaje final
BERTopic HDBSCAN	40	-0.57	0.98	0.73	0.93
BERTopic KMeans	40	-0.25	0.95	0.49	0.86
CTM W	40	-6.83	0.97	0.59	0.72
ZS-CL	40	-5.75	0.95	0.39	0.69
FLSA	40	-6.42	0.75	0.36	0.62
FLSA W	40	-14.95	1.0	1.0	0.61
LDA	40	-18.10	0.99	0.99	0.52

Tabla 10: Resultados obtenidos tras hiperparametrización.
Fuente: Elaboración propia.

Resultados sobre muestra de compradores					
Modelo	Tópicos	Umass	iRBO	TU	Puntaje final
BERTopic HDBSCAN	40	-0.48	0.99	0.76	0.94
BERTopic KMeans	40	-0.18	0.96	0.52	0.87
CTM W	40	-6.88	0.98	0.64	0.73
ZS-CL	40	-5.54	0.96	0.42	0.71
LDA	40	-18.67	0.86	0.99	0.47

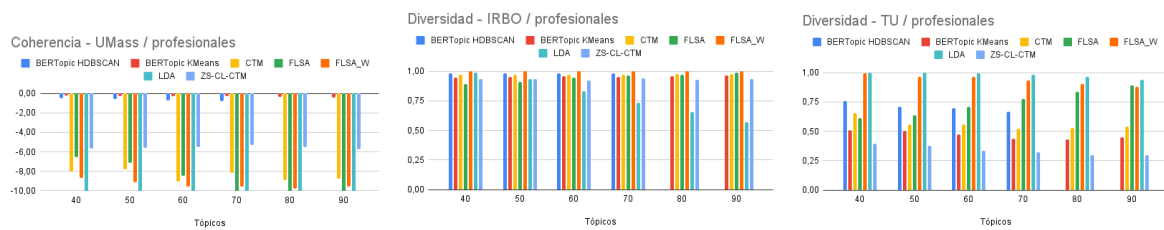


Figura 15: Resultados hiperparametrización de tópicos sobre profesionales
Fuente: Elaboración propia.

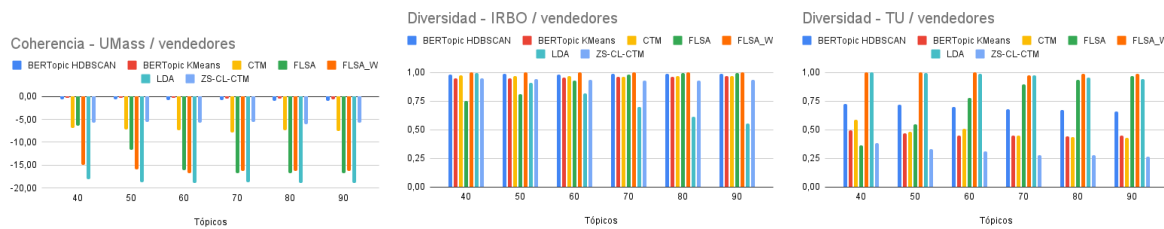


Figura 16: Resultados hiperparametrización de tópicos sobre vendedores
Fuente: Elaboración propia.

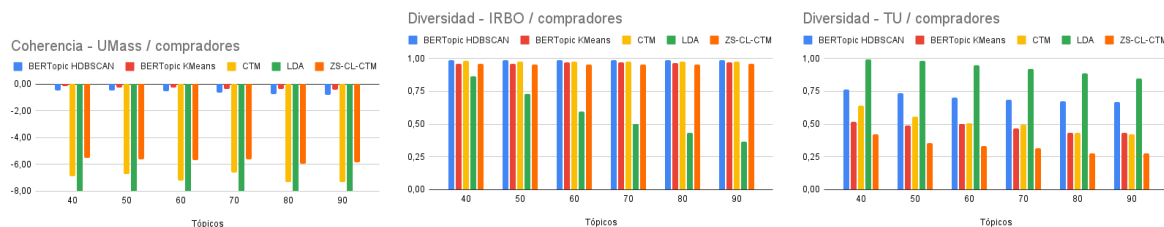


Figura 17: Resultados hiperparametrización de tópicos sobre compradores
Fuente: Elaboración propia.

En base a los resultados obtenidos es posible evidenciar que todos los modelos tienden a obtener mejores resultados al reducir el número de tópicos, esto es un indicativo de que ampliar el espacio de búsqueda a un menor número de tópicos puede ser una alternativa factible para mejorar los resultados obtenidos en un futuro en caso de que los resultados no reflejen lo esperado en la práctica. Por otro lado se destaca que los modelos basados en *BERTopic* presentan grandes diferencias en relación a la coherencia en comparación a otros modelos, obteniendo de tal manera los mejores resultados, sin embargo la diversidad de estos no sobresale en comparación al resto de modelos, esto en conjunto con la premisa de reducir el número de tópicos puede ser una alternativa a mejorar los resultados, dado que al hacer un análisis exploratorio sobre el resultado de los tópicos obtenidos es posible evidenciar que efectivamente (al menos para el caso de la muestra de profesionales) el tópico relacionado a problemas con los ejecutivos es un tema recurrente pero que es mencionado en conjunto a otras problemáticas, sin embargo es importante evidenciar que al realizar esto se puede llegar a desviar la atención de estos problemas secundarios que no necesariamente están relacionados entre sí, por lo que el análisis de su efectividad no debe pasar por alto. A su vez el uso de *HDBSCAN* mejora los resultados en comparación a *KMeans* (tanto en coherencia como diversidad), sin embargo no se opta por la puesta en producción de la versión que utiliza *HDBScan* debido a que este modelo considera como *outliers* alrededor del 60 % de los datos, los cuales tras una revisión manual se determino que si son relevantes para el negocio y por ende deben tomarse en consideración, ante esto se opta por utilizar la versión de *BERTopic* que utiliza *KMeans*, que si bien obtiene peores resultados en la diversidad de los tópicos modelados, no presenta grandes diferencias en relación a la coherencia. Por otro lado, es importante destacar que *LDA* logra presentar los mejores resultados en *TU* al momento de medir la diversidad de los modelos, sin embargo la coherencia de los mismos se aleja en gran medida de los resultados obtenidos por el resto de modelos (rondando alrededor de -18 siempre), convirtiéndolos en una alternativa infactible como propuesta de solución, como ya ha sido mencionado anteriormente este comportamiento era esperable dado la evidencia en la literatura que justifica estos resultados bajo la premisa de que estos modelos son dependientes de la coocurrencia de las palabras, lo cual es un elemento que se ve afectado ante la presencia de textos cortos. En relación con los resultados obtenidos por los modelos pertenecientes a la familia de modelos *Combined Topic Models* es posible observar que presentan mejores resultados en comparación a los modelos basados en *BoW* pero no logran acercarse a los resultados obtenidos por *BERTopic*, lo cual puede deberse a dos motivos, el primero corresponde al hecho de que estos modelos pueden manejar máximo un diccionario de 2000 palabras distintas, lo cual los lleva a podar palabras y por ende reducir en gran medida la cantidad de palabras que estos modelos toman en consideración a partir de la muestra original, por otro lado es posible que tras entrenar durante más épocas estos modelos se obtengan mejores resultados (durante esta etapa se opto por entrenar ambos modelos durante 10 épocas dado los largos tiempos de espera que implicaban extender el número de épocas). Finalmente si bien los modelos de la familia *FLSA* logran obtener buenos resultados tanto en la coherencia como diversidad en algunas ocasiones, sus resultados son variables en las distintas muestras, se destaca a su vez el problema del manejo de memoria que estos modelos presentan ante grandes volúmenes de datos (reflejándose en la incapacidad de

lograr su funcionamiento en la muestra de datos de compradores).

Tras la selección del modelo candidato se realiza la puesta en producción del mismo para la automatización del proceso.

4.6.1. Puesta en producción

Con el fin de determinar la frecuencia con la cual es requerido llevar a cabo la asignación de tópicos a las encuestas recibidas se evaluaron las necesidades del negocio, en base a reuniones con los *skateholders* se determinó que la solución a implementar debe ser capaz de cumplir las siguientes dos necesidades:

1. Asignar automáticamente cada 1 semana los tópicos a detectar sobre las respuestas recibidas.
2. Almacenar los datos en *BigQuery* para su posterior uso en *BI*.

Ante esto se implementa la solución haciendo uso de la herramienta *DataSuite*, la cual es una herramienta de automatización de procesos propia de la empresa que permite trabajar con distintas tecnologías de Base de datos, visualización, etc., esta automatización fue llevada a cabo mediante el uso de un *job* que se ejecuta todos los lunes de cada semana a las 6 AM, los pasos definidos dentro de este *job* son:

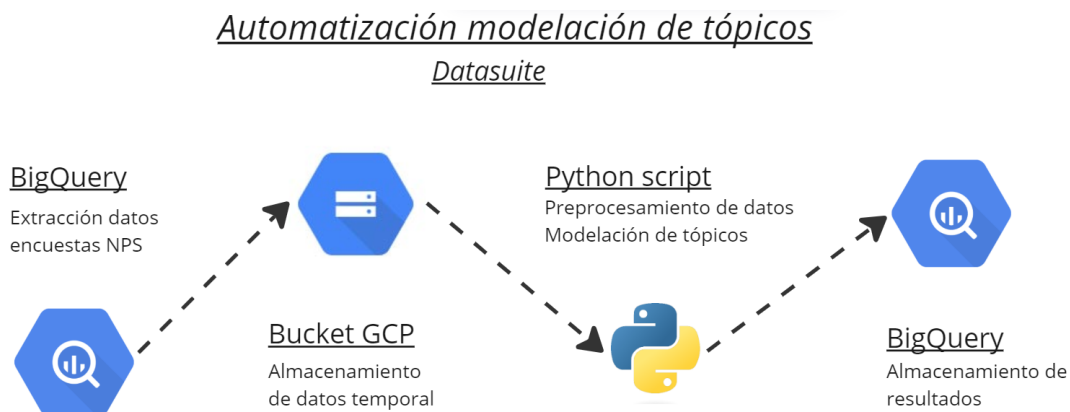


Figura 18: Pasos del *job* de automatización implementado en *DataSuite*
Fuente: Elaboración propia.

4.7. Interpretabilidad del tópico mediante ChatGPT

Con el fin de dar una interpretación semántica a los tópicos modelados y proveer de tal manera una vista resumida de las principales dolencias o bienestar que poseen los clientes de Mercado Libre VIS, se opta por utilizar la API de *ChatGPT*.

Para la tarea se utiliza la versión *GPT-3.5-turbo*, esta permite enviar textos con un máximo de 4096 tokens de largo, esta restricción es manejada mediante un submuestreo de los comentarios disponibles por cada tópico, dado que los comentarios son cortos esta restricción no supone gran problemática, a su vez se destaca la posibilidad de realizar análisis más finos que relajan en cierta medida el problema causado por esta restricción, segmentando de tal manera aún más los datos dependiendo de las verticales, países o rangos de fechas que son deseados en el análisis.

A continuación se presentaran el top 10 tópicos más frecuentes encontrados (para ver el resto revisar el anexo 6.9):

Tabla 11: Interpretación semántica de los resultados.

Fuente: Elaboración propia.

Tópico	c-TFIDF	Resumen hecho por ChatGPT
0	precio bajar costo paquete	Usuarios sugieren bajar precios, crear paquetes económicos, mejorar atención al cliente y resultados obtenidos, y ofrecer más promociones y descuentos. También se mencionan problemas con la plataforma y con la relación precio/calidad del servicio. Algunos usuarios sugieren mayor flexibilidad en los precios, más atención personalizada y mejores tiempos de respuesta a reportes de problemas. Otros comentan sobre la alta inflación en el país y la necesidad de adaptarse a las necesidades locales.
5	publicación precio paquete publicar	Se han recopilado opiniones sobre la plataforma, y muchas personas sugieren que se deben bajar los precios de las publicaciones, mejorar los paquetes y ofrecer promociones más atractivas. También se mencionan problemas en la plataforma para publicar desde el celular y la modificación de precios.
7	anuncio anuncio publicidad fazer	Algunos usuarios solicitan una mejora en la plataforma y en la atención al cliente, así como más promociones y publicidad para mejorar la visibilidad de los anuncios. También se pide una segmentación según el grupo objetivo y precios más accesibles para los corredores autónomos.
8	inmobiliario inmueble mercado propiedad	Sugerencias para mejorar la presencia y los resultados en el mercado inmobiliario de Mercado Libre, incluyendo promociones, mejoras en la plataforma y atención al cliente, así como precios y planes más accesibles para inmobiliarias pequeñas y la generación de más leads calificados. También se han mencionado dificultades en la plataforma y limitaciones en la búsqueda de propiedades, así como la necesidad de una mejor estrategia de marketing y la flexibilización de los precios y planes.

