2nd International Symposium on Master Engineering

Booklets

RENIECYT - LATINDEX - EBSCO - Research Gate - DULCINEA - CLASE - Sudoc - HISPANA - SHERPA UNIVERSIA - Google Scholar DOI - REDIB - Mendeley - DIALNET - ROAD - ORCID

Title: Text mining for question classification in the context of electronic commerce

Authors: DE LEÓN-BARRÓN, Melissa Alejandra & RÍOS-ALVARADO, Ana Bertha

Editorial label ECORFAN: 607-8695
BIMES Control Number: 2022-02
BIMES Classification (2022): 231122-0002

Pages: 39
RNA: 03-2010-032610115700-14

ECORFAN-México, S.C.
143 – 50 Itzopan Street
La Florida, Ecatepec Municipality
Mexico State, 55120 Zipcode
Phone: +52 1 55 6159 2296
Skype: ecorfan-mexico.s.c.
E-mail: contacto@ecorfan.org
Facebook: ECORFAN-México S. C.
Twitter: @EcorfanC

www.ecorfan.org

Holdings
Mexico Colombia Guatemala
Bolivia Cameroon Democratic
Spain El Salvador Republic
Ecuador Taiwan of Congo
Peru Paraguay Nicaragua

Introducción

- El Internet es uno de los principales requisitos para la existencia del comercio electrónico
- El comercio electrónico se está volviendo popular y ha creado tendencia en nuestro estilo de vida
- La cantidad de usuarios en el mercado mundial del comercio electrónico fue de 1.920 millones de personas en el 2019
- Debido a la pandemia por el COVID-19, el Internet se convirtió en el principal medio de compra – venta

Planteamiento del problema

- Necesidad de la clasificación de preguntas de un dominio de comercio electrónico
- El comercio electrónico enfrenta el reto de responder automáticamente preguntas realizadas por los clientes

37%

de los usuarios abandonan una compra al no obtener una respuesta inmediata

Planteamiento del problema

- Mejorar la precisión de los modelos de clasificación de preguntas

- En el español no se ha abordado por completo la tarea de clasificación de preguntas

Justificación

- El contar con un modelo de clasificación de textos cortos en español permitirá diseñar y desarrollar sistemas de respuestas automáticas a preguntas en español

- El 83% de los usuarios compran en aquellas tiendas en línea valoradas con un servicio de atención al cliente de calidad

Objetivo general

Desarrollar un método de clasificación automática de preguntas en español en el dominio de comercio electrónico basado en características léxicas.

Objetivos específicos

- Estudiar los modelos de representación de textos que se adapten al dominio de comercio electrónico en idioma español y seleccionar un modelo de representación

- Seleccionar e implementar algoritmos de clasificación de textos cortos

- Evaluar los algoritmos de clasificación de textos mediante las medidas de exactitud, precisión, cobertura y medida F

Hipótesis

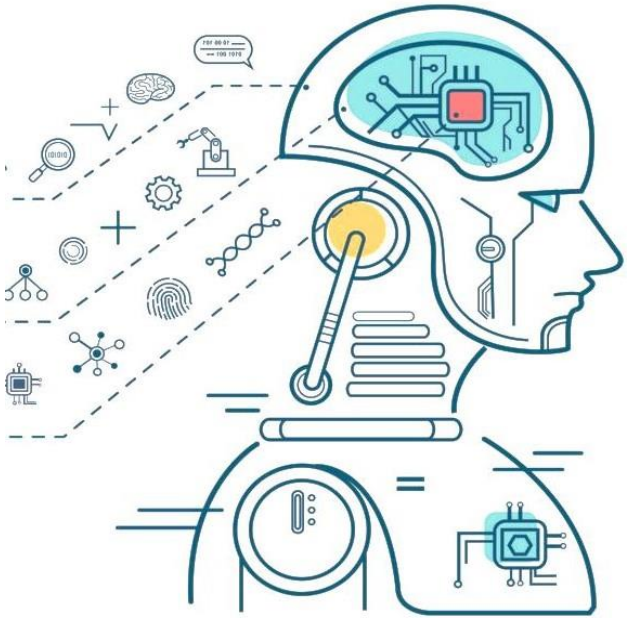
A partir del diseño de un modelo de representación de textos cortos basado en características léxicas y la selección de un modelo de clasificación de textos cortos en español se podrá clasificar preguntas con mayor precisión y cobertura que los reportados en la literatura.

Marco Teórico

Preprocesamiento de texto

Representación del texto

Aprendizaje Supervisado

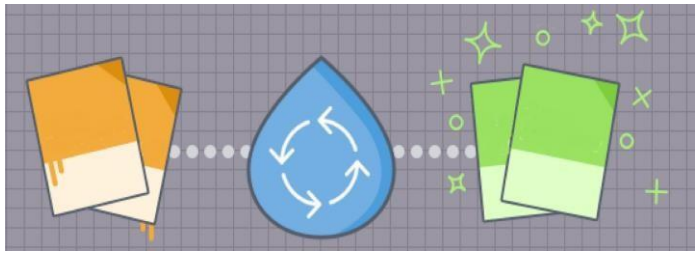


Marco Teórico

- Minería de Texto

- Extraer información útil e importante de formatos de documentos heterogéneos, tales como páginas web, correos electrónicos, medios sociales, artículos de revistas, etc.

- Esto se hace mediante la identificación de patrones dentro de los textos, tales como tendencias en el uso de palabras, estructura sintáctica, etc.



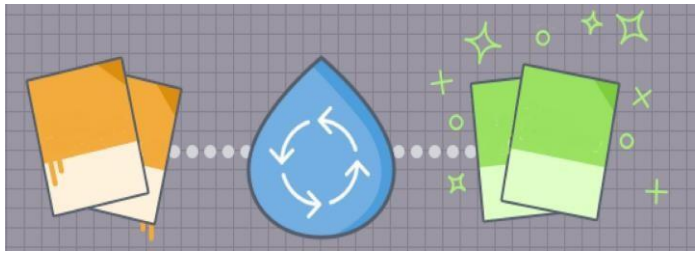
Marco Teórico

Preprocesamiento de texto

Representación del texto

Aprendizaje Supervisado

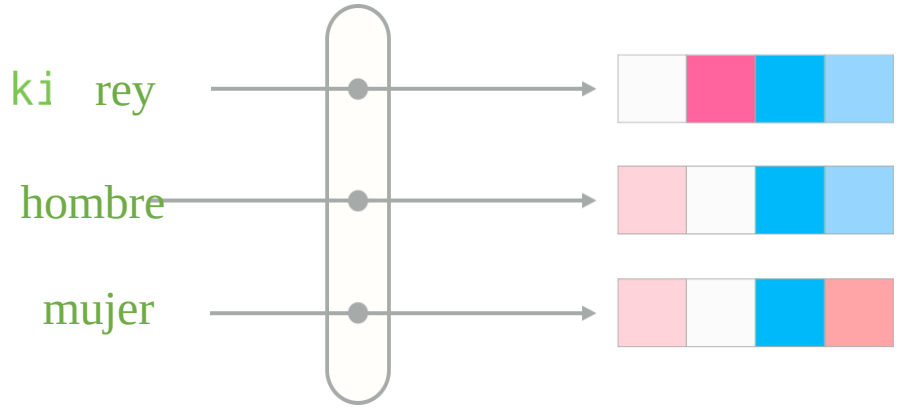
- Es el primer paso para la construcción de un modelo de clasificación de textos
- El preprocesamiento consiste en eliminar el ruido que pueda encontrarse en el texto
- Su objetivo principal es representar cada documento como una característica del vector de representación



Marco Teórico

- Transformar el texto a una representación estructurada

- Se utilizan técnicas para la extracción de características y esquemas de ponderación de términos para representar el texto



Preprocesamiento de texto

Representación del texto

Aprendizaje Supervisado

SEGUNDO Simposio Internacional de Maestría en Ingeniería 2022

Marco Teórico

Extracción de características

La extracción de características tiene el objetivo de reducir la dimensión del vocabulario del conjunto de datos.

