Benemérita Universidad Autónoma de Puebla

Facultad de Ciencias de la Computación

Doctorado en Ingeniería del Lenguaje y del Conocimiento

Método para el descubrimiento de tópicos con aprendizaje profundo

---

Presenta
Ana Laura Lezama Sánchez

Director
Dra. Mireya Tovar Vidal

Co-director
Dr. José Alejandro Reyes Ortiz

Enero 2024

Agradecimientos

- A CONAHCYT por la beca otorgada para obtener el grado académico de Doctorado en Ingeniería del Lenguaje y del Conocimiento con la beca número 788155.
- Al Laboratorio Nacional de Supercómputo del Sureste de México (LNS), perteneciente al padrón de laboratorios nacionales CONAHCYT, por los recursos computacionales, el apoyo y la asistencia técnica brindados, a través del proyecto No. 202103090c.
- Al Laboratorio Nacional de Supercómputo del Sureste de México (LNS), perteneciente al padrón de laboratorios nacionales CONAHCYT e INAOE, por los recursos computacionales, el apoyo y la asistencia técnica brindados.
- A mis directores de tesis la Dra. Mireya Tovar Vidal y al Dr. José Alejandro Reyes Ortiz por su respaldo durante todo este proceso.
- A mis revisores el Dr. José de Jesús Lavalle Martínez, a la Dra. Claudia Zepeda Cortés y a la Dra. Maricela Claudia Bravo Contreras por su colaboración y su trabajo como miembros del comité revisor.

Índice general

Índice de figuras

Índice de tablas

Resumen

En internet existe una cantidad masiva de información de la que es posible extraer datos importantes para algún propósito particular. Sin embargo, analizar o extraer información manualmente no es una tarea fácil de realizar. Analizar mucha información disponible en internet es una tarea difícil y obtener información relevante sigue siendo una tarea costosa. Por lo que, una de las áreas de la Inteligencia Artificial (IA) con la que es posible obtener la información esencial de un documento es a través del Procesamiento del Lenguaje Natural (PLN) en particular a través de la tarea del descubrimiento de tópicos. Esta tarea permite extraer la idea central de un documento; también puede ser aplicado a datos masivos de información, permitiendo descubrir la idea central de una colección de documentos.

Por lo tanto, en esta tesis doctoral se propone el desarrollo de un método para el descubrimiento de tópicos en los córpora *20-Newsgroup* y *Reuters*. El método consta de 5 fases. Las cuales están compuestas de: pre-procesamiento de textos, extracción de relaciones semánticas, construcción del modelo de incrustación, clasificación de texto y descubrimiento de tópicos. La evaluación de la clasificación se realiza con las métricas de precisión, *accuracy, recall* y medida-F_1, y el descubrimiento de tópicos con la métrica de coherencia del tópico normalizada.

Los mejores resultados de la evaluación de la clasificación utilizando el modelo de incrustación de hiponimia-hiperonimia fueron para el corpus *20-Newsgroup* que obtuvo un *accuracy* de 0.79. El corpus *Reuters*, obtuvo una medida-F_1 y un *recall* de 0.87 utilizando el modelo de incrustación con las tres relaciones semánticas. Para los córpora para el descubrimiento de tópicos los mejores resultados obtenidos fueron al descubrir 20 tópicos con 10 palabras representativas. El corpus *20-Newsgroup* obtuvo una coherencia del tópico normalizada de 0.1723 con *LDA*, 0.1622 con *LSA* y 0.1716 con *PLSA*. Para el corpus *Reuters* los mejores resultados fueron 0.1441 con *LDA*, con *LSA* 0.1360 y con *PLSA* 0.1436.

Capítulo 1

Introducción

En este capítulo se presentan las bases que forman este trabajo de investigación: introducción al problema de investigación propuesto, así como el planteamiento del problema, el objetivo general y los específicos. Además, la justificación, los antecedentes y la pregunta de investigación e hipótesis. Finalmente se expone la organización del documento de tesis.

1.1. Planteamiento del problema

Actualmente existe información masiva en internet lo que dificulta su análisis de manera manual. Por lo que en los últimos años contar con recursos computacionales que analicen la información automáticamente en tiempos de respuesta cortos se ha convertido en una herramienta útil.

El Procesamiento del Lenguaje Natural (PLN) estudia la relación entre el lenguaje humano y las computadoras [78]. Una de las áreas de estudio del *PLN* es el descubrimiento de tópicos que permite descubrir la idea central en un documento.

El descubrimiento de tópicos implica identificar las ideas principales dentro de un documento de texto. Los cuales indican tópicos recurrentes en los documentos, proporcionando una descripción general del texto. Los modelos actuales de descubrimiento de tópicos reciben el texto, con o sin preprocesamiento previo, como la eliminación de palabras vacías, etiquetas html, símbolos de puntuación, espacios en blanco adicionales y conversión a minúsculas. Un proceso de descubrimiento de tópicos que recibe texto de dominio general, es decir, sin una clasificación o agrupamiento previo, proporciona tópicos generales. Los tópicos generales no ofrecen descripciones

detalladas del texto ingresado y su categorización manual es tediosa y requiere mucho tiempo. Por lo que es necesario extraer de un texto previamente clasificado tópicos específicos formados con palabras que mantengan una mayor relación entre ellas.