20	factura servicio pagar mes	Quejas sobre la mala atención al cliente y problemas en la facturación en Mercado Libre. También se mencionan problemas en la contratación de planes, renovación automática no deseada, falta de solución a problemas y eliminación de herramientas útiles para los clientes.
32	cliente atención mejorar personalizar	Mejor atención al cliente, personalizada y rápida. Mayor exposición para potenciales clientes y opciones de paquetes más convenientes.
34	correo telefono contacto email	Problemas de comunicación y atención al cliente en la plataforma. Pedidos de atención telefónica, mejoras en la verificación de datos y actualización de publicaciones, entre otros. Algunos usuarios reportan falta de resultados en las ventas a través de la plataforma.
10	pago pagar factura tarjeta	Clientes insatisfechos por problemas de facturación, cobros indebidos y mal servicio postventa. Sugieren mejorar canales de pago y atención al cliente.
13	publicación publicar plataforma mejorar	Solicitudes de mejoras en la plataforma de publicación, mayor visibilidad y posicionamiento de las publicaciones, atención al cliente más rápida y eficiente, posibilidad de modificación de publicaciones e implementación de estadísticas. También se mencionan problemas técnicos y de comunicación, así como la necesidad de una mayor variedad de productos en el catálogo.
18	foto subir video publicación	Sugieren mejorar la calidad y cantidad de imágenes de las propiedades publicadas, permitir la subida de videos y tours virtuales, simplificar el proceso de carga de fotos y mejorar la plataforma en general para una mejor experiencia del usuario. También piden una mayor inversión en tecnología para mostrar los inmuebles y la posibilidad de guardar propiedades en un banner para futuras publicaciones.

Es posible observar en estos resúmenes la presencia de temas en común, este solapamiento de ideas se refleja en las métricas de diversidad vistas anteriormente, por otro lado este hecho puede ser justificado dado que los encuestados no siempre hacen referencia a un único tema por comentario, son estas ideas que se anexan a una idea principal las que terminan separando el tópico, un ejemplo de ello son los tópicos 20 y 10, donde si bien ambos denotan un problema con la facturación, el tópico 20 apunta a otros problemas no necesariamente relacionados a la facturación, como lo son la contratación de planes por ejemplo, mientras que el tópico 10 aborda el tema dando explicaciones de porque se reportan problemas con la facturación. Por otro lado es importante destacar que el tema del servicio al cliente es muy reiterado, siendo mencionado en la mayoría de los tópicos, esto puede ser justificado dado que cuando los clientes presentan un problema relacionado ya sea al mal funcionamiento de la página, las factura, las publicaciones o cualquier otro tema, estos son reportados con los ejecutivos y si su atención al cliente no es la esperada, se vuelve normal que se reporte tanto el problema raíz como la mala atención relacionad al servicio del cliente.

4.8. Propuesta de dashboards para BI mediante Tableau

Con el fin de disponibilizar los datos no solamente mediante *BigQuery*, se proveen los resultados obtenidos mediante visualizaciones con la herramienta *Tableau*, buscando de esta manera facilitar el consumo de los datos para los distintos analistas.

La automatización de estas visualizaciones sigue un flujo de trabajo similar al realizado para la automatización del modelo de detección de tópicos, este proceso puede ser descrito en tres pasos: Extracción de la información mediante *BigQuery*, almacenamiento en un *Bucket* de GCP y generación de un extracto de los datos para *Tableau*.

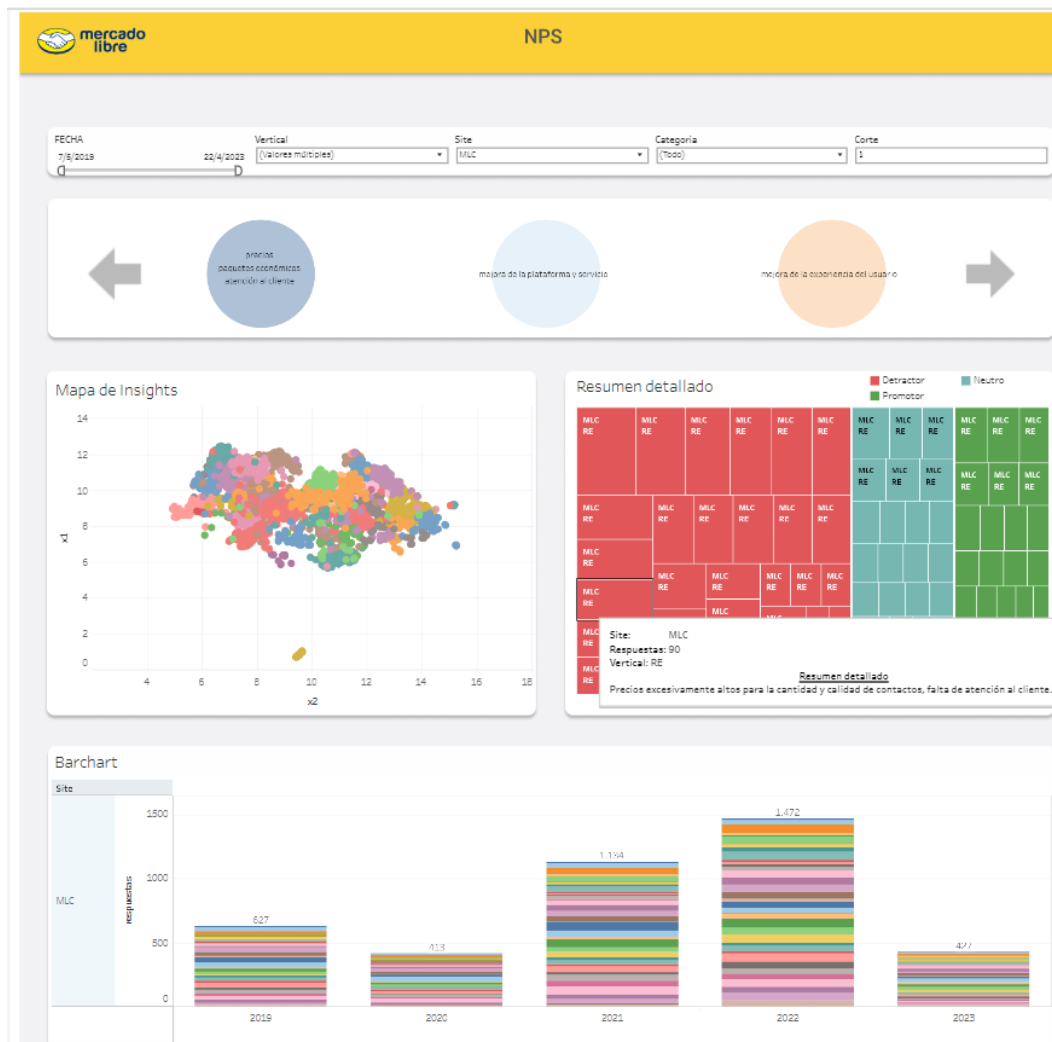


Figura 19: Dashboard estratégico
Fuente: Elaboración propia.

Este *dashboard* es propuesto como una prueba de concepto realizada como propuesta de

valor agregado para guiar la toma de decisiones basada en los datos, en esta se disponibilizaron filtros que permitan reducir la cantidad de datos visibles en pantalla, consolidando a su vez la información por país, vertical del negocio o fecha de análisis deseada, todas las visuales son interactivas y permiten leer el detalle de su significado (tópico, descripción breve y detallada del contenido).

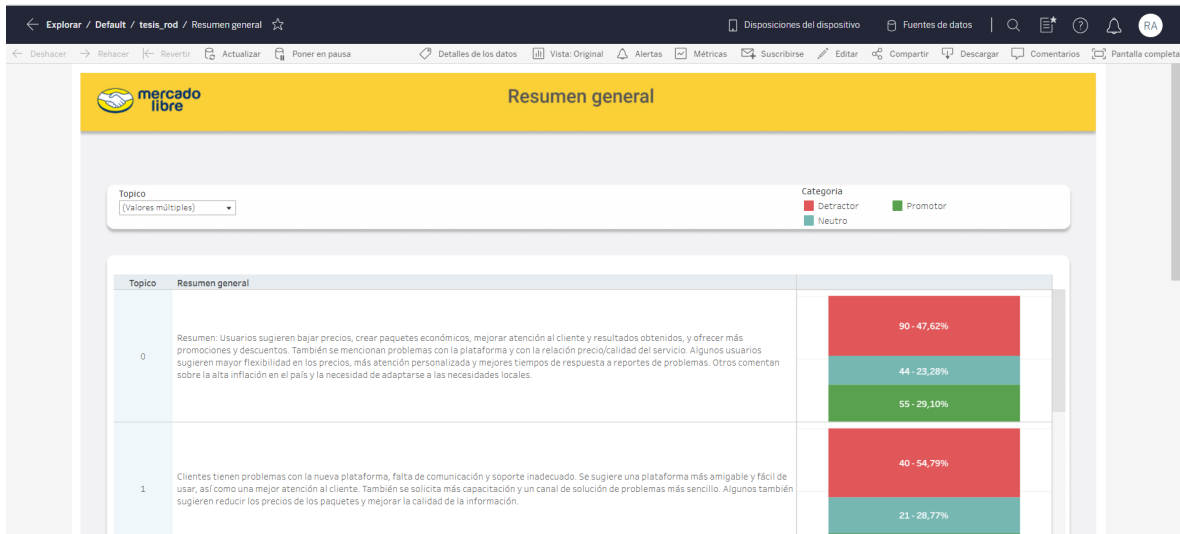


Figura 20: Resumen tópicos detectados

Fuente: Elaboración propia.

A su vez se propuso un segundo *dashboard* para centralizar de manera resumida la interpretación semántica obtenida por *ChatGPT* por cada tópico, permitiendo a su vez observar la distribución de sentimientos asociados a cada encuesta (categorías de sentimientos obtenidas por la definición del negocio que segmenta los comentarios de los encuestados en base a la nota que estos entregan).

Comentarios y conclusiones respecto a la utilidad y eficacia de estas visualizaciones serán revisadas durante la sección de validación de la solución 5.3.

CAPÍTULO 5

VALIDACIÓN DE LA SOLUCIÓN

5.1. Motivación

Durante esta etapa es importante recordar que el objetivo principal planteado en esta memoria es generar tópicos que permitan y faciliten la obtención de *insights* por y para el negocio, es por ello que con el fin de validar que los resultados obtenidos por el modelo candidato propuesto cumplan con los estándares deseados se llevaron a cabo diversas reuniones con los *skateholders* (equipo de NPS de Mercado Libre VIS). Ante esto se decide llevar a cabo una revisión manual en conjunto con el equipo de NPS sobre los tópicos detectados en las encuestas de vendedores profesionales.

Es importante mencionar que esta etapa se encuentra enfocada en analizar los resultados desde la perspectiva del área de negocio mas que por las métricas de coherencia y/o diversidad obtenidas anteriormente 4.6, dado que estas fueron utilizados para guiar la selección del mejor modelo posible en cuanto a la calidad de los tópicos, sin embargo ninguna de estas métricas puede asegurar que los tópicos modelados sean realmente relevantes o significativos para el negocio (para un análisis de estas métricas revisar 4.6).