Modelo bolsa de palabras: método no le importa cuántas veces aparece una palabra o el orden de las palabras, lo único que importa es si la palabra está presente en una lista de palabras $d = (x_1, x_2, x_3, \ldots, x_n)$.

El tiempo que perdiste por tu rosa hace que tu rosa sea tan importante.

Marco Teórico

- **Esquemas de ponderación de características** Esquemas de ponderación son utilizados para brindar un mayor valor o peso a aquellos términos que son importantes y reducir el peso de aquellos que son menos relevantes.

- **Ponderación TF-IDF (Term Frequency-Inverse Document Frequency):** los términos que aparecen en menos documentos tienden a ser más relevantes que aquellos que aparecen más veces.

$$tfidf(t, d, D) = tf(t, d)\ x\ idf(t, D)$$

Marco Teórico

- **Selección de características**
- Nos permite seleccionar los términos (características) más relevantes de los ya extraídos
- Permite reducir el espacio de características a representar y por consiguiente reducir el costo a nivel procesamiento

- **Chi-Cuadrado:** se aplica para examinar la independencia de dos eventos A y B, en donde, A son los términos y B es la clase de la pregunta.

$$x^2(t, C) = \sum_{t \in 0,1} \sum_{C \in 0,1} \frac{(N_{t,C} - E_{t,C})_2}{E_{t,C}}$$

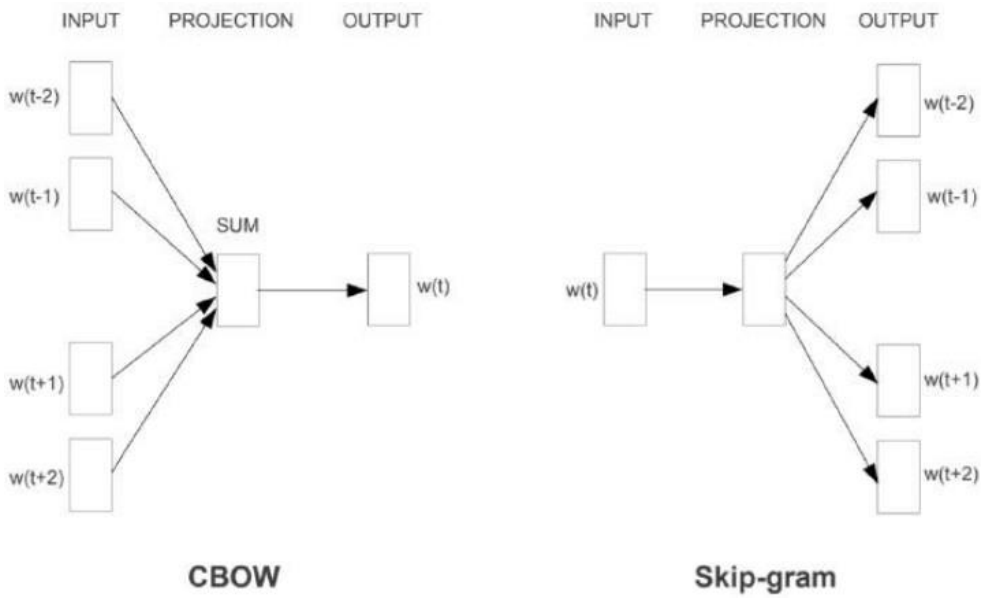
Marco Teórico

Preprocesamiento de texto

Representación del texto

Aprendizaje Supervisado

- **Word2vec**
- Se refiere a un grupo de modelos o arquitecturas desarrolladas por Mikolov et al [4]
- Se utiliza para crear y entrenar espacios vectoriales semánticos que constan de varios de cientos de dimensiones en un corpus de texto
- El modelo word2vec está basado en dos enfoques: Continuous Bag of Words (CBOW) y Skip-gram



Arquitecturas de Word2vec basadas en Mikolov et.al [4]

Preprocesamiento de texto

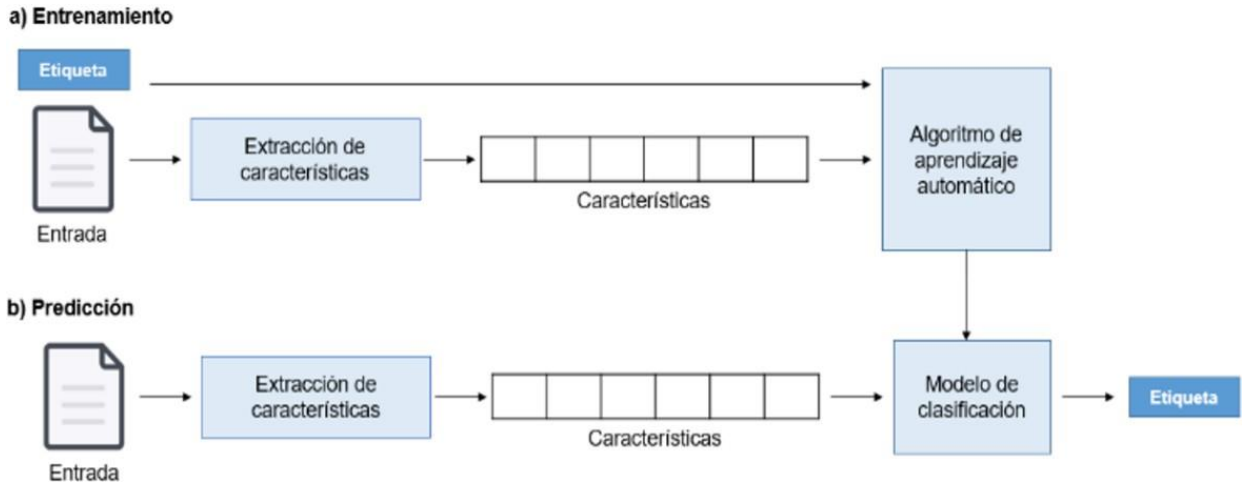
Representación del texto

Aprendizaje Supervisado

Marco Teórico

Se utilizan datos que ya hayan sido etiquetados u organizados previamente para indicar como tiene que ser categorizada la nueva información

En la tarea de clasificación se aplica un algoritmo ya ajustado y entrenado con todo el conjunto de datos que se necesita clasificar

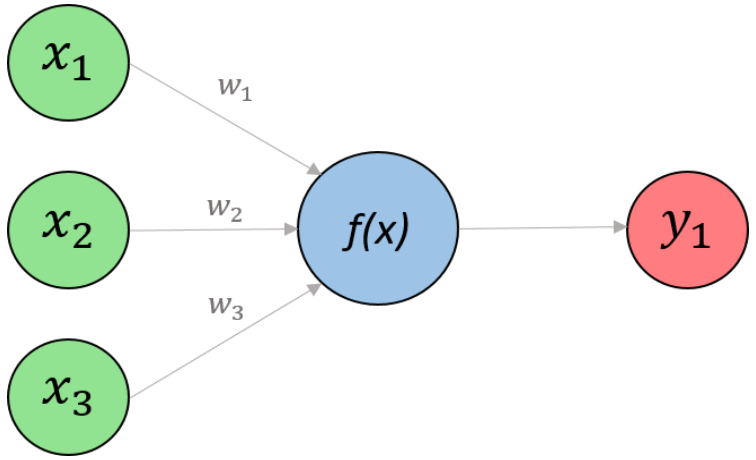


Marco Teórico

- **Red Neuronal Artificial**

Una red neuronal artificial (RNA) es un modelo matemático que intenta simular la estructura y funcionalidad de las redes neuronales biológicas.