Actualmente la representación de textos en lenguaje natural a nivel computacional puede llevarse a cabo utlizando un modelo de incrustación de palabras. Lo que ha permitido mejorar significativamente el descubrimiento de conocimiento en diversas tareas del *PLN* por ejemplo la clasificación de texto, recomendación de contenido o sistemas de recomendación y sistemas de preguntas-respuestas. Sin embargo, en el estado del arte la existencia de modelos de incrustación que incorporen semántica extraída de un base de conocimiento solo existe uno denominado *wnet2vec* [109]. Dicho modelo ha sido evaluado con la tarea de similitud semántica otorgando buenos resultados en comparación *word2vec* utilizado para el mismo fin.

La generación de modelos que incorporen semántica a las palabras que forman una oración permite a una computadora obtener información adicional. Lo que proporcionará resultados que apoyarán las necesidades del usuario. Las relaciones semánticas son parte integral en una oración proporcionando coherencia y generando ideas completas en los textos. Algunas de las relaciones semánticas existentes son hiponimia, hiperonimia y sinonimia.

Por lo tanto, es importante contar con un método para el descubrimiento de tópicos que genere tópicos específicos del corpus a analizar.

En este trabajo de tesis doctoral se plantea el desarrollo de un método para el descubrimiento de tópicos basado en la representación de los textos con relaciones semánticas de sinonimia, hiponimia e hiperonimia y aprendizaje profundo. El método incluye la clasificación de los córpora utilizando una red neuronal convolucional. La representación de los textos se realizó con un modelo de incrustación de relaciones semánticas. El resultado proporcionado por la red neuronal convolucional son las clases existentes en cada corpus. Posteriormente se descubren los tópicos de cada una de las clases que pertenecen a cada uno de los córpora. En este punto un modelo clásico utilizado en la literatura para descubrir tópicos es capaz de descubrir tópicos específicos, ya que el texto que recibe son clases previamente obtenidas por una red neuronal convolucional.

Por lo que este método está formado por 5 fases:

1. **Pre-procesamiento de textos:** en esta fase se descargan los córpora *20-Newsgroup* y *Reuters* para el descubrimiento de tópicos. Además de un corpus

de *Wikipedia* en inglés de 4,500,000 de documentos para la extracción de las relaciones semánticas.

2. **Extracción de relaciones semánticas:** en esta fase se genera un repositorio de patrones léxico sintácticos de relaciones de sinonimia, hiponimia e hiperonimia. Los cuales se convierten en expresiones regulares en Python.

3. **Construcción de modelo de incrustación:** en esta fase se desarrollan tres matrices con las relaciones semánticas previamente extraídas. La primera matriz representa las relaciones de sinonimia, la segunda las relaciones de hiponimia-hiperonimia y la tercera una combinación de las tres.

4. **Clasificación de texto:** esta fase involucra la tarea de clasificación de textos implementando una red neuronal convolucional y como representación de los textos los modelos de incrustación generados en la fase 3. La evaluación de esta fase se realiza con las métricas precisión, exhaustividad, medida-F_1 y exactitud.

5. **Descubrimiento de tópicos:** en esta fase el corpus de clases obtenido en la fase 4 son los datos de entrada en los modelos de descubrimiento de tópicos. Lo que genera tópicos particulares. La evaluación de los tópicos descubiertos se realiza con la métrica de coherencia del tópico normalizada.

1.2. Objetivos

A continuación, se presentan los objetivos generales y específicos de esta tesis doctoral.

1.2.1. Objetivo general

Proponer un método de descubrimiento de tópicos por medio de aprendizaje profundo e incrustación de palabras para la identificación de tópicos que están presentes en una colección de documentos.

1.2.2. Objetivos específicos

- Implementar un módulo computacional para el pre-procesamiento de los textos utilizados para el descubrimiento de tópicos.
- Desarrollar un método para la extracción automática de relaciones semánticas de tipo sinonimia, hiponimia e hiperonimia en la colección de documentos.
- Aplicar e implementar un modelo de incrustación de palabras con relaciones semánticas previamente identificadas.
- Proponer un método de descubrimiento de tópicos basado en aprendizaje profundo y la representación de los textos.
- Evaluar el método de descubrimiento de tópicos con la finalidad de comparar los resultados obtenidos con los trabajos existentes en la literatura.

1.3. Justificación

El aumento de la información disponible en internet dificulta la posibilidad de llevar a cabo un análisis manual. Por lo que el desarrollo de métodos que sean capaces de llevar a cabo esa tarea en poco tiempo ha sido de interés por especialistas en el área. En lingüística un tópico se define como la idea básica de un texto. El proceso de identificarlo puede ser una pieza clave en alguna aplicación que involucre el descubrimiento de tópicos [29]. Los trabajos relacionados sobre descubrimiento de tópicos con el uso de aprendizaje profundo e incrustación de palabras son analizados en esta tesis doctoral. Además de las técnicas de descubrimiento de tópicos tradicionales, y las técnicas de extracción de relaciones semánticas. Montoya [86] señala que incorporar aprendizaje profundo e incrustación de palabras proporciona resultados alentadores.