5.2. Análisis de resultados

Para esta fase de validación se decide utilizar la muestra de datos de vendedores profesionales (dejando fuera a compradores o vendedores) dado que esta muestra resulta más familiar para el equipo de NPS dado que durante el último periodo han enfocando más su trabajo sobre estas encuestas debido a su mayor relevancia a la hora de levantar problemas/accionables con los ejecutivos o el equipo de IT.

Tras varias revisiones en conjunto con el equipo de NPS, se llegaron a las siguientes conclusiones:

- Es posible hacer un análogo con los tópicos ya existentes y *trackeados* por el negocio (atención al cliente, precios, procesos, producto y resultados obtenidos), sin embargo se destaca un análisis más robusto ante la captura de más respuestas sobre estos tópicos u otros problemas que no están estrechamente relacionados a alguno de estos tópicos. Esto se atribuye dado que el funcionamiento anterior para dar seguimiento a los tópicos definidos por el equipo de NPS no era lo suficientemente robusto, dada la dependencia en la presencia de palabras específicas que no capturaban la semántica, variaciones o faltas ortográficas en las mismas.

- La presencia de tópicos conformados por más de una idea hace sentido (son problemas relacionados la mayor parte de los casos), destacándose de esta manera la coherencia de los distintos temas relacionados a un único tópico, sin embargo se deja como deseable a futuro trabajar más para mejorar la diversidad los mismos.
- Si bien la herramienta permite trabajar distintos idiomas, se detecta 1 tópico sesgado por el idioma Ingles lo cual llama bastante la atención, sin embargo aquellos conformados por portugués y español se mezclan con naturalidad, esto último se destaca como algo útil, ya que actualmente al tener que manejar distintas reglas específicas por idiomas, se gasta bastante tiempo en realizar la comparación y analogía de los tópicos detectados entre el Portugués y el Español.
- Se logra ahorrar tiempo de análisis, anteriormente el equipo de *NPS* tenía que revisar las encuestas de manera desordenada (para aquellas que no formaban parte de ningún tópico provisto por las reglas definidas), mientras que ahora con el a priori automático que se obtiene respecto al tema que pertenece la encuesta, se puede organizar de manera mucho más consistente, ordenada e inteligente la lectura de los comentarios, facilitando de esta manera la delegación de conjuntos de encuestas a personas para dividir el trabajo de análisis y uso de las mismas.

A lo largo de las diversas reuniones sostenidas con el equipo de *NPS*, estos han demostrado una clara disposición y conformidad ante la propuesta de solución llevada a cabo, evidenciando su utilidad. Entre los diversos tópicos que se han ido analizando para la generación de *insights* se destacan los siguientes ejemplos:

- La satisfacción y servicio del cliente; la cual es divisible en sub-temas como los canales de atención entre ejecutivos y clientes, sumado a una clara evidencia que demuestra la necesidades de facilitar esta comunicación mediante el uso de *whatsapp* o teléfono, la mayoría de los clientes muestra un descontento ante la nula o difícil contactabilidad que existe al momento de querer solucionar problemas mediante correos con los ejecutivos de la empresa.
- Los resultados que provee la plataforma; la cual es divisible en el múltiples temas, como por ejemplo un bajo número de visualizaciones obtenidas dada una poca visibilidad dentro de la misma y la falta de publicidad y exposición de las publicaciones.
- Problemas con la plataforma; este tópico se explica a través de diversos puntos que critican el funcionamiento de la página, tildándola de poco amigable, sugiriendo mejorar el sistema de republicación, mencionando problemas a la hora de publicar desde el celular y sugiriendo mejorar el sistema de búsqueda por mapa.
- Los clientes solicitan mejorar el posicionamiento de la página y sus publicaciones en *Google*, promocionar las publicaciones en redes sociales como *Instagram* y *Facebook* además de integrar la publicidad con *Google Ads* y *Facebook Ads*.

5.3. Validación propuestas *dashboards BI*

Con el fin de evaluar la calidad de los *dashboards* propuestos 4.8, se llevo a cabo una encuesta a dos expertos del negocio pertenecientes al equipo de *NPS*, como consideraciones a tener en cuenta es importante mencionar que:

- Los encuestados no colaboraron en la creación de las visualizaciones.
- Estos observaron por primera vez los *dashboards* al momento de realizar la encuesta.
- Cada encuestado no tuvo contacto con otro encuestado o encuestador al momento de evaluar las visuales.

Dado que ambos *dashboards* 4.8 cumplen funciones distintas, se opta por evaluar cada elemento por separado, definiendo secciones de preguntas distintas para cada uno, evaluando elementos tanto del aspecto visual como la utilidad de los mismos.

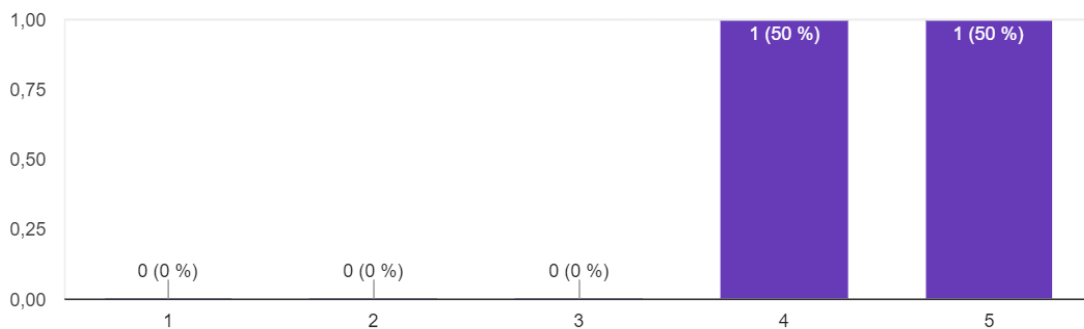
5.3.1. Feedback dashboard estratégico 19

Comodidad

En una escala del 1 al 5, donde 5 es un valor bueno y 0 un valor malo:

¿Qué tan bien o mal se encuentra estructurada la información?

2 respuestas



¿Es el nivel de detalle visualizado suficientemente útil para el análisis o te gustaría tener menos elementos/variables en pantalla?

2 respuestas

Es útil para realizar el análisis, pero requiere mucho conocimiento previo del modelo para poder entender, no es tan friendly para un usuario nuevo

El detalle de información permite encontrar y asociar de forma eficaz los datos sin embargo la asociación es genérica y se llegan a agrupar temas que a nivel de negocio no están relacionados.

¿Permite la visualización explorar de manera rápida el contenido de las encuestas? ¿Permite generar insights?

2 respuestas

Permite generar insights pero no es tan claro para poder dimensionar segmentos, ejemplo podría hacerse un mejor trabajo organizando todo lo relacionado a atención, lo de producto, etc

Si, al agrupar temas nos permite enfocar en soluciones específicas

¿Encuentra que existe un elemento claramente deficiente en las visualizaciones propuestas?

2 respuestas

No hay disponible un mapa de colores que permita diferenciar
Se podría detallar mejor que explica cada cuadro
En el mapa de insights, el resumen general y el resumen detallado siempre dicen lo mismo

No

5.3.2. Feedback dashboard resumen general 20

Dashboard resumen general

¿Permite esta visualización comprender sobre que hablan los usuarios de Mercado Libre VIS?

2 respuestas

La visualización está muy bien estructurada, lo que queda un poco confuso son los insights que termina arrojando el modelo. por ejemplo "Resultados obtenidos" puede salir en varios grupos de comentarios, lo que hace muy difícil dimensionar el peso real de esto ya que habría que indagar entre varios grupos

Si

¿En que ocuparías esta visual?

2 respuestas

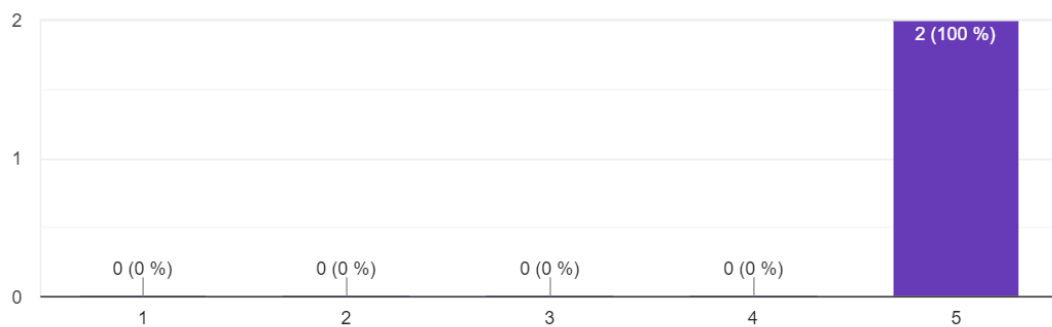
La visual está muy buena

generar propuestas para la mejora de producto

En un nivel del 1 al 5, donde 1 es totalmente disconforme y 5 es totalmente conforme.

¿Que tan conforme esta con la visualización provista?

2 respuestas



A partir del feedback obtenido es posible rescatar una satisfacción general respecto a como los elementos fueron presentados visualmente, sin embargo entre los elementos más importantes a mejorar se destaca que:

- La visual (*dashboard* estratégico 19) puede llegar a ser poco intuitiva en caso de no poseer conocimiento previo en el área.
- La diversidad de algunos tópicos puede ser mejorada.

Ambos elementos son importantes a tener en cuenta y no deben ser descuidados, el primer punto, si bien es posible mejorarlo en función del propio comentario adjunto que indica que sería bueno la agregación de breves descripciones que soporten la idea principal de cada visualización, es posible que existan mejores soluciones, una visualización debe poseer un diseño intuitivo inherente, el objetivo de una visual debe ser el de transmitir información clara y efectiva sin la necesidad de que esta sea explicada, es por ello que en futuras iteraciones realizar *mockups* y encuestar una mayor cantidad de analistas puede ser una buena alternativa para mejorar la construcción de estas visuales. Por otro lado, el segundo punto requiere de un mayor esfuerzo para ser mejorado, diversos motivos pueden causar esta baja diversidad entre algunos tópicos, ya sea esto una característica inherente de los datos donde cada comentario puede poseer más de un tópico o por un déficit en el modelo utilizado, al observar las métricas es posible evidenciar que la diversidad del mismo es mejorable, esto puede ser llevado a cabo tanto desde una hiperparametrización más fina del modelo empleado o por modificaciones en el mismo, como por ejemplo reemplazar *K-Means* por *fuzzy c-Means* para tener en cuenta la posibilidad de pertenencia de un comentario a múltiples tópicos.

Tanto estas como otras alternativas pueden ser propuestas para guiar la mejora en los resultados, es por ello que se da paso a la sección de trabajo a futuro.

5.4. Trabajo a futuro

Si bien se obtuvieron buenos resultados, se deja como trabajo a futuro y complementario la propuesta de múltiples mejoras a la solución propuesta; entre estas se propone:

- Expandir diccionario de *stopwords*, tras una inspección simple de las nubes de palabras 14 es posible evidenciar la presencia de palabras que no generan un valor nuevo a la hora de buscar *insights* para el negocio.
- Realizar *POS Tagging* para detectar la presencia de entidades y realizar la detección de tópicos sobre estos.
- Evaluar modelos de *embedding* creados por la empresa.

- Realizar análisis de sentimiento, ya que a pesar de obtener calificaciones positivas, a veces estas pueden estar destacando aspectos negativos de la página o viceversa.
- Especializar el modelo mediante una hiperparametrización más profunda.
- Realizar *fine-tune* del modelo de *embedding* para ajustarlo al contexto de los datos.
- Probar nuevamente *HDBSCAN* como modelo *clusterizador* y determinar si es posible reducir el número de *outliers* sin comprometer la calidad de sus métricas.
- Separar textos largos en oraciones, si bien la mayoría de las respuestas provistas por los encuestados son cortas, algunas de estas se alejan de la media.
- Evaluar el uso de un menor número de tópicos.
- Mejorar los *dashboards* provistos, reduciendo la cantidad de información por pantalla o presentándola de otra manera.
- *Indexar* la información por *sentence embeddings* para facilitar la extracción de información en base al conocimiento previo.
- Evaluar la factibilidad de realizar detección de tópicos especializados por país y vertical del negocio.
- Incorporar el tiempo como variable en la detección de tópicos.