La forma más común de una red neuronal es el perceptrón simple que consta de una sola neurona.



$$\sum_{i=1}^{n} w_i \cdot x_i$$

Marco Teórico

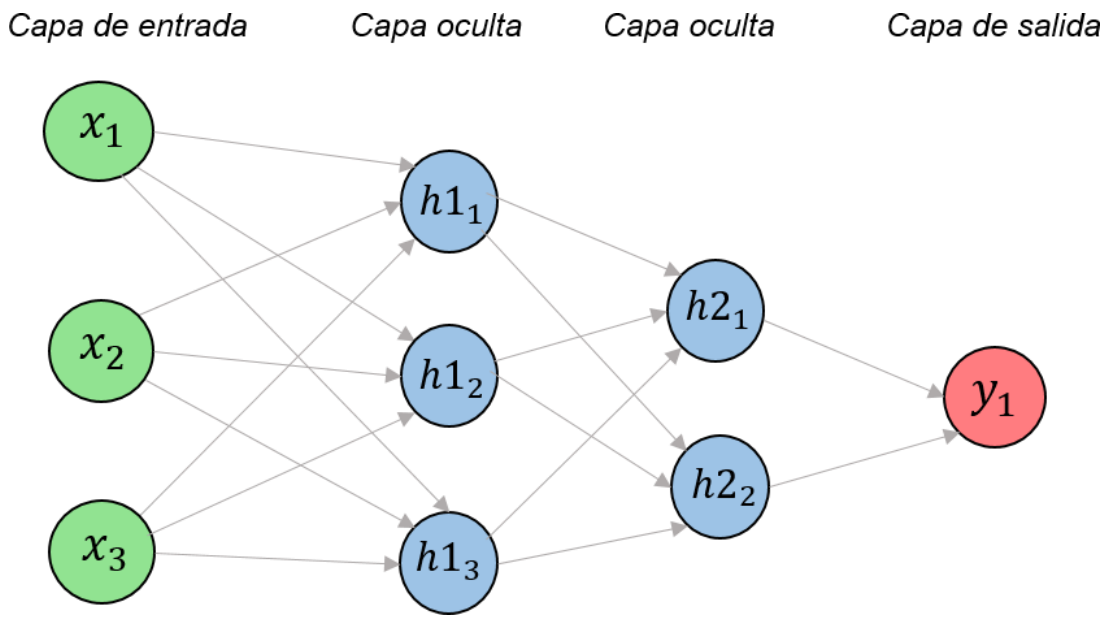
Preprocesamiento de texto

Representación del texto

Aprendizaje Supervisado

- **Perceptrón Multicapa (MLP)**

Es un tipo de red neuronal entrenada con el algoritmo *backpropagation*. MLP esta totalmente conectado y consta de neuronas divididas en capas.



Marco Teórico

- **Red Neuronal Convolucional (CNN)**

Una red convolucional consiste en múltiples capas de filtros convolucionales de una o más dimensiones.

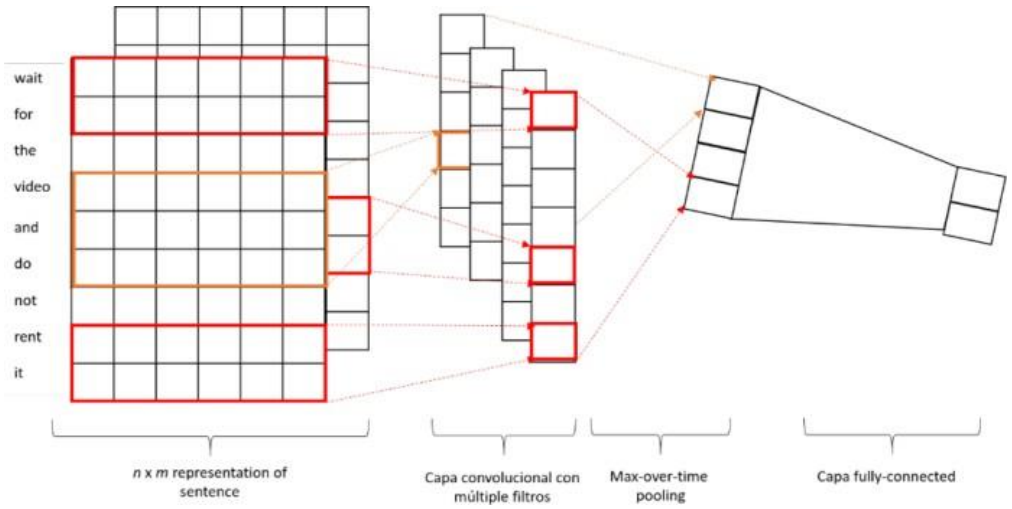
Se compone de tres tipos de capas: convolucionales, pooling y fully-connected.

Las primeras capas pueden detectar líneas, curvas y se van especializando hasta llegar a capas más profundas que reconocen formas complejas como un rostro o la silueta de un animal.

Preprocesamiento de texto

Representación del texto

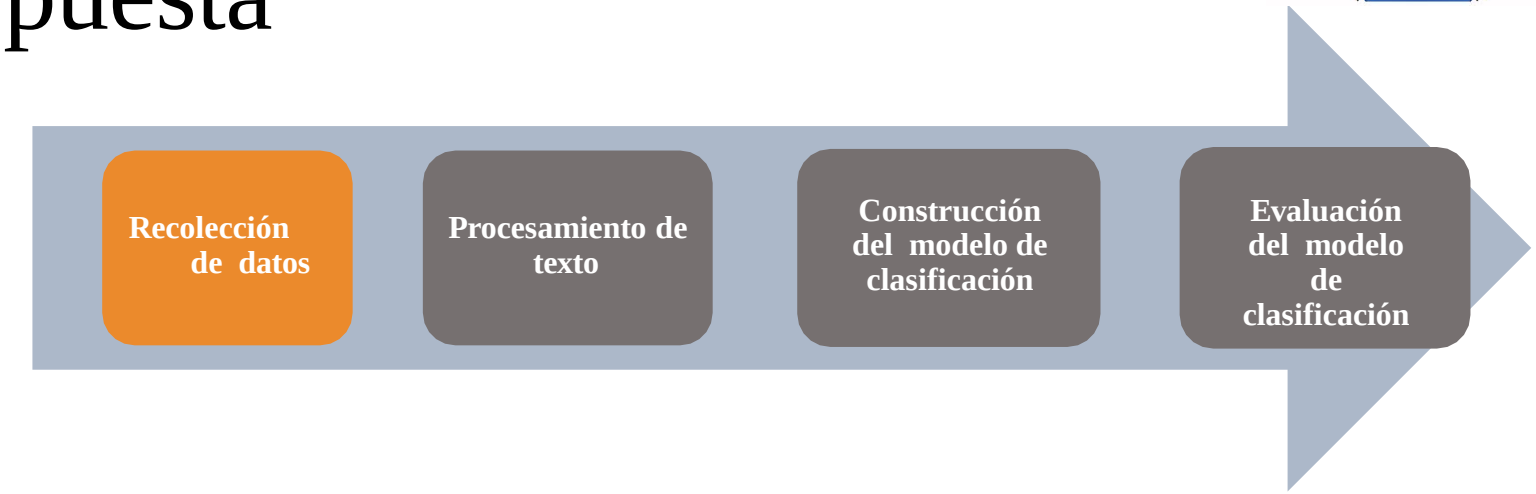
Aprendizaje Supervisado



Basada en Kim[3]