Por otro lado, Lezama Sánchez et al. [71, 70] señalan que las relaciones semánticas de sinonimia, hiponimia e hiperonimia permiten incorporar conocimiento adicional sobre una palabra en un texto mejorando los resultados esperados. Por lo que resulta pertinente abordar desde la perspectiva de la Ingeniería de Lenguaje y del Conocimiento, una solución basada en un modelo de aprendizaje profundo, incorporando relaciones semánticas para el desarrollo de un modelo de incrustación de relaciones.

Además de un algoritmo tradicional para el descubrimiento de tópicos. Este último toma como textos de entrada las clases obtenidas al incorporar un modelo de aprendizaje profundo. Los resultados son tópicos particulares que son clave importante en el desarrollo de una aplicación que necesite extraer este tipo de información en textos de gran tamaño y sin una clasificación previa.

1.4. Antecedentes

El aumento exponencial de la información ha generado la necesidad de contar con herramientas computacionales que lleven a cabo este proceso. El descubrimiento de tópicos es un área en constante desarrollo. Los métodos de descubrimiento de tópicos son las herramientas computacionales que llevan a cabo este proceso sólo proporcionando un conjunto de datos pre-procesados. El resultado será un conjunto de tópicos contenidos en el córpora, que posteriormente podrán ser el eje central en aplicaciones que impliquen el descubrimiento de los tópicos en un conjunto de datos para el desarrollo de posteriores mecanismos. Algunas investigaciones como en Fuentes-Pineda et al. [33] proponen un método para el descubrimiento de tópicos basándose en *min-hashing*. El método consiste en estimar la similitud entre dos conjuntos. Los autores no proporcionan el número de tópicos por descubrir, sino que su método es capaz de generarlos de manera automática. Los autores realizan pruebas con córporas en inglés y español. En cambio, en Srivastava et al. [125] presentan un método para la misma tarea pero incorporando un *autoencoder*. Los autores experimentan únicamente con textos en el idioma inglés y hacen uso del algoritmo de agrupamiento k-means y similitud coseno.

1.5. Pregunta de investigación

1. ¿De qué manera benefician el aprendizaje profundo y un modelo de incrustación de palabras basado en relaciones semánticas en la tarea de descubrimiento de tópicos?

1.6. Hipótesis

1. La calidad de un método de descubrimiento de tópicos que incluye la tarea de clasificación y basado en aprendizaje profundo depende de la representación de los textos considerando las relaciones semánticas.

1.7. Organización del documento

El resto del documento está organizado de la siguiente manera: En el capítulo 2 se realiza una revisión de los conceptos necesarios para introducir el problema de investigación. Los niveles del procesamiento del lenguaje natural, métodos de descubrimiento de tópicos, aprendizaje profundo e incrustación de palabras. Así como las métricas de evaluación para la tarea de clasificación y descubrimiento de tópicos. En el capítulo 3 se lleva a cabo una revisión del estado del arte con los enfoques de descubrimiento de tópicos basado en incrustación de palabras, aprendizaje profundo y relaciones semánticas. Así mismo en el capítulo 4 se describe la metodología de solución propuesta para el desarrollo del método de descubrimiento de tópicos. Además, en el capítulo 5 se presentan los resultados experimentales de esta investigación. Finalmente, en el capítulo 6 se incluyen las conclusiones y posteriormente las referencias consultadas para el desarrollo de este proyecto.

Capítulo 2

Marco teórico

En este capítulo se presentan y explican los conceptos teóricos que darán sustento al presente trabajo de tesis. Los cuales se basan en el Procesamiento del Lenguaje Natural (PLN), los niveles de procesamiento del PLN, las relaciones semánticas, los modelos de representación de texto, el aprendizaje profundo y automático para la clasificación de texto y 4 métricas usadas para su evaluación. Además, de los modelos de descubrimiento de tópicos más empleados en la literatura y una métrica para la evaluación de los tópicos.

2.1. Procesamiento del Lenguaje Natural

El área de lingüística computacional (LC) se encuentra formada por diferentes disciplinas y estudia la interacción entre el lenguaje natural y las computadoras denominado también como Procesamiento del Lenguaje Natural (PLN) [88].

El *lenguaje natural (LN)* es descrito como el medio usado para establecer comunicación entre personas (lenguaje). El LN permite nombrar a cosas o personas y razonar sobre ellas. Su sintáxis se modela por medio de un lenguaje formal, similar a las matemáticas y la lógica [136].

Por lo tanto, el lenguaje es considerado como un sistema de signos usados por el ser humano para comunicarse en su día a día [24]. El cual se expresa por medio del sonido o símbolos denominados como código oral y escrito.

Por lo que el PLN consiste en la capacidad de una máquina para procesar el lenguaje natural, es decir, el PLN utiliza una expresión en lenguaje natural (LN) que pueda tener comunicación con la computadora de manera natural por medio de

voz o texto [88]. El *PLN* incluye 5 niveles de estudio necesarios para que la extracción y comprensión de la información se realicen de manera correcta, debido a su relación entre sí [108].

A continuación se mencionan y describen los niveles del *PLN*. Posteriormente algunas tareas donde es necesario el tratamiento del lenguaje natural. Por último dada la complejidad del *LN* existen diferentes aplicaciones donde se aplican algunas tareas del *PLN*.