Es posible evidenciar que existen muchas formas de extender, guiar y/o mejorar el trabajo realizado, el área de la minería de datos y el procesamiento del lenguaje natural es vasto y extenso, distintos enfoques pueden ser empleados para llegar a una misma solución, sin embargo es importante destacar que dentro del contexto de la industria no siempre es posible darse el tiempo de probar todas las posibilidades, saber guiar y discernir entre los múltiples caminos existentes es una habilidad importante a desarrollar para la toma de decisiones efectivas y obtener un éxito en los proyectos, ante esto y con el fin de aportar un flujo de trabajo futuro más claro, se destacan los siguientes pasos como los principales y más importantes a ser probados en un futuro para guiar una mejora en los resultados:

- Reemplazar *K-Means* por *Fuzzy c-Means* y/o *HDBSCAN*.
- Hiperparametrizar en mayor detalle el modelo candidato con una reducción de los tópicos a modelar.
- Realizar *fine-tuning* del modelo de *embedding* utilizado.

En principio (nada asegurado) estos cambios deberían guiar y resultar en una mejora de los resultados, el primer cambio se encuentra enfocado en adaptar la arquitectura ante supuestos más cercanos y lógicos respecto a como se pueden distribuir los datos en la vida real

(los comentarios pueden tener un grado de pertenencia a múltiples tópicos y/o los tópicos pueden estar conformados de múltiples subtópicos), a su vez el segundo punto se destaca dada la tendencia observada en las métricas 4.6 que denotan que a menor número de tópicos se obtienen mejores resultados, es posible que intrínsecamente los datos no posean tantos tópicos como los modelados, motivo que puede aportar a la baja diversidad obtenida, a su vez optar por un menor número de tópicos puede mejorar la efectividad de las visuales al reducir el número de variables por pantalla que la persona tiene que observar y comprender. Finalmente el último punto es recomendado dado que el contexto de las encuestas trabajadas esta sujeto al negocio y no corresponde a un lenguaje usual o coloquial, por lo que adaptar el contexto del modelo puede traer mejoras significativas en la calidad de sus representaciones y por ende en la detección de tópicos basados en la semántica de los mismos.

CAPÍTULO 6

CONCLUSIONES

6.1. Metas propuestas

Durante la presente memoria se ha realizado un análisis de diversos modelos pertenecientes al área del procesamiento del lenguaje natural, esto con el fin de tanto facilitar como guiar la generación de *insights* que pueden ser obtenidos a partir de los resultados de las encuestas de satisfacción que realizan las empresas a sus clientes, para ello se definieron los objetivos 1.4 que proponen como una posible solución al problema el uso de la modelación de tópicos y visualizaciones de los mismos para dar una mejor interpretabilidad del contenido asociado a las encuestas, posibilitando así una entendimiento más expedito de los mismos.

6.2. Contexto y entorno del trabajo

Esta memoria fue llevada a cabo como parte de un proyecto dentro del área de Vehículos, Infraestructura y Servicios dentro de la empresa Mercado Libre, la solución propuesta nace a partir de la problemática que representa para el equipo de *NPS* el análisis de estas encuestas tras cada cuatrimestre del año, donde la asignación de tópicos para las encuestas, es un proceso guiado por una definición de reglas que evalúan la presencia de palabras específicas dentro de cada encuesta; esta práctica guía la obtención de conclusiones erróneas y a la necesidad de una validación manual exhaustiva por parte del equipo de *NPS* a la hora de realizar reportes y levantar los accionables correspondientes para implementar mejoras en el área. Ante esto se realizó un entendimiento del problema en conjunto con el equipo de *NPS* para comprender las principales dificultades y necesidades del equipo para guiar y definir de esta manera los alcances del proyecto ya mencionados en los objetivos 1.4 de la tesis.

6.3. Dificultades enfrentadas

Entre las principales dificultades a manejar se destaco que:

- El contenido de las encuestas esta mayormente conformado por la presencia de textos cortos, esto se traduce en una dificultad a la hora de obtener *insights* útiles para el negocio y una posible degradación en los resultados obtenidos por los modelos que si toman en consideración la semántica del texto.
- La informalidad presente en los textos, que se visibiliza ante la existencia de errores ortográficos y modismos asociados a cada país induce una reducción en la concurrencia

de las palabras y significado de las mismas, reduciendo posiblemente el desempeño tanto de modelos basados en *BoW* y *embeddings*.

- Se destaca que dado el contexto del problema, la semántica asociadas a las palabras se encuentra sesgada fuertemente por las terminologías y definiciones del negocio, esto implica una posible reducción en el rendimiento de modelos basados en *embeddings* dado que la semántica de estas palabras no es similar a la que el modelo vio durante su entrenamiento.

Todo esto fue revisado y evaluado durante un análisis exploratorio preliminar de la data 4.3, donde es posible evidenciar que la gran mayoría de las respuestas provistas por los encuestados ronda alrededor de 20 palabras, a su vez se realizó nubes de palabras para evidenciar la presencia de faltas ortográficas y visualizar posibles candidatos a *stopwords*.

6.4. Propuesta de solución

Como posibles candidatos a propuesta de solución se evaluaron los modelos conocidos como *Latent Dirichlet Allocation*, *Fuzzy Topic Modeling*, *Combined Topic Models* y *BERTopic*, estos fueron elegidos dadas sus características descritas en 3.1, 3.3, 3.4 y 3.5, donde en breve palabras:

- **LDA:** Fue elegido como modelo de comparación base al ser uno de los pioneros en el área de la modelación de tópicos, sus resultados fueron los peores para las tres muestras de datos, se destaca que la coherencia que logran estos modelos no logra estar a la par de los demás modelos, esto puede ser explicado ante la poca concurrencia de las palabras producto tanto del hecho de que los textos sean cortos en su mayoría y a las faltas ortográficas presentes.
- **Fuzzy Topic Modeling:** Se seleccionaron modelos pertenecientes a esta familia de modelos, ya que a pesar de estar basados en técnicas de análisis sintáctico latente logran demostrar buenos resultados en comparación a otras variantes de los modelos de *LDA* y es una de las pocas técnicas que propone el uso del modelo *fuzzy c-means* (una técnica de *clustering* que si bien fue inventada en 1970 [MacQueen, 1967]), ha demostrado buenos resultados en la actualidad para la tarea de modelación de tópicos [Rijcken et al., 2022]. Durante la experimentación fue posible observar una degradación en sus resultados ante la disminución de palabras por oraciones (se obtuvo mejores resultados en el conjunto de datos de vendedores profesionales que vendedores), si bien estos resultados pueden motivar una explicación relacionada al hecho de que este método está condicionado a la concurrencia de las palabras, es necesario mayor experimentación para sostener tal hipótesis. Otro elemento a destacar es que este método ocupa bastante memoria, incluso llegando a dar problemas para su ejecución con el *dataset* de compradores.

- **Combined Topic Models:** Este modelo fue elegido dado que hace mejoras sobre el modelo *ProdLDA*, incorporando así información semántica sobre los mismos, sin embargo sigue haciendo uso de la técnica *BoW*, por lo que se convierte en un candidato atractivo a la hora de comparar los beneficios de incorporar información semántica a los modelos y por otra parte demostrar si sigue siendo beneficiario mantener la co-ocurrencia de las palabras como una característica para el modelo. Tras los experimentos realizados se observaron buenos resultados en todas las muestras de datos en comparación a las dos técnicas mencionadas anteriormente.
- **Zero Shoot Topic Modeling:** Este modelo fue elegido, ya que si bien forma parte de la familia de los *Combined Topic Models*, entrega un enfoque distinto, donde ya no es necesario el uso de *BoW* y abre la posibilidad de manejar palabras nuevas en conjuntos de datos nuevos, este modelo al igual que *CTM* mostró resultados consistentes en todas las muestras de datos, sin embargo se destaca que al igual que su contraparte (*CTM*) es necesario disponibilizar de bastante tiempo para su entrenamiento.
- **BERTopic:** A diferencia de los modelos mencionados anteriormente, *BERTopic* utiliza un enfoque distinto a todos los métodos anteriores, este método es modular, convirtiéndose así en una opción flexible que permite probar bastantes variantes para el mismo modelo, por su parte la incorporación de técnicas como *HDBSCAN* le otorgan el beneficio de poder optar por análisis a distintos niveles de detalles sin la necesidad de re entrenar al modelo. Durante la fase de experimentación mostró buenos resultados en todos los conjuntos de datos, sin embargo el alto número de *outliers* obtenidos no lo convierten en una alternativa viable para el negocio, es por ello que se opta por hacer uso de *KMeans* dado que este método no deja de lado ninguna encuesta (aún así se destaca la obtención de peores resultados en cuanto a diversidad de los tópicos modelados).

6.5. Limitaciones enfrentadas durante la experimentación

Durante el proceso de experimentación de los modelos se tomo la decisión de limitar el espacio de búsqueda de los hiperparámetros a únicamente la variable que define el número de tópicos a modelar, el motivo detrás de esta decisión se basa ante el hecho de que los tiempos de ejecución eran demasiado elevados como para permitir una hiperparametrización exhaustiva, al llevarse a cabo este proceso, cada modelo debe ser reentrenado con el nuevo conjunto de parámetros a probar, esto combinado con el hecho de que cada modelo debe ser entrenado en tres muestras de datos distintas (datos correspondientes a compradores, vendedores y vendedores profesionales) resulto en una infactibilidad para llevar a cabo una hiperparametrización más detallada y profunda.

6.6. Selección del modelo en base a métricas

Con el objetivo de evaluar la calidad de los tópicos obtenidos por los distintos modelos y así comparar la calidad de los mismos, se utilizaron métricas de coherencia y diversidad 4.6, si bien los resultados obtenidos indican que el modelo obtuvo un mejor desempeño a la hora de generar tópicos más coherentes y diversos, la elección y puesta en producción de un modelo requiere llevar a cabo una validación manual que corrobore que el modelo candidato para la tarea cumpla con las expectativas deseadas 4.2 y permita cumplir con los objetivos definidos, ante ello se optó por evaluar la calidad de los tópicos modelados por *BERTopic*, debido a que como ha sido mencionado anteriormente, este presenta resultados mejores y más consistentes en las tres muestras de datos, se suma entre los motivos el hecho de que esta técnica a diferencia de los modelos como *LDA* o *FuzzyTM* permite manejar distintos idiomas haciendo uso de un único modelo.

6.7. Validación e interpretabilidad de los resultados

Durante la validación del modelo se realizaron diversas reuniones con el equipo de *NPS*, durante ellas fue posible comprobar que los tópicos modelados se relacionan con conocimiento del negocio que se tenía a priori, sin embargo se destacó la visibilidad de tópicos los cuales no se tenían en cuenta con anterioridad, destacándose así una solicitud de la mejora de los medios como *whatsapp* y la comunicación telefónica que brinda la compañía a la hora de dar servicio al cliente a los vendedores profesionales, otro ejemplo permite descubrir una disconformidad con el funcionamiento de la app en ciertos aspectos relacionados al proceso de pago de inmuebles (que puede presentar fallos en algunos dispositivos), una disconformidad con el sistema que define automáticamente el orden de las fotos a mostrar de los productos, una clara discrepancia con el método actual que impide añadir marcas de agua a las imágenes (donde según vendedores destacan una inseguridad por publicaciones de otros usuarios sobre sus propias publicaciones), problemas graves con el sistema de renovación automática del pago de planes de venta, una clara disconformidad ante la incapacidad de añadir vínculos a las redes sociales del vendedor.