Metodología propuesta

Metodología propuesta



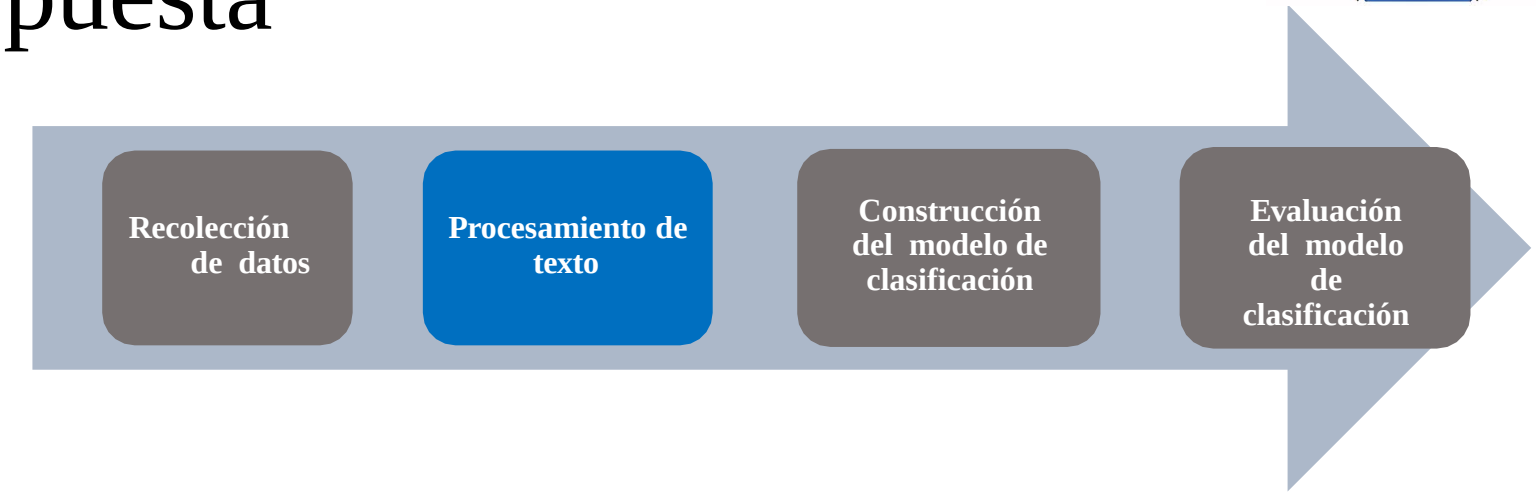
Sitio de comercio electrónico

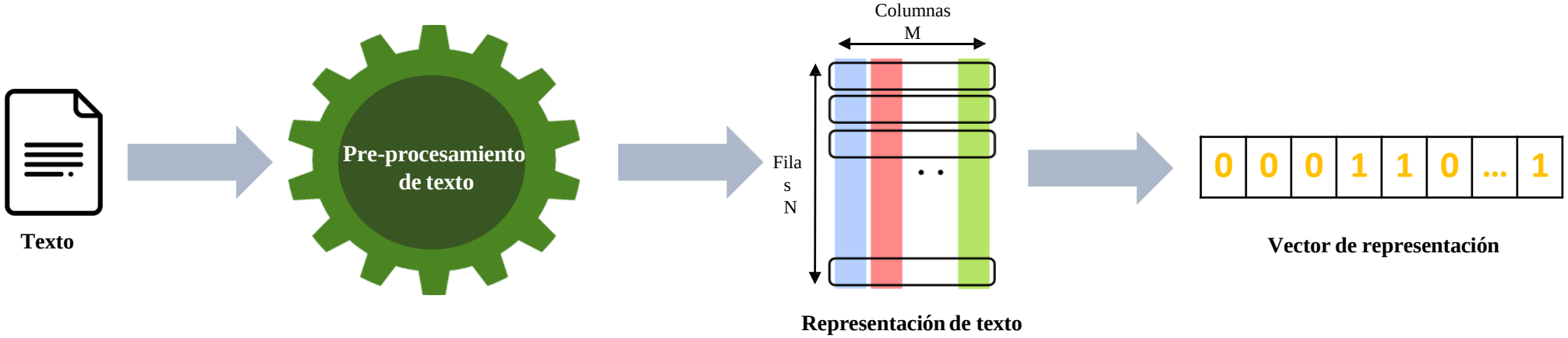
Extracción de preguntas

Etiquetado manual de preguntas

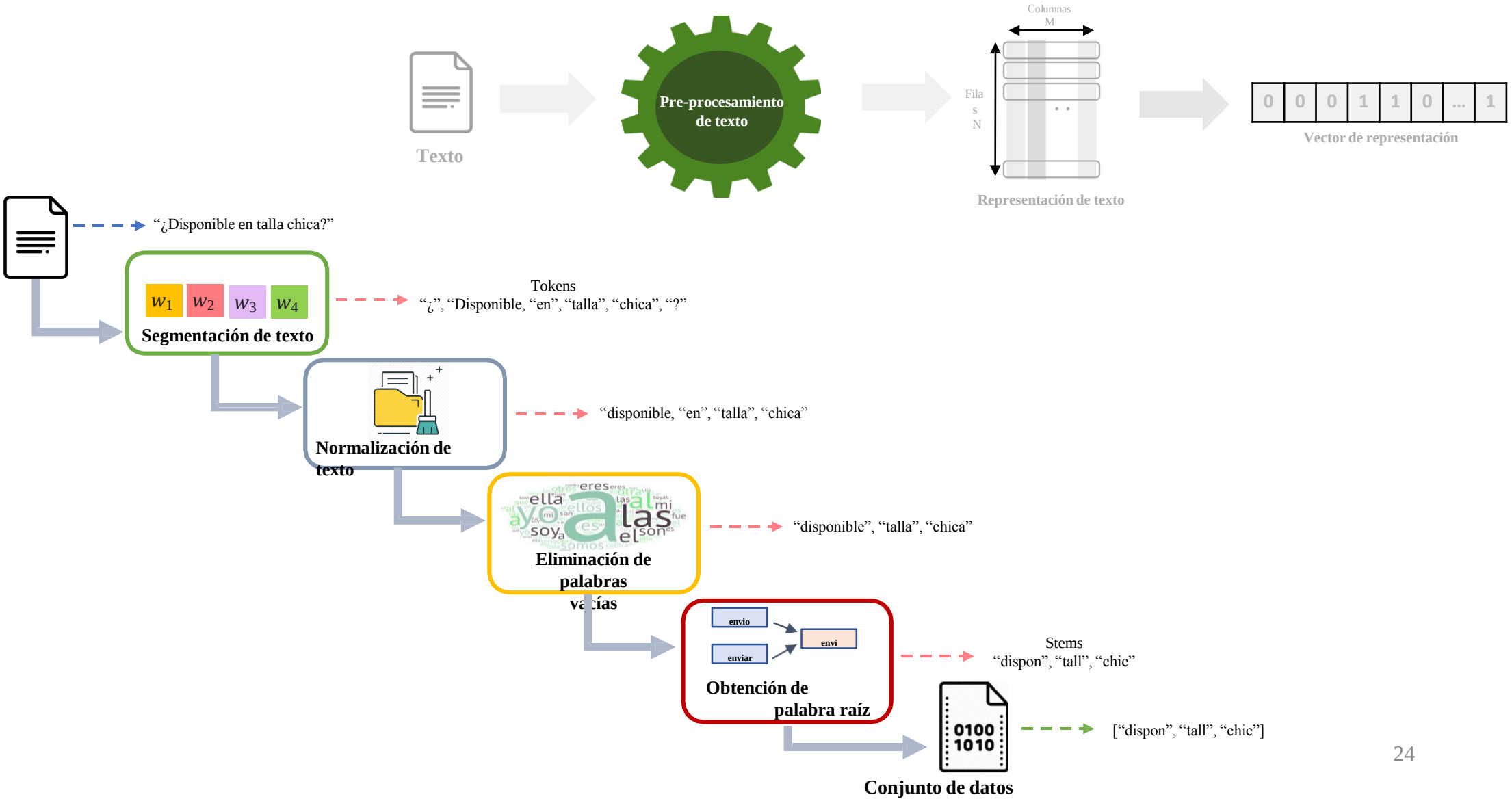
Conjunto de datos etiquetado

Metodología propuesta

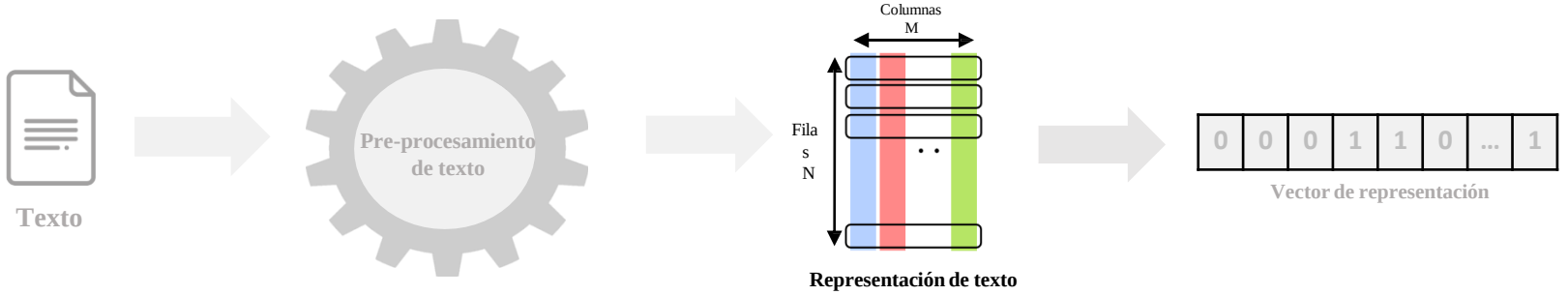


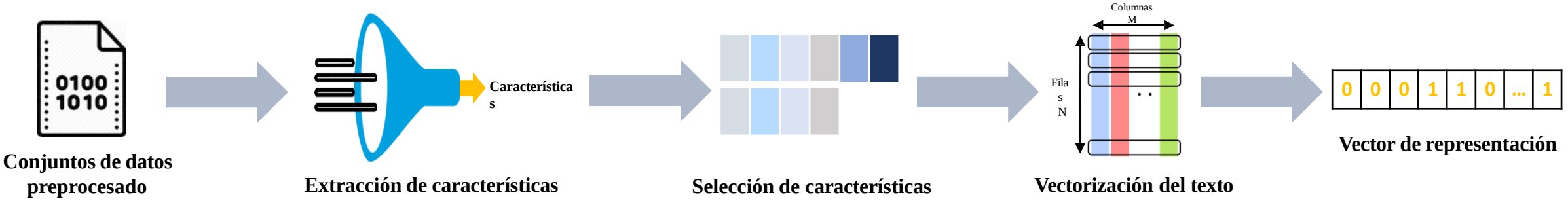


Metodología propuesta



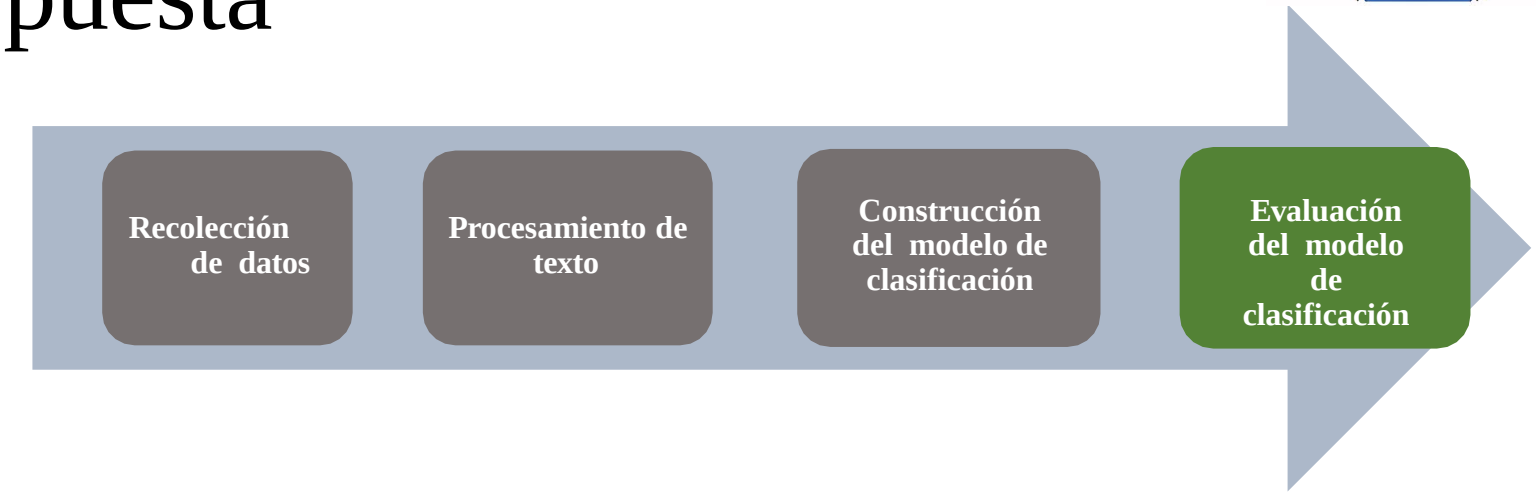
Metodología propuesta

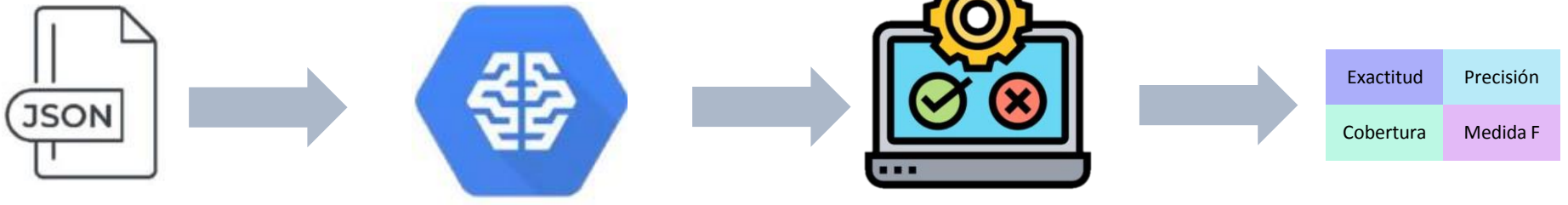




Metodología propuesta

Metodología propuesta





Vectores de representación

Modelo de clasificación de preguntas

Predicción de la clase de las preguntas

Métricas de evaluación del modelo de clasificación

Experimentos

- El conjunto de datos esta compuesto por 500 preguntas extraídas de Mercado Libre México. Las preguntas están distribuidas en dos clases: operativa y técnica.

- El conjunto de datos fue dividido en dos subconjuntos: uno con el 80% preguntas para el entrenamiento y otro con el 20% para las pruebas (predicción).

Distribución de las preguntas del conjunto de datos por clase

No.Clase	Nombre Clase	No. de Preguntas	
		Entrenamiento	Predicción
1	Operativa	203	47
2	Técnica	197	53
Total de preguntas		400	100

Experimentos

Descripción del escenario de evaluación

- Validación cruzada de ”k” iteraciones y se define ”k=10”, es decir, el conjunto de datos de las preguntas se divide en 10 grupos o subconjuntos del mismo tamaño.

- En cada iteración, el conjunto de datos se divide en 80% para entrenamiento y 20% para pruebas.

Experimentos

- **Métricas de evaluación**

$$exactitud = \frac{total\ VP + total\ VN}{total\ ejemplos} \qquad precision = \frac{VP}{VP + FP} \qquad cobertura = \frac{VP}{VP + FN}$$

$$medida - F = \frac{2 * precision * cobertura}{precision + cobertura}$$

VP: Verdaderos Positivos *VN*: Verdaderos Negativos *FP*: Falsos Positivos *FN*: Falsos Negativos

- Se utilizó la matriz de confusión con el objetivo de visualizar el comportamiento de los modelos de clasificación.

Matriz de confusión de 2x2 para dos clases: P y N.