2.1.1. Niveles del procesamiento del lenguaje natural

La arquitectura de un sistema de *PLN* se rige en la definición del *LN* por niveles: morfológico, sintáctico, fonológico, semántico, y pragmático [136]. Los cuáles serán descritos detalladamente a continuación:

Nivel morfológico

El nivel morfológico estudia la estructura interna de las palabras, su propósito es identificar una palabra en particular. Este nivel consiste en determinar que categorías morfológicas posee cada palabra, así como sus características, su forma base o lema [23].

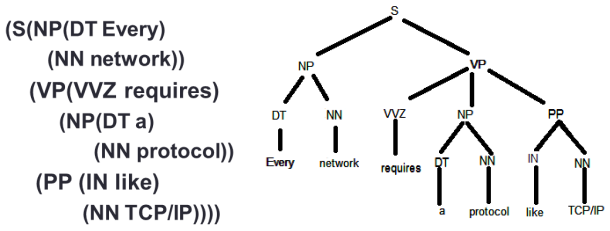
Por lo que existen tres procesos morfológicos que son el morfema, *stemming* y lematización, que serán descritos a continuación:

1. morfema: unidad lingüística más pequeña que tiene un significado
2. *stemming*: transformar una palabra dada en su forma canónica o raíz [23]
3. lematización: indica la acción de encontrar el lema de una palabra.

La diferencia entre un morfema y un lema es que un morfema puede tener diferentes significados, pero no un lema.

Nivel sintáctico

El nivel sintáctico, llamado *parsing* estudia las relaciones estructurales presentes entre palabras dentro de una oración. Un conjunto de reglas permite a un lector reconocer el significado de las palabras. Por lo que dependiendo de su lugar en la



(a) Forma de lista (b) Forma de árbol

Figura 2.1: Ejemplo de la estructura del nivel sintáctico. Imagen tomada de [107].

oración y sus palabras vecinas, pueden contar con diferentes categorías gramaticales. Además, proporciona información sobre la forma en cómo suena una palabra siendo útil en tareas del *PLN*. Algunas de ellas: recuperación de información (*RI*) y desambiguación del sentido de las palabras [23].

Una oración se compone de diferentes sub-oraciones lingüísticas. Un ejemplo de estos tipos de frases: frase sustantiva (*NP*), frase verbal (*VP*), frase preposicional (*PP*) y frase adjetiva (*AP*) [107].

La estructura de una sentencia puede representarse como una lista o árbol. En [107] proporcionan un ejemplo de la sentencia *"Every network requires a protocol like TCP/IP"* que se muestra en la Figura 2.1(a) en forma de lista y en forma de árbol 2.1(b).

Nivel semántico

El nivel semántico define el significado de las oraciones y las palabras que la forman [107]. Además, se basa en el estudio de las reglas y principios sobre la creación de expresiones sintácticas que se puedan interpretar a partir de expresiones más simples [23]. La representación de la semántica puede llevarse a cabo mediante predicados lógicos, grafos dirigidos o marcos semánticos [23]. A continuación, se presenta un ejemplo con una oración de las representaciones indicadas.

La Figura 2.2 expone el grafo dirigido de la oración: *A computer is a machine that has a processor* (Una computadora es una máquina que tiene un procesador). La Tabla 2.1 muestra la representación mediante lógica de predicados y la Tabla 2.2 mediante marcos semánticos, del significado de la oración: *A computer is a machine that has a processor* (Una computadora es una máquina que tiene un procesador).

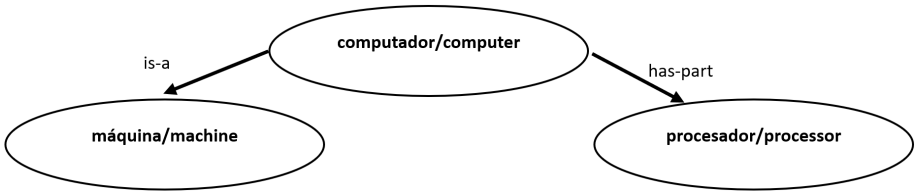


Figura 2.2: Representación mediante grafo dirigido. Imagen tomada de [23].

Tabla 2.1: Representación mediante lógica de predicados. Imagen tomada de [23].

is a (computer, machine)
has-part (computer, processor)

Tabla 2.2: Representación mediante marcos semánticos. Imagen tomada de [23].

computer *isa: machine*
has-part: processor

Nivel pragmático

El nivel pragmático establece la identidad de las personas y objetos de las que se narra en un texto. La estructura del discurso determina este nivel y gestionan el diálogo en una conversación [138].

El análisis de los niveles del *PLN* en un corpus puede ser desarrollado aplicando las herramientas proporcionadas por el lenguaje de programación Python.

2.2. Relaciones semánticas

Las relaciones semánticas pertenecen al nivel de análisis semántico el cual se expuso en 2.1.1.

Las relaciones semánticas son segmentos que pueden ser usados en varias tareas como formación de conceptos, establecer jerarquías y definir relaciones no jerárquicas. Además, se relacionan de acuerdo con su significado [40]. Algunas relaciones semánticas son: hiperonimia, hiponimia, meronimia, sinonimia y antonimia que serán

descritas en las siguientes secciones.

2.2.1. Hiponimia e hiperonimia

Las relaciones semánticas de hiponimia se define como la relación de inclusión que se establece entre unidades léxicas de la misma categoría gramatical dentro de áreas conceptuales determinadas [40]. La hiperonimia es una relación que se establece entre una palabra de carácter más general y otra de carácter más específico [40].