Por otro lado, con el fin de dar un valor agregado y facilitar la generación de *insights* se propusieron dos *dashboards* 4.8. Las visualizaciones desarrolladas otorgan visibilidad tanto a los tópicos modelados, sus significados e interpretación y a otras variables ajenas al análisis en la construcción de los tópicos, como lo son la distribución de los encuestados por país, vertical de negocio, evolución por fechas y sentimientos asociados. Dado el alto número de aperturas y tópicos, proponer una visualización que capture todos los elementos es una tarea compleja de abordar y por ende es fácil caer en visualizaciones saturadas de información y difíciles de interpretar, para solventar esta problemática se incorporan filtros que reduzcan la cantidad de información vista en pantalla en conjunto con múltiples visuales que buscan presentar una apertura distinta de los datos. A pesar de esta problemática se logra una satisfacción general tras encuestar a expertos del negocio,

evaluando positivamente tanto la construcción de las mismas como su utilidad, a pesar de ello quedan como puntos a mejorar la facilidad de uso y la diversidad de los tópicos.

6.8. Puesta en producción

El modelo implementado fue automatizado con el uso de herramientas internas de la empresa, la flujo definido durante la automatización del modelo es un proceso modular el cual permite y da la flexibilidad de modificar piezas mismo en un futuro, ya que cada parte del proceso es un paso que se ejecuta por separado de manera secuencial tras la finalización del paso anterior, por lo que en caso de querer modificar el preprocesamiento de los datos, el modelo utilizado o la fuente de datos utilizadas no supondrá un problema en el futuro.

6.9. Puntos de mejora

El área del la minería de datos y el procesamiento del lenguaje natural es basto y extenso, diversas soluciones o enfoques distintos a los propuestos pueden guiar a una solución similar o mejor, sin embargo el tiempo es un factor clave para el negocio, por lo que muchas veces probar todas las soluciones no es una tarea factible, es por ello que discernir y seleccionar los siguientes pasos para una eventual mejora del trabajo realizado es clave para reducir tiempos y evitar reprocesamientos innecesarios. Entre los principales elementos a incorporar dentro del flujo de trabajo llevado a cabo, se propone como trabajo a futuro:

- Reemplazar *K-Means* por *Fuzzy c-Means* y/o *HDBSCAN*.
- Hiperparametrizar en mayor detalle el modelo candidato con una reducción de los tópicos a modelar.
- Realizar *fine-tuning* del modelo de *embedding* utilizado.

Todos estos puntos están enfocados en mejorar la calidad de los tópicos modelados, tanto en coherencia como diversidad. Si bien durante este trabajo se obtuvieron buenos resultados en coherencia, aún es posible mejorar la diversidad de los tópicos modelados, es por ello que mejorar la calidad de las representaciones vectoriales y los hiperparámetros del modelo puede ser un buen punto de partida para la mejora de estos resultados.

ANEXOS

Tabla 12: Interpretación semántica de los resultados.

Fuente: Elaboración propia.

Tópico	c-TFIDF	Resumen hecho por ChatGPT
0	precio bajar costo paquete	Usuarios sugieren bajar precios, crear paquetes económicos, mejorar atención al cliente y resultados obtenidos, y ofrecer más promociones y descuentos. También se mencionan problemas con la plataforma y con la relación precio/calidad del servicio. Algunos usuarios sugieren mayor flexibilidad en los precios, más atención personalizada y mejores tiempos de respuesta a reportes de problemas. Otros comentan sobre la alta inflación en el país y la necesidad de adaptarse a las necesidades locales.
1	plataforma cliente usuario ca- pacitacion	Clientes tienen problemas con la nueva plataforma, falta de comunicación y soporte inadecuado. Se sugiere una plataforma más amigable y fácil de usar, así como una mejor atención al cliente. También se solicita más capacitación y un canal de solución de problemas más sencillo. Algunos también sugieren reducir los precios de los paquetes y mejorar la calidad de la información.
2	mapa pro- piedad búsqueda ubicación	Sugerencias para mejorar el portal incluyen la búsqueda por mapa, codificación de propiedades, filtros más precisos y una plataforma más amigable. También se pide mejorar la ubicación geográfica de las propiedades y las opciones de búsqueda por ubicación.
3	encuesta problema hacer duda	Las principales sugerencias son mejorar la atención al cliente, aumentar la cantidad de leads, simplificar las evaluaciones y mejorar el sistema de republicaciones. También hay solicitudes de solución de problemas y mayor exposición de los anuncios.
4	problema servicio cliente respuesta	Clientes insatisfechos con el servicio debido a problemas técnicos, falta de comunicación personalizada y mala atención al cliente.
5	publicación precio paquete publicar	Se han recopilado opiniones sobre la plataforma, y muchas personas sugieren que se deben bajar los precios de las publicaciones, mejorar los paquetes y ofrecer promociones más atractivas. También se mencionan problemas en la plataforma para publicar desde el celular y la modificación de precios.
6	whatsapp responder teléfono poder	Clientes piden más opciones de contacto directo y atención personalizada, como WhatsApp y teléfono, con respuesta inmediata y soluciones a problemas planteados. También solicitan mejoras en la plataforma, como opciones de chat online, enlazar consultas a un mensaje por WhatsApp y facilitar el proceso de contratación de anuncios. Además, solicitan la activación del número de Whatsapp Business y la mejora en la atención al cliente por parte de los ejecutivos comerciales.

7	anuncio anuncio publicidad fazer	Algunos usuarios solicitan una mejora en la plataforma y en la atención al cliente, así como más promociones y publicidad para mejorar la visibilidad de los anuncios. También se pide una segmentación según el grupo objetivo y precios más accesibles para los corredores autónomos.
8	inmobiliario inmueble mercado propiedad	Sugerencias para mejorar la presencia y los resultados en el mercado inmobiliario de Mercado Libre, incluyendo promociones, mejoras en la plataforma y atención al cliente, así como precios y planes más accesibles para inmobiliarias pequeñas y la generación de más leads calificados. También se han mencionado dificultades en la plataforma y limitaciones en la búsqueda de propiedades, así como la necesidad de una mejor estrategia de marketing y la flexibilización de los precios y planes.
9	lead re- sultado contra- tación cliente	The majority of feedback suggests the need to improve the quality and quantity of leads generated by the platform, as well as providing better value for the cost of the service. Some customers also call for more flexibility in the platform and advertising options. Others appreciate the current visibility and results provided by the platform. The importance of customer service and post-sales support is also highlighted by some respondents.
10	pago pagar factura tar- jeta	Clientes insatisfechos por problemas de facturación, cobros indebidos y mal servicio postventa. Sugieren mejorar canales de pago y atención al cliente.
11	propiedad publicación publicar publicado	Usuarios necesitan mejor atención al cliente y gestión de publicaciones, además de opciones de pago más flexibles y filtrar por fecha de publicación.
12	publicación publicar eliminar dar	Sugerencias para mejorar el servicio de publicaciones de MercadoLibre: dar tiempo de expiración, fomentar la creatividad, tener control interno por fecha de publicación, dar facilidad para republicar y eliminar publicaciones, mejorar la API, evitar repeticiones de publicaciones y tener una mejor atención personalizada. También se reportan algunos problemas como errores en la creación de publicaciones, bloqueos de publicaciones sin explicación y dificultades para encontrar los avisos.
13	publicación publicar plataforma mejorar	Solicitudes de mejoras en la plataforma de publicación, mayor visibilidad y posicionamiento de las publicaciones, atención al cliente más rápida y eficiente, posibilidad de modificación de publicaciones e implementación de estadísticas. También se mencionan problemas técnicos y de comunicación, así como la necesidad de una mayor variedad de productos en el catálogo.
14	contacto comunica- ción canal comunicar	Necesidad de mejorar la comunicación y tener canales de contacto más directos como un teléfono de atención al cliente.

15	ejecutivo atención contar cuenta	Mejorar atención y capacitación de ejecutivos de cuenta para solucionar problemas y mejorar comunicación con clientes.
16	agrícola maquinaria agricola agro	Se necesita mejorar la presencia y categorización de maquinarias agrícolas en la plataforma de Mercado Libre. También se sugiere generar más publicidad en el rubro y adaptar los medios de pago a las costumbres de los clientes rurales.
17	auto vehículo precio modelo	Usuarios solicitan liberar duplicidades de anuncios, mejorar opciones de búsqueda y categorización, y corregir errores en modelos, precios y ubicación de vehículos en Mercado Libre Vehículos. También se pide mejorar el soporte al cliente, eliminar el apartado online, aumentar la cantidad de fotos para algunos rubros, y quitar la fecha de publicación de los vehículos. Algunos usuarios sugieren incluir la opción para vehículos blindados y trabajar en el posicionamiento de las publicaciones.
18	foto su- bir video publicación	Sugieren mejorar la calidad y cantidad de imágenes de las propiedades publicadas, permitir la subida de videos y tours virtuales, simplificar el proceso de carga de fotos y mejorar la plataforma en general para una mejor experiencia del usuario. También piden una mayor inversión en tecnología para mostrar los inmuebles y la posibilidad de guardar propiedades en un banner para futuras publicaciones.
19	portal propiedad inmobilia- rio portal inmobilia- rio	Sugerencias para mejorar el Portal Inmobiliario: mejorar la atención al cliente, modernizar la página de búsqueda, eliminar la opción de destacar propiedades, eliminar la sobreexposición de grandes franquicias, tener chat en línea, mejorar la gestión de publicaciones y contactos. También se recomienda mejorar la plataforma de publicación y el filtrado de propiedades.
20	factura ser- vicio pagar mes	Quejas sobre la mala atención al cliente y problemas en la facturación en Mercado Libre. También se mencionan problemas en la contratación de planes, renovación automática no deseada, falta de solución a problemas y eliminación de herramientas útiles para los clientes.
21	mercado libre livre venta	Opiniones diversas sobre Mercado Libre. Algunos consideran que es una excelente plataforma para vender, pero otros piensan que ha empeorado con el tiempo. Se mencionan problemas con el servicio al cliente, el cobro de comisiones y la atención a los corredores de propiedades. También se discute la falta de segmentación en la oferta y la necesidad de mejorar la comunicación y el soporte técnico. Sin embargo, algunos usuarios siguen confiando en Mercado Libre como una herramienta útil para incrementar las ventas digitales.
22	plataforma innovar seguir excelente	Opiniones variadas sobre la plataforma inmobiliaria.

23	celular app poder móvil	Usuarios solicitan una app más eficiente y amigable para dispositivos móviles, que permita publicar y editar propiedades desde el celular y que tenga las mismas funcionalidades que la versión de escritorio. También piden un mejor servicio al cliente y precios más bajos.
24	portal filtro pagina búsqueda	Sugerencias para mejorar la plataforma de publicación y búsqueda de propiedades, así como la atención al cliente y la accesibilidad. También se pide la implementación de filtros más precisos y un diseño más amigable.
25	problema atención asesor solución	Resumen: - Comunicación y atención personalizada - Mayor rapidez en responder y solucionar problemas - Mayor cantidad de contactos y consultas - Mejorar la plataforma y estadísticas de anuncios - Capacitación de asesores de cuenta
26	inmueble inmobiliario portal imóvil	Sugieren mejorar la plataforma y la exposición de propiedades para captar más clientes e incrementar la eficacia en la venta de inmuebles. También se recomienda mayor atención al cliente y más capacitación para los usuarios.
27	inmobiliario inmueble propiedad cliente	Clientes se quejan de la falta de atención y respuesta, altos precios, problemas en la facturación y fallas en la publicación de propiedades en Portal Inmobiliario.
28	ejecutivo nunca atender solucionar	Mejorar la atención al cliente con ejecutivos personalizados y un contacto telefónico directo para resolver problemas y dudas.
29	servicio excelente buen seguir	Clientes satisfechos con el servicio, desde la atención al cliente hasta la plataforma y resultados obtenidos. Algunas sugerencias incluyen promociones especiales para clientes leales y mejorar la calidad de los datos proporcionados.
30	argentino inmobiliario portal mercado	Sugerencias para mejorar atención al cliente y servicios de publicación, así como problemas con facturación y pausa de publicaciones debido a precios bajos.
31	publicación plataforma problema hacer	Clientes piden una mejor atención al cliente, mejoras en la plataforma, solución de problemas técnicos y mayor eficacia en la publicación y promoción de sus productos o servicios. También solicitan más comunicación y capacitación en el uso de la plataforma.
32	cliente atención mejorar personalizado	Mejor atención al cliente, personalizada y rápida. Mayor exposición para potenciales clientes y opciones de paquetes más convenientes.