	Predicción Positivo	**Predicción Negativo**
Positivo actual	TP	FP
Negativo actual	FN	TN

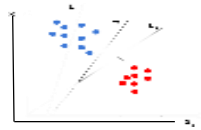
Experimentos

Enfoque aprendizaje automático

- Se utilizaron diferentes técnicas de extracción de características: unigramas hasta cuatrigramas
- Para cada representación se realizó un experimento para diferentes cantidades de características (100 hasta 500)

- Se representaron las características en un espacio vectorial con ponderación TF-IDF
- Se implementaron cuatro diferentes
- Algoritmos de clasificación:
  - Perceptrón Multicapa (MLP) – 8 diferentes configuraciones
  - Naive Bayes (NB)
  - Máquina de Soporte Vectorial (SVM)
  - Árbol de decisión (DT)





Experimentos

Aprendizaje profundo

- En este enfoque se utilizaron las dos diferentes arquitecturas de Word2vec: CBOW y Skip-gram
- Se utilizó la representación de texto basada en unigramas
- El modelo de Word2vec entrenado utiliza un conjunto de 2,000 preguntas y cuenta con 300 dimensiones
- Se utilizó una Red Neuronal Convolucional (CNN) basada en la propuesta por Kim [3]

Resultados finales

- **Unigramas**

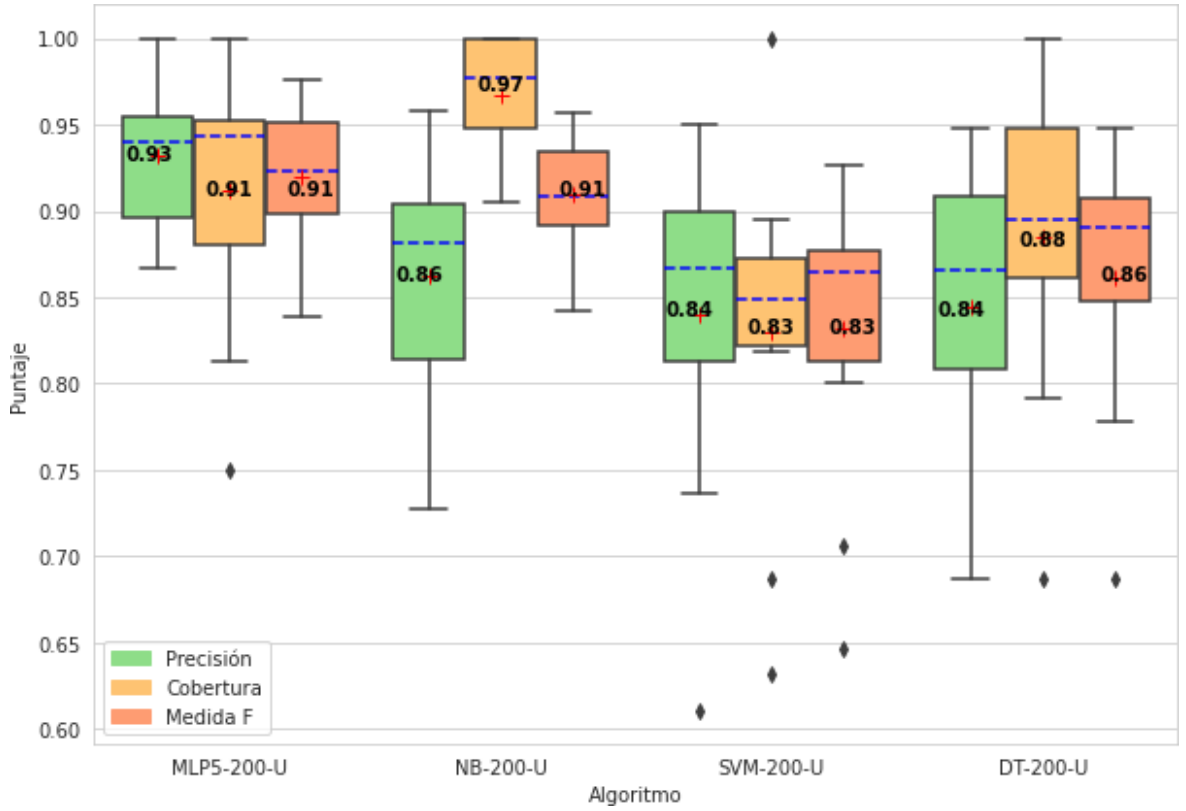
Resultados de configuraciones de redes neuronales MLP para unigramas de 200 características evaluadas mediante validación cruzada

Conf.	Capas Ocultas	No.Neuronas ocultas	Factor aprendizaje	Épocas	Exactitud	Error	Método
MLP1-200-U	1	101	0.001	10,417	0.9150	0.1377	L & F
MLP2-200-U	1	199	0.001	7,604	0.8975	0.1315	Tamura
MLP3-200-U	1	6	0.001	14,671	0.9225	0.1387	Hunter
MLP4-200-U	1	133	0.001	8,123	0.8675	0.1565	RB2
MLP5-200-U	**1**	**20**	**0.001**	**13,780**	**0.9225**	**0.1202**	**PG**
MLP6-200-U	1	115	0.001	9,231	0.9150	0.1220	RB1
MLP7-200-U	1	80	0.001	7,809	0.8150	0.1701	RB1
MLP8-200-U	1	180	0.001	7,717	0.8725	0.1444	RB1

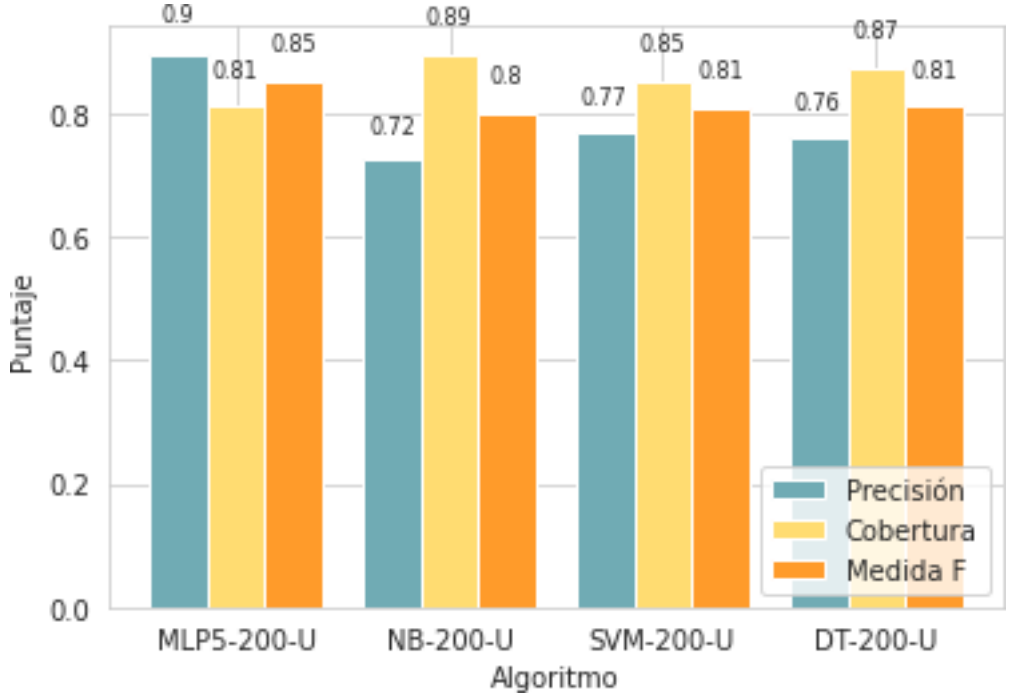
*_L & F:_ Lawrence & Fredrickson;_PG:_ Pirámide geométrica;_RB1:_ Regla Básica 1;
_RB2:_ Regla Básica 2

Resultados finales

Unigramas (entrenamiento)



Unigramas (predicción)



Resultados finales

- **Word2vec**

Resultados de configuraciones de redes neuronales CNN para la clasificación de preguntas