Por ejemplo, Tierra es un hipónimo de Planeta y árboles es un hiperónimo de árbol de: aguacate, durazno, café [82].

Algunos de los métodos existentes en la literatura para la extracción automática de este tipo de relación se basan en:

1. Diccionarios: Para el desarrollo de este método es necesario contar con un diccionario como *Wordnet* [40].

2. Agrupamiento: Este proceso emplea un algoritmo de agrupamiento bajo la premisa de que palabras similares comparten contextos similares [40].

3. Patrones léxico sintácticos: Este proceso emplea expresiones del lenguaje natural que están formados de frases nominales y conceptos del léxico de dominio [40]. Hearst [45] expone algunos de los patrones más empleados denominados patrones léxico sintácticos para extraer relaciones entre conceptos. Los patrones léxico sintácticos son más sólidos para representar propiedades de clase que las características basadas en términos, ya que no se inclinan hacia términos frecuentes [3]. Por otro lado, los patrones léxico sintácticos pueden representar más de un concepto de dominio en el mismo patrón, al igual que una representación basada en frases [3].

2.2.2. Meronimia

Las relaciones semánticas de meronimia son una relación semántico-conceptual asimétrica [112]. La cual forma parte del proceso cognitivo de categorización y composición del significado [112].

Por ejemplo, este tipo de relación se da entre las partes y los todos como mano y brazo que sigue el patron X es una parte de Y (mano es una parte de brazo) [82].

2.2.3. Sinonimia

Las relaciones semánticas de sinonimia son las que presentan una relación entre dos o más palabras que tienen el mismo significado. Además de pertenecer a la misma parte del discurso, pero se escriben de manera diferente [112]. Los pares de palabras que presentan una relación de sinonimia comparten rasgos semánticos. Los sinónimos son probablemente la relación semántica más estudiada en tareas de *PLN* [40].

Algunos de los métodos existentes en la literatura para la extracción automática de este tipo de relación se basan en:

1. Extracción de frases clave: Este proceso extrae las palabras relevantes de cada documento y en base a ellas es posible encontrar las frases que comparten una relación de sinonimia [40].

2. Redes neuronales convolucionales: Este proceso entrena una red neuronal convolucional con las frases clave extraídas previamente [40].

3. Patrones léxico sintácticos: Este proceso se basa en patrones o expresiones del lenguaje natural formadas por frases nominales y conceptos presentes en un corpus [40]. Algunos ejemplos se presentan en la Tabla 4.1.

2.2.4. Antonimia

Las relaciones semánticas de antonimia son las que tienen significados opuestos o contrarios entre sí [112]. Por ejemplo, esta relación existe entre dos palabras cuyo significado es opuesto como grande y pequeño [82].

2.2.5. Tareas del Procesamiento del Lenguaje Natural

El *PLN* juega un papel preponderante en una amplia variedad de tareas como:

– Traducción automática *(Machine translation)*: su propósito es la traducción automática de un documento de un lenguaje a otro [129].

– Recuperación de la información (*RI*): tarea relacionada con la localización de documentos, es decir recuperar documentos relevantes asociados a una consulta recibida por un usuario [23].

- Extracción de Información (*EI*): necesaria para el análisis eficiente de los textos [137].

- Generación de resúmenes: en esta tarea se involucra el análisis y comprensión de un texto en lenguaje natural del que debe de extraerse solo la información que aporta las ideas principales del texto recibido de forma natural [139].

Otra tarea del *PLN* es el descubrimiento de tópicos que será descrita en la sección 2.3.

El resultado del uso de estas tareas ha dado paso a la generación de aplicaciones que se han convertido en una herramienta fundamental en el análisis de grandes volúmenes de información en poco tiempo. Un ejemplo de algunas de ellas se exponen en la siguiente sección.

2.2.6. Aplicaciones del procesamiento del lenguaje natural

En la actualidad existen aplicaciones que involucran el uso del procesamiento del lenguaje natural. Los cuales reciben información que debe ser tratada para algún propósito en particular. Algunos ejemplos de ellas son:

- Google Translator, iTranslate Converse para iOS o Microsoft Translator App son ejemplos de aplicaciones para Traducción automática *(Machine translation)*

- Motor de búsqueda de *Google* es un ejemplo de Recuperación de la información (*RI*)

- *TLDR This* han aplicado algunos algoritmos de de *PLN* para la generación automática de resúmenes [137]

- Cloud Speech-to-Text de Google es un ejemplo de aplicación de text to speech [137]

- Gmail y Outlook aplican filtros de correo electrónico [137]

2.3. Descubrimiento de tópicos

El descubrimiento de tópicos es una tarea de *PLN* que permite extraer automáticamente el significado de los textos mediante la identificación de tópicos recurrentes [128].

Los tópicos exponen de manera general y resumida el contenido de un documento y proporcionan resultados para el desarrollo de tareas con mayor complejidad como establecer relaciones entre conceptos [139].

El objetivo del descubrimiento de tópicos es extraer información a partir de textos para encontrar el tópico o idea central del cual tratan. La idea es que ciertos tópicos aparezcan en documentos relevantes.La tarea de descubrimiento de tópicos se realiza de acuerdo con la literatura con modelos, tales como:

- Análisis Semántico Latente (*LSA* por sus siglas en inglés) [73]
- Análisis de Dirichlet Latente (*LDA* por sus siglas en inglés) [73]
- Análisis Semántico Latente Probabilístico (*PLSA* por sus siglas en inglés) [73]

A continuación, se exponen de manera detallada cada uno de los modelos anteriormente mencionados.