33	reserva comprador vendedor cliente	Clientes piden mejor atención al cliente, geolocalización precisa, prevención de fraudes, promociones para clientes antiguos y más filtros contra publicaciones engañosas.
34	correo teléfono contacto email	Problemas de comunicación y atención al cliente en la plataforma. Pedidos de atención telefónica, mejoras en la verificación de datos y actualización de publicaciones, entre otros. Algunos usuarios reportan falta de resultados en las ventas a través de la plataforma.
35	pandemia pandemio momento crisis	Los clientes necesitan descuentos en los precios debido a la crisis de la pandemia y una mejor atención al cliente con soluciones rápidas a los problemas de la plataforma.
36	inmueble propiedad inmo- biliario publicación	Sugerencias incluyen no permitir publicaciones repetidas, filtrar publicaciones de inmobiliarias no autorizadas, permitir publicar propiedades sin precio y mejorar la renovación de publicaciones.
37	facebook google instagram social	Recomendaciones: mejorar el posicionamiento en Google, promocionar en redes sociales como Instagram y Facebook, mejorar el servicio al cliente y bajar los precios de publicación. También sugieren integración con Google Ads y Facebook Ads y ofrecer opciones de publicidad más atractivas.
38	momento gracias seguir bien	Comentarios o valoraciones de clientes acerca de un servicio o producto.
39	portal pla- taforma propiedad cliente	El Portalinmobiliario podría mejorar la facilidad de uso, los precios y resultados para captar más clientes y mejorar su imagen de confianza. También se pide una atención personalizada, una plataforma más amigable y que la información personal sea confidencial.

6.1. Componentes de los modelos neuronales de *embeddings* de texto

Existe una gran diversidad de modelos neuronales que permiten la generación de *embeddings* de textos, a pesar de esta diversidad la gran mayoría son construidos bajo los mismos principios y arquitecturas bases que son dispuestas/unidas de formas diferentes, es por ello que se destacan arquitecturas como las *CNN*, *RNN*, *LSTM's* y los mecanismos de atención.

6.1.1. Convolutional Neural Networks, CNN

El origen de este tipo de redes neuronales surge gracias a la necesidad de capturar información de las imágenes de tal forma que el contenido de los datos (píxeles) que conforman

a la imagen, capturen la relación de proximidad que estos poseen; antes de que estas redes surgieran era común aplanar el contenido de las imágenes para que estas pudieran adaptarse y entrar como *input* a las redes neuronales *fully connected*, sin embargo al llevar a cabo esta práctica, se perdía la relación de proximidad que el contenido de estos píxeles guardaban entre sí. Si bien este tipo de arquitecturas ha demostrado buenos resultados en tareas como el procesamiento de imágenes o audio, su uso no se ha visto limitado a tales campo, este tipo de arquitecturas ha dado lugar a modelos capaces de procesar texto al nivel de palabras [Gong y Ji, 2018] e incluso al nivel de detalle de letras [Zhang *et al.*, 2016], se destaca incluso su presencia en el modelo *FastText*.

Entre las componentes principales que conforman una *CNN* es posible encontrar las *convolutional layers* y las *pooling layers*, estas capas pueden ser apiladas en un número indeterminado de veces (donde las *pooling layers* suelen ser ubicadas después de las *convolutional layers*). En términos generales a mayor profundidad las capas convolucionales serán capaces de detectar características más complejas (por ejemplo objetos en una imagen) mientras que en las primeras capas serán capaces de detectar elementos más simples (por ejemplo esquinas o colores de una imagen).

¿Qué es y cómo funciona una Convolutional Layer?

Las capas convolucionales son un bloque fundamental de las *CNN*, estas se encuentran conformadas únicamente de un *input* y un filtro (conocido también como *kernel*); para su funcionamiento, el *kernel* puede ser pensado como un detector de características, el cual tiene un tamaño bidimensional o unidimensional (su tamaño es un hiper parámetro que suele ser ajustado como un matriz de 3x3 en casos bidimensionales), este *kernel* recorrerá todo el *input* (dando saltos determinados por un *stride*) realizando operaciones del tipo producto punto, donde los pesos de este operan con los valores del *input*, almacenando de esta manera el resultado en un arreglo, el cual es conocido como un mapa de características. Es importante remarcar que tras cada operación convolucional, la *CNN* aplica una transformación *ReLU* a este mapa de características para introducir una no linealidad al modelo [ibm, 2023].

¿Qué es y cómo funciona una pooling Layer?

Estas capas son usadas con el fin de reducir la dimensionalidad de un *input* (el cual suele ser el *output* de una capa convolucional). Esta capa, al igual que el *kernel* recorre todo el *input*, sin embargo, realiza una operación de agregación sobre este de tal forma que el *output* posee un tamaño menor al original. Esta operación suele ser llevada a cabo de dos maneras por lo general:

- **Average pooling:** Este unifica el valor mediante el calculo de un promedio de los mismos.

- **Max pooling:** A diferencia de *Average pooling*, se conserva solo el mayor valor (esta opción suele ser más usada por los autores en la literatura).

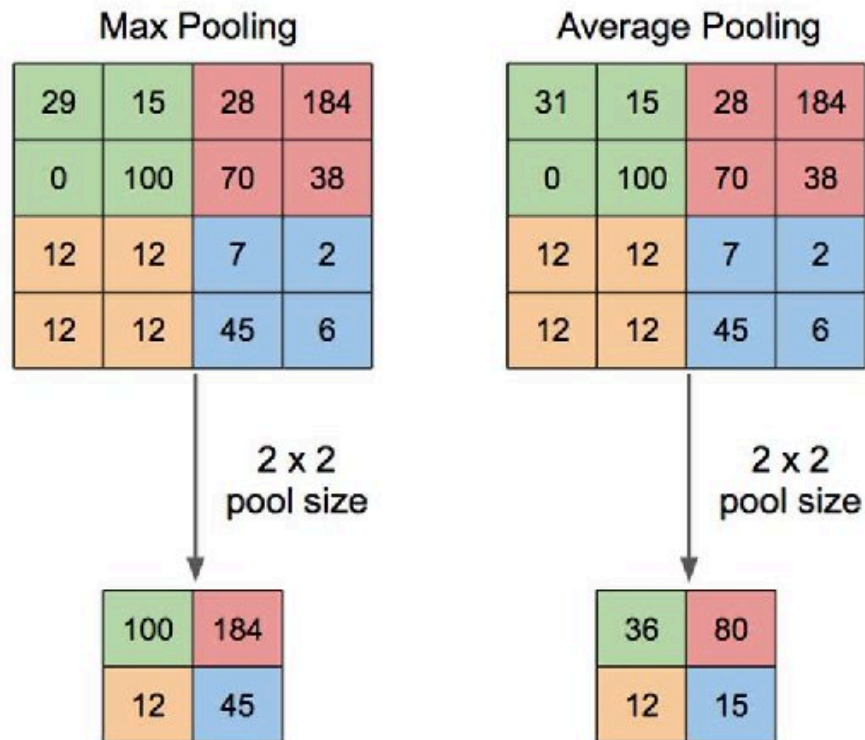


Figura 21: Pooling Layers

Fuente: Application of Transfer Learning Using Convolutional Neural Network Method for Early Detection of Terry's Nail, 2019, Muhamad YaniS, Si., M.T. Budhi Irawan

Estas capas son usadas con el fin de mejorar la eficiencia y reducir el riesgo de *overfitting*, a costa de una pérdida de información.

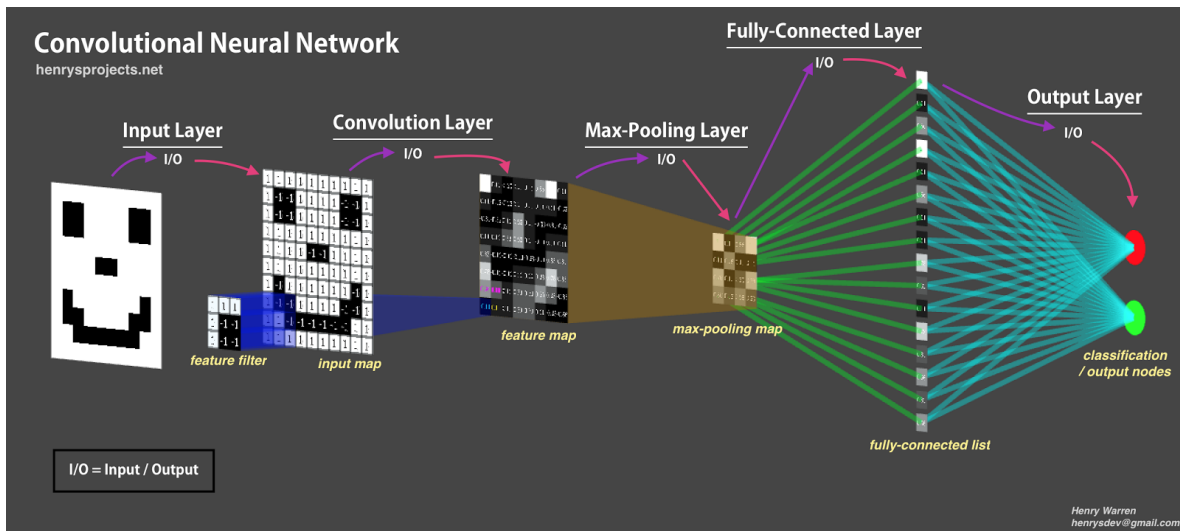


Figura 22: Explicación visual de las *Convolutional Neural Networks*
 Autor: Henry Warren.

6.1.2. Recurrent Neural Network, RNN

Las redes neuronales recurrentes son una variación de las redes *feed-forward-networks*, a diferencia de estas, este tipo de redes poseen una memoria interna la cual usan a su favor para procesar y entender de mejor manera relaciones secuenciales presentes en la data. Este tipo de redes surgen de manera casi natural en 1986 dada una necesidad de mejorar el rendimiento de las redes *feed-forward* que asumían una independencia entre los datos, derivando así en bajos rendimientos ante la presencia de datos secuenciales donde existe una clara dependencia entre los datos (como por ejemplo, audio, texto, etc.).

Este tipo de redes son llamadas recurrentes debido a que el output producido por una celda servirá como *input* para la siguiente, teniendo así cada neurona dos *inputs* como tal, produciéndose una dependencia / recurrencia entre las mismas que desemboca en un procesamiento más natural de los datos de tipo secuencial.

Existen cuatro tipos de redes RNN:

- **One to One:** Este tipo de redes es la más básica y sencilla de todas. A partir de un *input* se produce un *output*, un ejemplo de esto puede ser la clasificación de una imagen para responder si esta pertenece o no a un gato.
- **One to Many:** Esta arquitectura es usada para problemas donde a partir de un único *input* se busca predecir múltiples *outputs*, un ejemplo de esto puede ser la generación de música o texto a partir de una única nota o palabra inicial, donde la red busca predecir el siguiente elemento en base a predicciones anteriores.

- **Many to One:** Se busca predecir en base a una serie de *inputs* un único elemento, un ejemplo de esto puede ser la clasificación de las opiniones que los clientes realizan en las encuestas sobre el servicio de una empresa para detectar cuales han sido sus *feedbacks* respecto al servicio.
- **Many to Many:** A diferencia de las arquitecturas mencionadas anteriormente, este tipo de redes busca a partir múltiples *inputs* generar múltiples *outputs*, un ejemplo de esto puede ser la traducción de texto desde un idioma a otro.