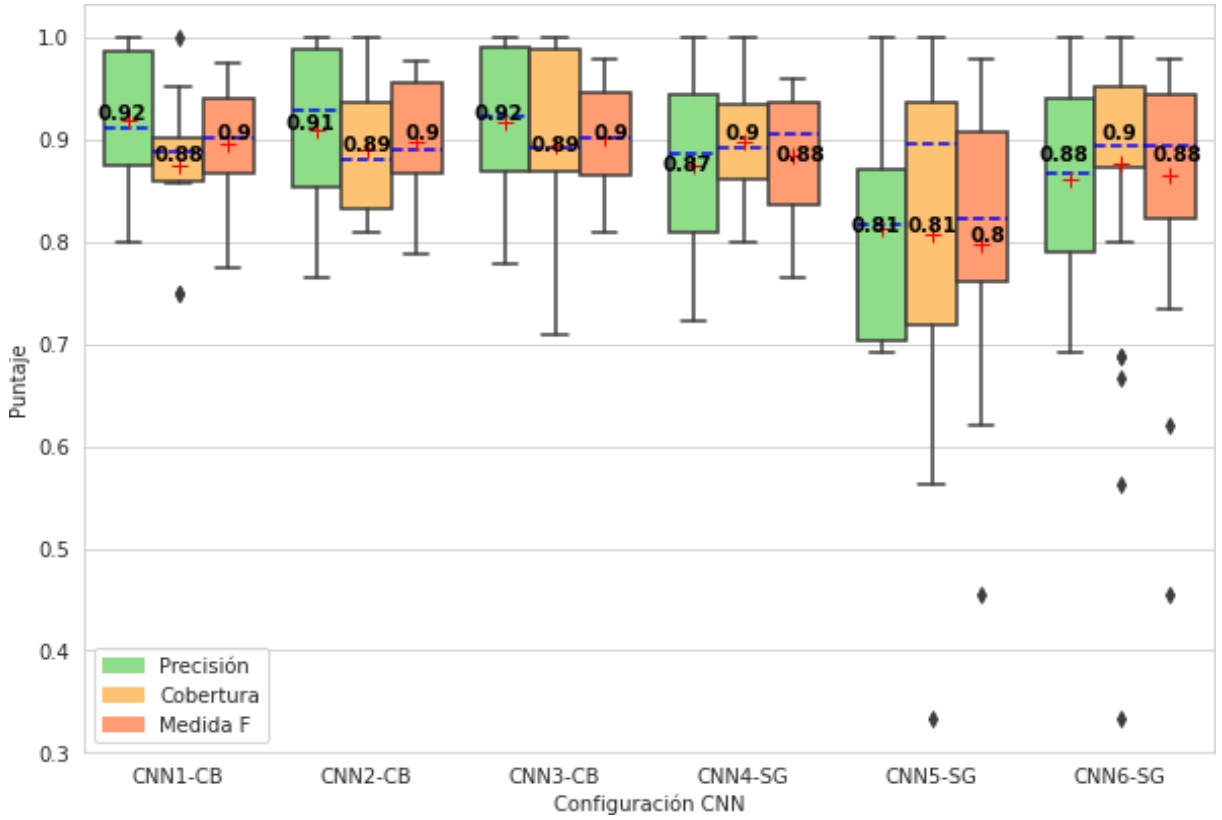
Conf	Tamaño No.filtros	Tamaño filtros	Tamaño embedding	Optimizador	Factor aprendizaje	Épocas	Tamaño batch	Exactitud	Error
CNN1-CB	3,4,5	300	300	adadelta	0.001	30	64	0.903	0.3204
CNN2-CB	3,4,5	300	300	adadelta	0.001	25	50	0.903	0.3252
CNN3-CB	**3,4,5**	**300**	**300**	**adadelta**	**0.001**	**30**	**50**	**0.908**	**0.3390**
CNN4-SG	3,4,5	300	300	adadelta	0.001	30	64	0.888	0.3735
CNN5-SG	3,4,5	300	300	adadelta	0.001	25	50	0.823	0.4840
CNN6-SG	3,4,5	300	300	adadelta	0.001	30	50	0.895	0.3705

**CB*: Continuous Bag of Words (CBOW)

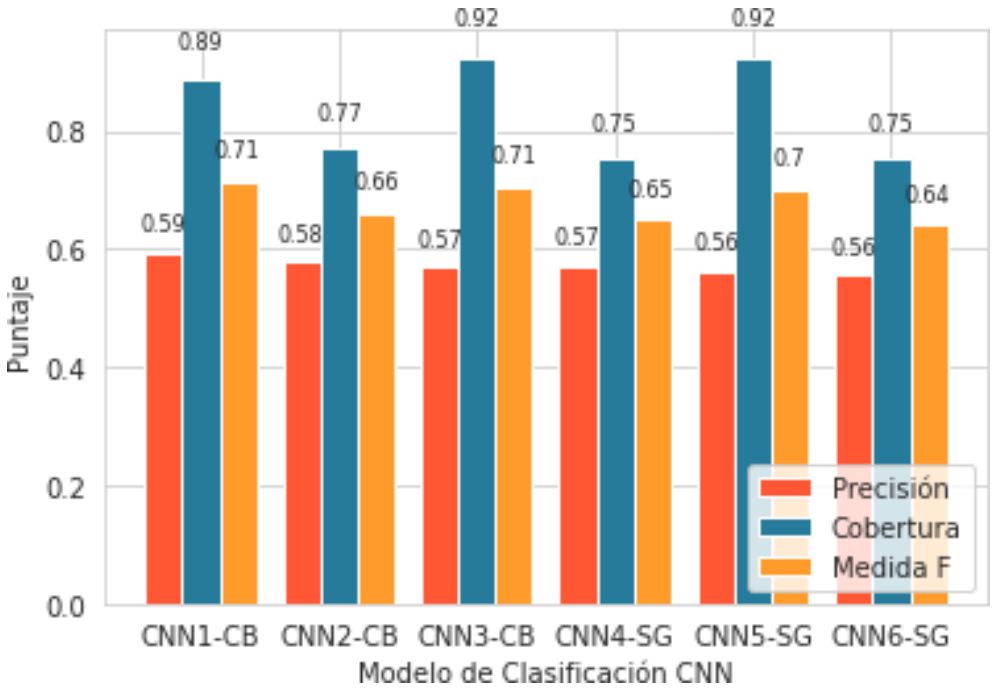
**SG*: Skip-gram

Resultados finales

Word2vec (entrenamiento)