2.3.1. Análisis Semántico Latente

El Análisis Semántico Latente es descrito como *un modelo usado para aprender representaciones de palabras densas* [124]. Este modelo es típicamente aplicado para factorizar una matriz de co-ocurrencia palabra-palabra obtenida de un corpus. Por lo que *LSA* calcula la frecuencia con la que aparecen las palabras en los documentos en todo el corpus. Lo que permite reconocer que documentos similares contendrán aproximadamente la misma distribución de frecuencias de palabras.

La información sintáctica (orden de las palabras) y la información semántica (multiplicidad de significados de una palabra determinada) se ignoran y cada documento se trata como una bolsa de palabras [4].

El *LSA* tiene su origen en *LSI* (del inglés *Latent Semantic Indexing*). Además es un modelo automático de recuperación de información, el cual incorpora la descomposición de valores singulares (*SVD Singular Value Decomposition*) [43].

El *LSA* es usado para determinar la similitud semántica entre los elementos analizados a partir de la co-ocurrencia con que son empleados en determinados contextos verbales [43]. La Figura 2.7 expone un ejemplo de la matriz generada por el modelo *LSA*.

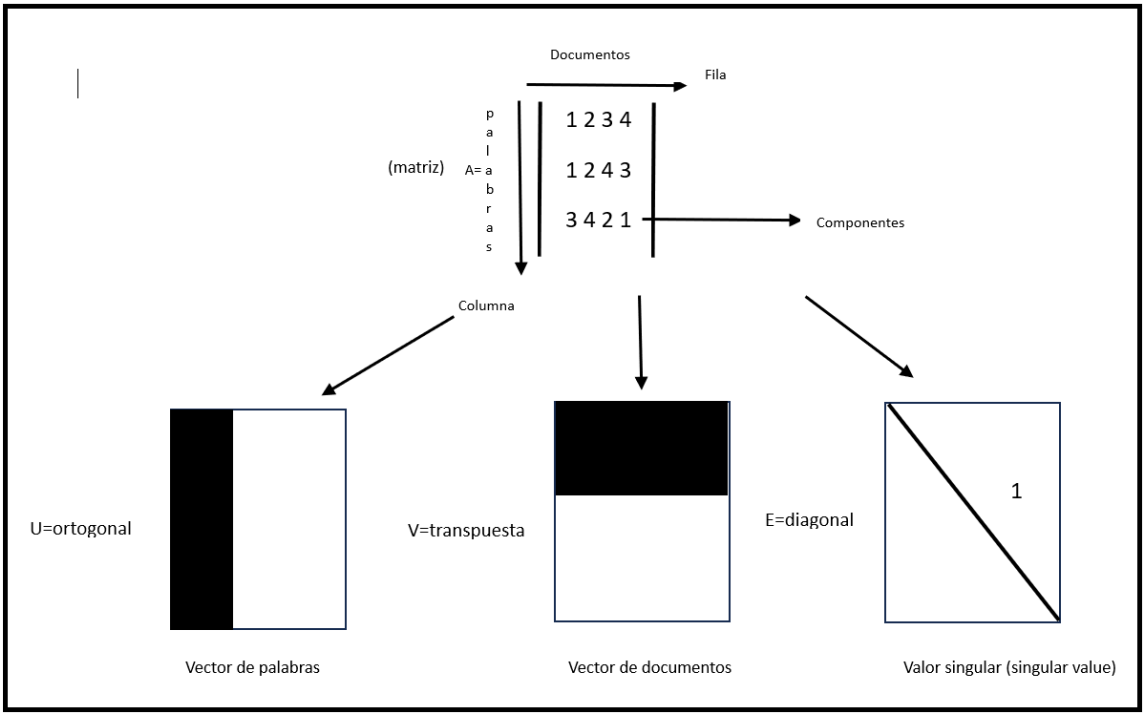


Figura 2.3: Un ejemplo de la matriz generada por *LSA*. Imagen tomada de [124].

Las matrices resultantes después de aplicar *SVD* son las siguientes:

- Matriz ortogonal (U): obtenida procesando linealmente el número de columnas de la matriz original (ortogonal) [124].

- Matriz transpuesta (V): obtenida intercambiando las filas con las columnas, proporcionando una disposición ortogonal de los elementos de la fila [124].

- Matriz diagonal (E): obtenida procesando linealmente el número de filas, columnas y dimensiones de la matriz original (A); la matriz diagonal representa el valor singular de (A), y en esta, todos los elementos que no pertenecen a la diagonal son nulos o iguales a cero [124].

2.3.2. Asignación Latente de Dirichlet

La Asignación Latente de Dirichlet (LDA) es *un modelo generativo*, es decir obtiene todos los resultados posibles para colecciones de datos discretos. El modelo fue desarrollado en 2003 donde destacaron las deficiencias de *tf-idf*, porque no puede comprender la semántica de las palabras [4].