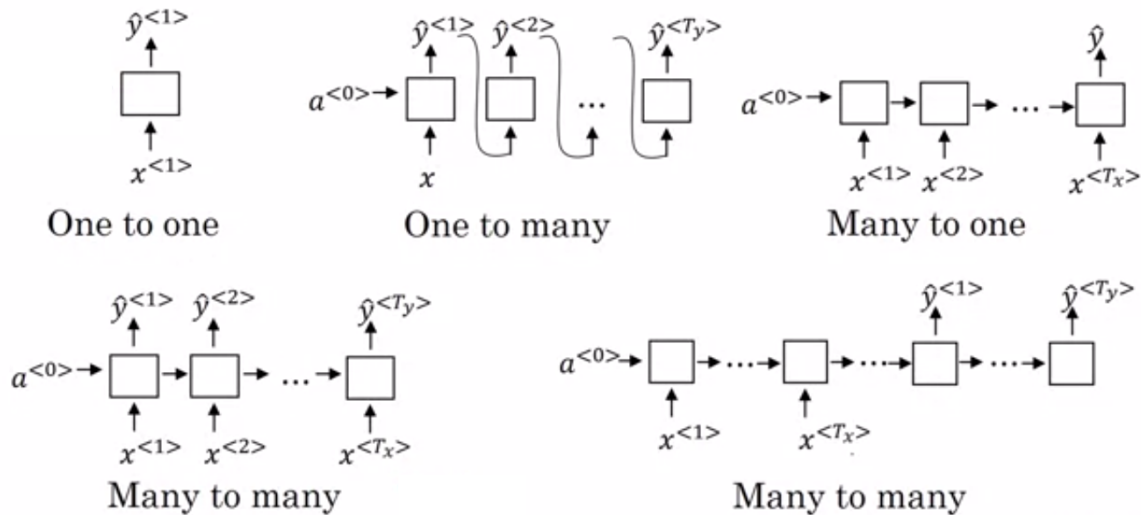


Figura 23: Types of RNN (Recurrent Neural Network)

Autor: Vijay Choubey

Uno de los problemas que presentan este tipo de arquitecturas ocurre con el gradiente de estas (elemento que regula el aprendizaje de las redes neuronales), problema el cual produce que en algunas ocasiones este tipo de redes presente malos resultados. Entre las principales desventajas de este tipo de arquitecturas se destacan los problemas conocidos como el desvanecimiento y explosión del gradiente [Bengio *et al.*, 1994].

- **Explosión del gradiente:** Este problema puede llegar a surgir durante la multiplicación de los gradientes en el proceso conocido como *backpropagation through time (BPTT)*, proceso el cual es usado para actualizar los pesos de la red neuronal durante su aprendizaje. Cuando el gradiente toma valores mayores a uno, este irá multiplicándose (y por ende propagándose) en cada paso de la *BPTT*, convirtiéndose cada vez en un valor más grande, desviando de esta manera la convergencia del modelo a resultados óptimos. Entre las principales soluciones a esta problemática se destaca truncar mediante la definición de un *threshold* los gradientes, normalización de los pesos o técnicas como *Adam* o *RMSProp* (algoritmos de optimización del gradiente) que ayudan a estabilizar los gradientes.

- **Desvanecimiento del gradiente:** Este problema es similar y ocurre bajo el mismo contexto que la explosión del gradiente, sin embargo, a diferencia de este último aquí el gradiente toma cada vez valores menores, produciendo que en el proceso de *back propagation* el modelo deje de aprender (o se demore demasiado en converger), entre las principales soluciones a este problema se destaca el uso de las funciones de activación, *batch normalization*, *skip connections* o usar *LSTM's*.

6.1.3. Long Short Term Memory, LSTM

A raíz de los múltiples problemas que presentaban las *RNN*, en 1997 a mano de Jürgen Schmidhuber surge una nueva variación y alternativa llamada arquitectura *LSTM*. Este tipo de redes poseen la habilidad de elegir que información recordar u olvidar, mejorando de esta la limitante que poseían las *RNN* ante problemas que presentan una dependencia a largo plazo.

Este tipo de arquitectura se encuentra conformado principalmente por tres componentes:

- **Compuerta de entrada:** Sirve para elegir que información nueva debe ser añadida o actualizada al estado de la celda *LSTM*, esta toma un *input* y el estado oculto previo de la celda anterior, los concatena horizontalmente y realiza la operación del producto punto con una matriz de pesos (la cual es aprendida durante el entrenamiento) para luego sumar un *bias* y pasar el resultado por una función de activación *sigmoid*, este resultado representa la decisión de la compuerta para elegir que tanto del input debe ser tomado en cuenta, sin embargo, el valor es multiplicado con un vector candidato como estado nuevo de la celda que surge a partir de aplicar la función de activación *tanh* tras sumar un *bias* a la multiplicación de una matriz de pesos aprendible con la concatenación del estado oculto anterior y el input.
- **Compuerta de olvido:** Esta se encarga de elegir que información almacenada ya no es relevante para el contexto, esta compuerta concatena el estado anterior con el *input* actual para multiplicarlos con una matriz de pesos aprendible, luego de eso pasa el resultado por una función de activación *sigmoid*, el resultado de esto indica que porcentaje del estado oculto anterior ha de ser retenido u olvidado.
- **Compuerta de salida:** Se encarga de seleccionar que información almacenada en el estado de la celda debe ser liberada. Básicamente se concatena el estado oculto de la celda anterior y el input actual para ser multiplicados por una matriz de pesos aprendibles, a este resultado se le suma un *bias* y es pasado por una función *sigmoid*, lo cual es multiplicado por la función de activación *tanh* del estado actual de la celda para así obtener el estado oculto de esta (el cual es se interpreta como una elección selectiva basada en el estado actual de la celda para apoyar la toma de decisiones futuras).

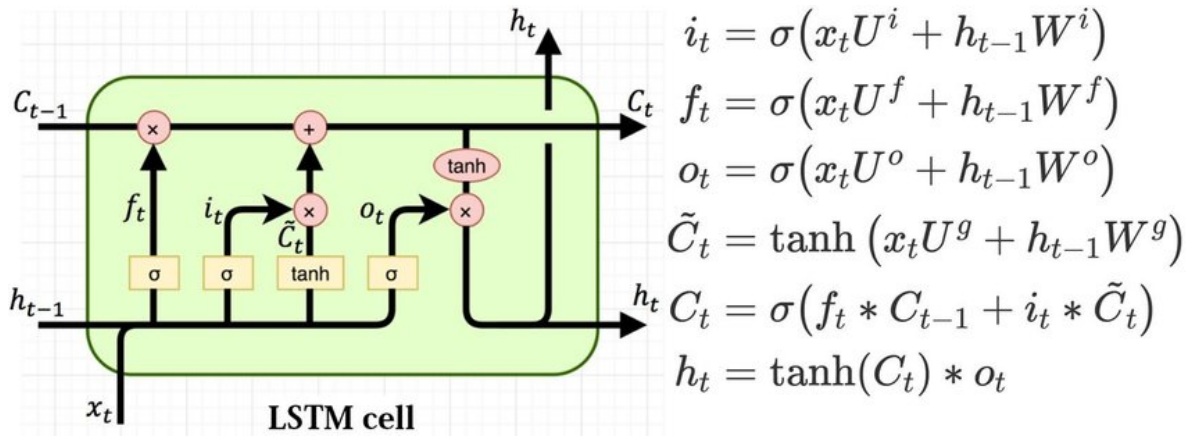


Figura 24: LSTM's

Fuente: Designing neural network based decoders for surface codes, Savvas Varsamopoulos, Koen Bertels, Carmen G. Almudever, November 2018

Si bien esta arquitectura presenta grandes mejoras en comparación a las *RNN*, no se encuentra exenta de poseer desventajas, entre los puntos más importantes a destacar se encuentra: un alto nivel de coste computacional durante el entrenamiento, la dificultad de elegir el número de unidades o bloques de memoria (el cual si bien al aumentar puede ayudar a detectar patrones más complejos también puede resultar en un mayor riesgo a *overfitting*) y el hecho de que cuando estas se enfrentan a datos con demasiado largo (como por ejemplo textos de un libro), su rendimiento se verá degradado, resultando más favorable por optar por otro tipo de arquitecturas como los *transformers* que se encuentran basados en mecanismos atencionales.

6.1.4. Attention mechanism

Este tipo de arquitectura surgió primeramente a raíz de la necesidad y limitaciones que presentaban los modelos *Seq2Seq* (*Sequence-to-Sequence*) tradicionales, los modelos *Seq2Seq* fueron diseñados inicialmente para tareas como la traducción de texto, donde las secuencias de entrada y de salida pueden presentar largos muy variables; la forma de abordar este tipo de problemas por estos modelos fue la de utilizar *encoders*; que eran responsables de llevar a una representación vectorial de largo fijo la sentencia de entrada (los cuales solían ser implementados con *RNNs*, *LSTMs* o arquitecturas *GRUs*) para que así un *decoder* tomara esta representación vectorial de largo fijo y la usara en conjunto con la representación vectorial de la palabra predicha anteriormente para predecir la siguiente palabra y así ir construyendo la sentencia de salida. El principal problema de este *approach* yace en cómo se construye la representación vectorial, esta es formada solamente por el estado oculto final, dejando de lado los estados intermedios; si bien eso supone una reducción en la memoria usada por los modelos, también implica una reducción en su desempeño, ya que para secuencias muy largas, el estado oculto final no es capaz de retener suficiente información de los estados

ocultos previos, desembocando así en una representación vectorial que ha perdido información y que no es lo suficientemente buena para el *decoder*. A raíz de estos problemas es que surgen los modelos atencionales, donde la idea principal detrás de ellos es resolver el cuello de botella generado por solamente utilizar el último estado oculto como la representación vectorial de las sentencias de entradas; dando así también uso a los estados ocultos intermedios previos a la hora de construir el vector de contexto que hará uso el *decoder*. Inicialmente la forma de llevar a cabo esto fue agregando una capa *feed-forward* en el *encoder* que tomara como *input* todos los estados ocultos presentes en el *encoder* en conjunto con el estado oculto del paso previo retornado por el *decoder* para obtener pesos (que representan la atención que se le debe dar a cada palabra) los cuales son pasados por una función *softmax*, a partir de esto se realiza un producto punto entre estos valores y los vectores de estados ocultos, dando así lugar a una representación vectorial rica en información relevante para el *decoder*, sin embargo, con el pasar de los años surgió otro *approach*, conocido como la atención basa en *Query, Key and value (QKV)*, donde la principal diferencia con el *approach* anterior radica en como se calculan los pesos de atención, donde se sigue manteniendo la idea de usar un *encoder* y *decoder*, sin embargo estos no utilizan una red *feed-forward* para calcular la representación vectorial intermedia entre ambos; si no mas bien se opta por el uso los siguientes tres elementos:

- **Query:** Corresponde a una proyección del estado oculto provisto por el *decoder*, este es calculado como un producto punto entre una matriz de pesos aprendibles y el estado oculto del *decoder*, la idea detrás de esto es llevar a cabo una alineación con las dimensiones de los demás componentes.
- **Key:** Análoga a la *query*, es una proyección del estado oculto del *encoder* y es calculada de la misma manera pero con una matriz de pesos aprendibles propia.
- **Value:** Funciona de la misma manera a la *key*, sin embargo, se usos son distintos.

La *query* y la *key* son usadas en conjunto para calcular un puntaje de similaridad entre ambas, este valor representa la atención que debe dársele a la llave, ya que entre más cercanas o similares son las proyecciones de los estados ocultos, se puede entender que se ha aprendido una mejor forma de representar, enfocar y dar uso a estos mismos sin presentar una perdida de información entre lo que el *encoder* y el *decoder* ven (la información que manejan y tienen disponible). Estos puntajes de similaridad son pasadas por una función de normalización *softmax*, resultado el cual es multiplicado con el *value* (proyección que representa el *encoder*); esto se hace ya que, en el primer paso se entiende a cual elemento es importante dar mayor atención, pero ahora se "filtra o enfoca" solo la información más relevante de la representación vectorial del *encoder* (originándose así la "atención"), una vez realizado esto se concatena este resultado con el *input* del *decoder* (por ejemplo la palabra que se busca traducir), lográndose así obtener una representación vectorial rica en información basada en el contexto y contenido.