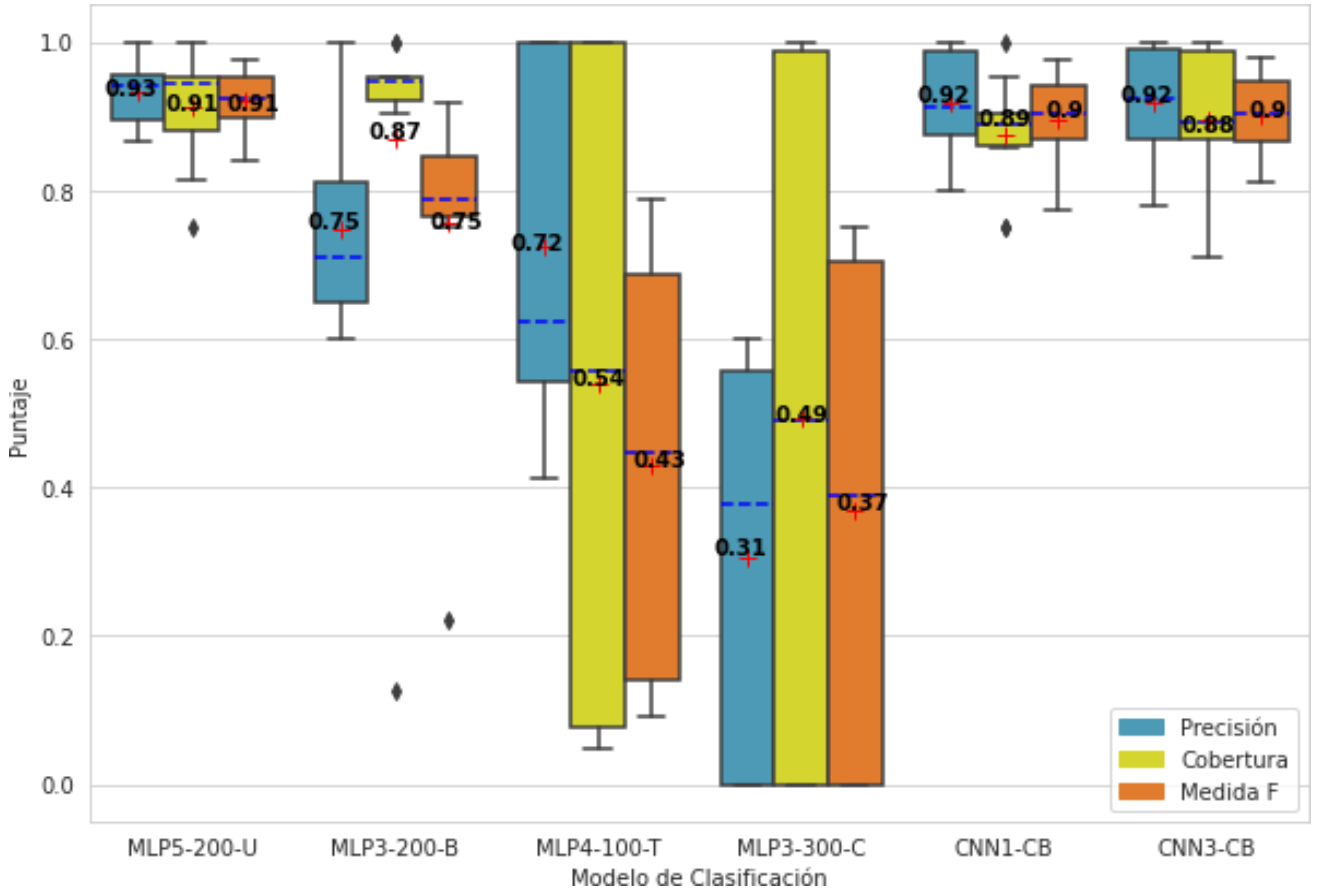


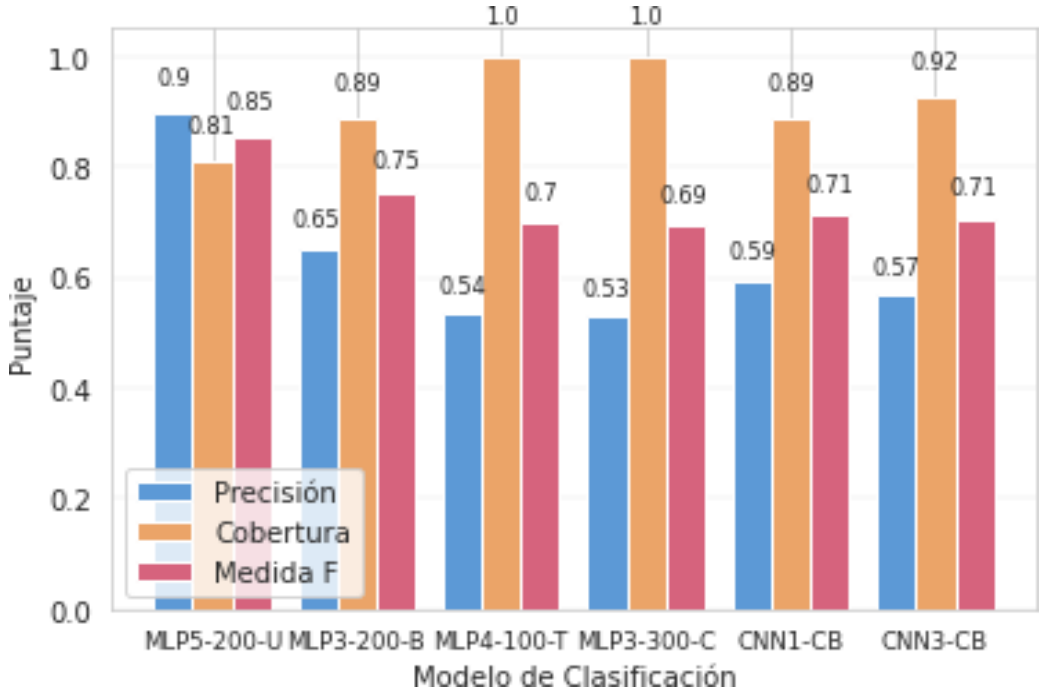
Word2vec (predicción)



Resultados finales

- **Resumen**





Resultados finales

- **Análisis Cualitativo**

Características léxicas más representativas en los unigramas

Característica léxica	Ejemplo	Total
Sustantivo	chaleco	492
Verbo	enviar	107
Adjetivo	bueno	38
Total		637

Ejemplos de preguntas clasificadas incorrectamente por los modelos de clasificación entrenados con una representación basada en unigramas

Pregunta	Clase Real	Clase Predicha
Hola, ¿cuáles son las medidas del audífono?	Técnica	Operativa
Gracias amigo llego el paquete todo bien, están bien padres, solo una duda ¿vendes solo los discos?	Operativa	Técnica
Hola, me interesa ¿tienes disponible ?, ¿se puede usar con un Xperia XA Ultra? , gracias	Operativa	Técnica
¿Qué tipo de chip debo usar?	Técnica	Operativa
¿Cuántos días de garantía tiene amigo?	Técnica	Operativa
Entonces, ¿serian 94 pesos en total ya con el envío?. De la batería del note 4 910v de Verizon	Operativa	Técnica
¿Cuál es la longitud ?	Técnica	Operativa

Resultados finales

- **Análisis Cualitativo**

Características léxicas más representativas en los bigramas

Colocación léxica	Ejemplo	Total
Verbo + Sustantivo	enviar Acapulco	452
Sustantivo + Sustantivo	colchón inflable	417
Sustantivo + Adjetivo	entrega inmediata	283
Verbo + Adjetivo	tendrá disponible	64
Verbo + Adverbio	pediría ahorita	40
Adjetivo + Sustantivo	disponibilidad aparato	38
Determinante Indefinido + Sustantivo	cuanto cuesta	25
Sustantivo + Verbo	color maneja	23
Total		1,342

Ejemplos de preguntas clasificadas incorrectamente por los modelos de clasificación entrenados con una representación basada en bigramas

Pregunta	Clase Real	Clase Predicha
Disculpa, ¿no tienes disponible la mica cristal para el iPhone 8?	Operativa	Técnica
¿Tienes disponible los modelos 4 ,5 y 6 ?	Operativa	Técnica
Buenas noches , ¿tienes disponible en color negro de 10 oz ? Para	Operativa	Técnica
¿Tienes disponible?	Operativa	Técnica
¿Y cuánto aumenta del envío?	Operativa	Técnica
Buenas tardes, ¿funciona para, gracias monitor de 15“?, gracias	Técnica	Operativa
¿En cuántos días hacen el envío a Acapulco?	Técnica	Operativa
Hola, buenas tardes, disculpa ¿funciona para el Modelo SM-G531H?, ¿Cuánto tarda en llegar?	Técnica	Operativa

Conclusiones

- Se presentó un método para el diseño de un modelo de representación y clasificación de textos cortos en el idioma español.
- Se aplicó un enfoque de aprendizaje automático tradicional y otro de aprendizaje profundo.
- Con base en los experimentos realizados se confirma la hipótesis de que las redes neuronales permiten obtener mejores resultados.
- Aplicar un enfoque de aprendizaje profundo que permita obtener modelos de clasificación por encima del 90 % de exactitud necesita un conjunto de datos mucho más grande.

ECORFAN®