El *LDA* considera a un tópico como una distribución en un vocabulario fijo. El *LDA* es un modelo bayesiano jerárquico de tres niveles (documento, palabra y tópico), que toma una cantidad de tópicos predefinida para toda la colección y se seleccionan las palabras que pertenecen a los tópicos seleccionados inicialmente [4].

El modelo consiste en identificar en qué medida los tópicos están presentes en los documentos [4]. Primero seleccionando una distribución sobre los tópicos seleccionados, es decir, el conjunto de tópicos predefinidos con sus palabras más probables. Segundo, para cada palabra del documento se escoge una asignación de tópicos y se selecciona la palabra para el tópico correspondiente [4]. La Figura 2.4 expone gráficamente el modelo LDA.

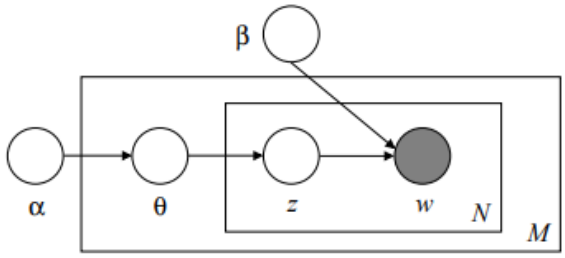


Figura 2.4: Representación gráfica del modelo LDA. Imagen tomada de [4].

En la Figura 2.4 M denota el número total de documentos y N denota el número de palabras en un documento determinado (el documento i tiene N_i palabras). El parámetro α denota el parámetro de Dirichlet antes de las distribuciones de tópicos por documento con un valor predeterminado de 0.1. El parámetro β denota el parámetro del Dirichlet antes de la distribución de palabras por tópico con un valor predeterminado de 0.01. θ_i denota la distribución de tópicos para el documento i. φ_k denota la distribución de palabras para el tópico k, z_{ij} denota la palabra j-th en el documento i y w_{ij} denota la palabra específica.

2.3.3. Análisis Semántico Latente Probabilístico

El Análisis Semántico Latente Probabilístico conocido también como Indexación Semántica Latente Probabilística (por sus siglas en inglés *PLSI Latent Semantic Indexing* o *PLSA*). Este modelo es una técnica estadística para el análisis de datos de co-ocurrencia. A partir de este modelo se puede derivar una representación de baja dimensión de las variables observadas en términos de su afinidad con ciertas variables ocultas.

En la Figura 2.5 M denota el número de textos y N denota el número de palabras en un texto determinado. El parámetro d indica un texto y z denota la variable latente u oculta (tópico). El parámetro w denota una palabra específica en el corpus.

El *PLSA* interviene en aplicaciones como recuperación y filtrado de información, procesamiento del lenguaje natural y aprendizaje automático [6]. El desarrollo de las tareas mencionadas involucran modelos de incrustación de palabras para la representación de los textos, aprendizaje automático y aprendizaje profundo [4].

La Figura 2.5 expone el modelo PLSA.

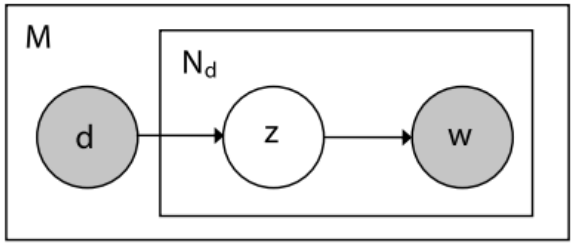


Figura 2.5: Representación gráfica del modelo PLSA. Imagen tomada de [6].

En las secciones 2.3.1, 2.3.2 y 2.3.3 se expusieron algunos de los modelos existentes más empleados en la literatura para el descubrimiento de tópicos. Cada modelo representa el texto como una matriz de tipo documento-término para llevar a cabo el descubrimiento de tópicos según el procedimiento de cada modelo. Sin embargo, para tareas como clasificación de texto la representación puede llevarse a cabo utilizando un modelo de incrustación de palabras.

2.4. Modelos de incrustación

Todas las áreas de estudio de la Inteligencia Artificial *(IA)* están diseñadas para procesar con números lo que representa un desafío para transformar palabras y texto; lo que ha originado el desarrollo de algoritmos que convierten palabras en números conocidos como *word embeddings* o incrustación de palabras. Las cuales hacen que sea mucho más sencillo aplicar las tareas de la *IA* para analizar el lenguaje natural y por ende las características existentes en el texto a analizar [5].

La representación de palabras de longitud fija, densas y distribuidas se denominan modelos de incrustación de palabras. Las cuales codifican información sintáctica y semántica, demostrando ser útiles como características adicionales en muchas tareas posteriores del *PLN* [5].

La representación de palabras y documentos es parte importante en las tareas de *PLN*. En general, se ha encontrado útil representarlos como vectores, que tienen una interpretación atractiva e intuitiva [5].

Los modelos de incrustación han surgido como un tema de investigación en sí mismos, al ser utilizados como características independientes en muchas tareas de *PLN* ya que codifican palabras precisas, relaciones de palabras sintácticas y semánticas [5].

Dos enfoques han sido propuestos para su implementación el primero basado en palabras que incorpora el contenido de un corpus de texto para generar un vector numérico para cada palabra que forma el corpus; el segundo basado en relaciones semánticas que incorpora pares de relaciones semánticas en una matriz para formar un vector numérico para cada palabra que forma la relación semántica. Los modelos previamente mencionados serán descritos a continuación.