Los modelos atencionales poseen una larga lista de variaciones, como por ejemplo los modelos:

- **Self-Attention:** A diferencia del *approach* convencional, el mecanismo de atención tiene acceso a un contexto global y no solo a una parte del *input*, atendiendo mejor relaciones de larga duración.
- **Multi-Head Attention:** Esta variación aplica en paralelo varias veces el mecanismo de atención (las cuales se les llaman como cabezas) donde cada una aprende proyecciones distintas a partir del mismo *input*.
- **Masked-Attention:** Esta variante difiere de las demás, dado que esta oculta partes de la secuencia con el fin de aprender a como modelar la generación de la secuencia, existen subvariantes como la *Local-Attention* o la *Sparse-Attention* que buscan ocultar secciones de la secuencia, ayudando a reducir el coste computacional cuando las secuencias son muy largas y/o también reduciendo la posibilidad de *over-fitting*.

REFERENCIAS BIBLIOGRÁFICAS

- [ibm, 2023] (2023). What are convolutional neural networks?
- [Albalawi *et al.*, 2020] Albalawi, R., Yeap, T. H., y Benyoucef, M. (2020). Using topic modeling methods for short-text data: A comparative analysis. *Frontiers in Artificial Intelligence*, 3.
- [Alghamdi Rubayyi, 2015] Alghamdi Rubayyi, A. K. (2015). A survey of topic modeling in text mining. *International Journal of Advanced Computer Science and Applications*, 6(1).
- [Alkhamees y Fasli, 2016] Alkhamees, N. y Fasli, M. (2016). Event detection from social network streams using frequent pattern mining with dynamic support values. pp. 1670–1679.
- [Babar, 2013] Babar, S. (2013). Text summarization: An overview.
- [Bengio *et al.*, 1994] Bengio, Y., Simard, P., y Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, 5(2):157–166.
- [Berba, 2020] Berba, P. (2020). Understanding hdbscan and density-based clustering.
- [Bhadane *et al.*, 2015] Bhadane, C., Dalal, H., y Doshi, H. (2015). Sentiment analysis: Measuring opinions. *Procedia Computer Science*, 45:808–814. International Conference on Advanced Computing Technologies and Applications (ICACTA).
- [Bianchi *et al.*, 2021a] Bianchi, F., Terragni, S., y Hovy, D. (2021a). Pre-training is a hot topic: Contextualized document embeddings improve topic coherence. En *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pp. 759–766, Online. Association for Computational Linguistics.
- [Bianchi *et al.*, 2021b] Bianchi, F., Terragni, S., Hovy, D., Nozza, D., y Fersini, E. (2021b). Cross-lingual contextualized topic models with zero-shot learning. En *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 1676–1683, Online. Association for Computational Linguistics.
- [Blei *et al.*, 2001] Blei, D., Ng, A., y Jordan, M. (2001). Latent dirichlet allocation. volumen 3, pp. 601–608.
- [Camacho-Collados y Pilehvar, 2018] Camacho-Collados, J. y Pilehvar, M. T. (2018). On the role of text preprocessing in neural network architectures: An evaluation study on text categorization and sentiment analysis.
- [Cataldi *et al.*, 2010] Cataldi, M., Caro, L. D., y Schifanella, C. (2010). Emerging topic detection on twitter based on temporal and social terms evaluation.

- [Chiche y Yitagesu, 2022] Chiche, A. y Yitagesu, B. (2022). Part of speech tagging: a systematic review of deep learning and machine learning approaches. *Journal of Big Data*, 9(10).
- [Coenen y Pearce,] Coenen, A. y Pearce, A. Understanding umap.
- [Dakhel et al., 2023] Dakhel, A. M., Majdinasab, V., Nikanjam, A., Khomh, F., Desmarais, M. C., Ming, Z., y Jiang (2023). Github copilot ai pair programmer: Asset or liability?
- [Ding et al., 2018] Ding, R., Nallapati, R., y Xiang, B. (2018). Coherence-aware neural topic modeling.
- [Drus y Khalid, 2019] Drus, Z. y Khalid, H. (2019). Sentiment analysis in social media and its application: Systematic literature review. *Procedia Computer Science*, 161:707–714. The Fifth Information Systems International Conference, 23-24 July 2019, Surabaya, Indonesia.
- [Garcia y Berton, 2021] Garcia, K. y Berton, L. (2021). Topic detection and sentiment analysis in twitter content related to covid-19 from brazil and the usa. *Applied Soft Computing*, 101:107057.
- [Gong y Ji, 2018] Gong, L. y Ji, R. (2018). What does a textcnn learn?
- [Grootendorst, 2022] Grootendorst, M. (2022). Bertopic: Neural topic modeling with a class-based tf-idf procedure.
- [Hajjem y Latiri, 2017] Hajjem, M. y Latiri, C. (2017). Combining ir and lda topic modeling for filtering microblogs. *Procedia Computer Science*, 112:761–770. Knowledge-Based and Intelligent Information Engineering Systems: Proceedings of the 21st International Conference, KES-20176-8 September 2017, Marseille, France.
- [Hernandez, 2019] Hernandez, A. A. R. (2019). Detección de polaridad por tópicos para textos cortos en español.
- [Hinton y Salakhutdinov, 2009] Hinton, G. E. y Salakhutdinov, R. R. (2009). Replicated softmax: an undirected topic model. En Bengio, Y., Schuurmans, D., Lafferty, J., Williams, C., y Culotta, A., editores, *Advances in Neural Information Processing Systems*, volumen 22. Curran Associates, Inc.
- [Hu et al., 2016] Hu, W., Zhang, J., y Zheng, N. (2016). Different contexts lead to different word embeddings. En *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pp. 762–771, Osaka, Japan. The COLING 2016 Organizing Committee.
- [IBM, 2021] IBM (2021). What is Natural Language Processing? | IBM.
- [Ismailov et al., 2016] Ismailov, A., Jalil, M. A., Abdullah, Z., y Rahim, N. A. (2016). A comparative study of stemming algorithms for use with the uzbek language. En *2016 3rd International Conference on Computer and Information Sciences (ICCOINS)*, pp. 7–12.

- [Kanerva *et al.*, 2020] Kanerva, J., Ginter, F., y Salakoski, T. (2020). Universal lemmatizer: A sequence to sequence model for lemmatizing universal dependencies treebanks.
- [Karami *et al.*, 2017] Karami, A., Gangopadhyay, A., Zhou, B., y Kharrazi, H. (2017). Fuzzy approach topic discovery in health and medical corpora. *International Journal of Fuzzy Systems*, 20(4):1334–1345.
- [Kavlakoglu,] Kavlakoglu, E. NLP vs. NLU vs. NLG: the differences between three natural language processing concepts - IBM Blog — ibm.com.
- [Li *et al.*, 2010] Li, S., Lee, S., Chen, Y., Huang, C.-R., y Zhou, G. (2010). Sentiment classification and polarity shifting. pp. 635–643.
- [Liu y Xu, 2020] Liu, Y. y Xu, Q. (2020). Short text classification model based on multi-attention. En *2020 13th International Symposium on Computational Intelligence and Design (ISCID)*, pp. 225–229. IEEE.
- [MacQueen, 1967] MacQueen, J. B. (1967). Some methods for classification and analysis of multivariate observations. En Cam, L. M. L. y Neyman, J., editores, *Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability*, volumen 1, pp. 281–297. University of California Press.
- [Mahendra *et al.*, 2021] Mahendra, M., Luo, Y., Mills, H., Schenk, G., Butte, A. J., y Dudley, R. A. (2021). Impact of different approaches to preparing notes for analysis with natural language processing on the performance of prediction models in intensive care. *Critical care explorations*, 3(6).
- [McInnes *et al.*, 2020] McInnes, L., Healy, J., y Melville, J. (2020). Umap: Uniform manifold approximation and projection for dimension reduction.
- [Mimno *et al.*, 2011] Mimno, D., Wallach, H., Talley, E., Leenders, M., y McCallum, A. (2011). Optimizing semantic coherence in topic models. En *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pp. 262–272, Edinburgh, Scotland, UK. Association for Computational Linguistics.
- [Moens *et al.*, 2014] Moens, M.-F., Li, J., y Chua, T.-S. (2014). *Mining User Generated Content*. Chapman Hall/CRC.
- [Monera, 2019] Monera, V. (2019). LENGUAJE, LENGUA, HABLA Y NORMA - DIVINAS PALABRAS. Victoria Monera — victoriamonera.com.
- [Murakami y Chakraborty, 2022] Murakami, R. y Chakraborty, B. (2022). Investigating the efficient use of word embedding with neural-topic models for interpretable topics from short texts. *Sensors*, 22(3). Cited by: 3; All Open Access, Gold Open Access, Green Open Access.
- [Nan *et al.*, 2019] Nan, F., Ding, R., Nallapati, R., y Xiang, B. (2019). Topic modeling with wasserstein autoencoders.

- [Radford y Narasimhan, 2018] Radford, A. y Narasimhan, K. (2018). Improving language understanding by generative pre-training.
- [Reimers y Gurevych, 2019a] Reimers, N. y Gurevych, I. (2019a). Sentence-bert: Sentence embeddings using siamese bert-networks. *CoRR*, abs/1908.10084.
- [Reimers y Gurevych, 2019b] Reimers, N. y Gurevych, I. (2019b). Sentence-bert: Sentence embeddings using siamese bert-networks.
- [Rijcken *et al.*, 2022] Rijcken, E., Mosteiro, P., Zervanou, K., Spruit, M., Scheepers, F., y Kaymak, U. (2022). Fuzzytm: a software package for fuzzy topic modeling. En *2022 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pp. 1–8.
- [Roy, 2021] Roy, A. (2021). Recent trends in named entity recognition (ner).
- [Sbalchiero y Eder, 2020] Sbalchiero y Eder, S. (2020). Topic modeling long texts and the best number of topics. some problems and solutions. 54:1095–1108.
- [Shen, 2009] Shen, D. (2009). *Text Categorization*, pp. 3041–3044. Springer US, Boston, MA.
- [Sia *et al.*, 2020] Sia, S., Dalmia, A., y Mielke, S. J. (2020). Tired of topic models? clusters of pretrained word embeddings make for fast and good topics too! En *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1728–1736, Online. Association for Computational Linguistics.
- [Sinharay, 2010] Sinharay, S. (2010). Continuous probability distributions. En Peterson, P., Baker, E., y McGaw, B., editores, *International Encyclopedia of Education (Third Edition)*, pp. 98–102. Elsevier, Oxford, third edition edición.
- [Snyder, 2015] Snyder, R. M. (2015). An introduction to topic modeling as an unsupervised machine learning way to organize text information.
- [Srivastava y Sutton, 2017] Srivastava, A. y Sutton, C. (2017). Autoencoding variational inference for topic models.
- [Stoy, 2021] Stoy, L. (2021). Topic modelling: A deep dive into lda, hybrid-lda, and non-lda approaches.
- [Tomar y Sharma, 2016] Tomar, D. S. y Sharma, P. (2016). Polarity analysis using sentiwordnet based an algorithm.
- [Xia *et al.*, 2019] Xia, L., Luo, D., Zhang, C., y Wu, Z. (2019). A survey of topic models in text classification. En *2019 2nd International Conference on Artificial Intelligence and Big Data (ICAIBD)*, pp. 244–250.
- [Yin y Wang, 2014] Yin, J. y Wang, J. (2014). A dirichlet multinomial mixture model-based approach for short text clustering. En *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '14*, p. 233–242, New York, NY, USA. Association for Computing Machinery.

- [Zhang y Liu, 2014] Zhang, L. y Liu, B. (2014). *Aspect and Entity Extraction for Opinion Mining*, pp. 1–40. Springer Berlin Heidelberg, Berlin, Heidelberg.
- [Zhang *et al.*, 2016] Zhang, X., Zhao, J., y LeCun, Y. (2016). Character-level convolutional networks for text classification.
- [Zhao *et al.*, 2021] Zhao, H., Phung, D., Huynh, V., Jin, Y., Du, L., y Buntine, W. (2021). Topic modelling meets deep neural networks: A survey.