2.4.1. Modelo de incrustación basado en palabras

Un modelo de incrustación basado en palabras son proyecciones en un espacio continuo de palabras que preservan las similitudes semánticas y sintácticas entre ellas. En la literatura existe evidencia que son una herramienta importante para tareas de *PLN*.

La incrustación de palabras es el proceso de convertir el texto en números siendo un proceso vital para que la computadora procese el lenguaje natural. La incrustación de palabras permite resolver problemas relacionados con el *PLN*. Una hipótesis detrás

de las incrustaciones de palabras es que son representaciones genéricas que se adaptan a la mayoría de las aplicaciones. Los vectores representan coordenadas en un espacio vectorial lo que posibilita el cálculo de palabras próximas en función de la distancia que exista entre sus vectores [37]. El cálculo de la similitud entre los vectores puede hacerse usando medidas como por ejemplo distancia euclídea o similitud coseno [37].

Los expertos en *PLN* y lingüística computacional han desarrollado técnicas que logran identificar similitudes entre palabras en función de la co-ocurrencia entre ellas [37].

Algunos de los modelos de incrustación de palabras existentes son: *word2vec, Glove y fastText* y han sido utilizados en tareas de *PLN* con el objetivo de mejorar los resultados de evaluación.

Word2vec

Word2vec es un modelo que recibe una palabra y genera su representación de forma vectorial mejor conocido como embedding. El modelo se encuentra disponible en dos versiones diferentes: *Continuous Bag-of-Words (CBOW)* y *Skip-Gram*. El modelo *Skip-Gram* analiza las palabras que forman una sentencia en un corpus y trata de usar cada palabra para predecir que palabras serán vecinas. Además suele proporcionar mejores resultados que *CBOW*. Por ejemplo a la palabra *coca* le seguirá *cola* con más probabilidad que cualquier otra palabra [19].

GloVe

El modelo *Global Vectors for Word Representation* o *GloVe* fue desarrollado por la Universidad de *Stanford*. El modelo es capaz de capturar conceptos como que "masculino" es "femenino" como "rey" es "reina", relaciones existentes entre verbos y tiempos verbales, o la vinculación entre países con capitales [100].

El modelo es similar a *word2vec*. La diferencia entre ambos es que *GloVe* sólo considera la información contextual, es decir, genera una matriz de tipo palabra a palabra, incluyendo la probabilidad $P(a|b)$ [100]. El objetivo de *GloVe* es obtener una representación de los vectores generando una probabilidad logarítmica de sus productos puntuales igual a la co-ocurrencia [100].

fastText

El modelo *fastText* es un conjunto de herramientas desarrollado por el equipo de investigación de *Facebook*. El propósito fue el aprendizaje de las representaciones de palabras y la clasificación de textos. La principal contribución del modelo de incrustación de *fastText* es que cuenta con la estructura interna de las palabras mientras aprende las representaciones de estas. El enfoque de representación de palabras del modelo de incrustación de *fastText* difiere de otras incrustaciones de palabras como *word2vec* [56].

El uso de cada palabra como la unidad más pequeña es aplicado en *word2vec*. Por el contrario *fastText* asume que una palabra está compuesta por n-gramas de caracteres donde la longitud de n puede cambiar de uno a la longitud de la palabra. El beneficio de este enfoque es que el método mantiene los vectores de palabras como n-gramas de caracteres [56]. El modelo puede encontrar representaciones vectoriales para palabras que directamente no se encuentran en el diccionario. Por lo que las clases de cada documento se determinan utilizando una función de pérdida f que permite calcular la distribución de probabilidad sobre las etiquetas de clase predefinidas [56].

2.4.2. Modelo de incrustación basado en relaciones semánticas

Un modelo de incrustación de relaciones semánticas tiene la particularidad que está formado por algún tipo de relación de sinonimia, hiponimia, hiperonimia, meronimia, antonimia, entre otras.

Al igual que un modelo de incrustación de palabras, estos modelos tienen la característica de ser una herramienta importante para tareas de *PLN*. Hasta ahora solo se ha encontrado evidencia de un modelo de incrustación de relaciones semánticas propuesto por Saedi et al. [109] llamado *wnet2vec*.

Saedi et al. [109] expone un modelo de incrustación de relaciones semánticas extraídas de la base de datos léxica *WordNet* [83], el cual será descrito a continuación.

Wnet2vec

Saedi et al. [109] presentan un modelo de incrustación de relaciones semánticas llamado *wnet2vec*. El modelo está formado por relaciones semánticas obtenidas

de la base de datos léxica *WordNet* [83]. El conjunto de datos base lo obtuvieron extrayendo solo un subgrafo de *WordNet* formado por 60,000 palabras [109].

La base de datos léxica *WordNet* está formada por nodos que están relacionados por diferentes tipos de relaciones semánticas (por ejemplo, hiperonimia, meronimia, sinonimia, entre otros). Por lo que para convertir el grafo en una matriz de relaciones solo fueron agregadas las relaciones de diferentes tipos con idéntico peso. Posteriormente con el objetivo de enriquecer la matriz creada aplicaron el siguiente procedimiento:

1. Enriquecimiento de la matriz de relaciones M para representar la fuerza de la afinidad semántica de relaciones o nodos identificados que no están conectados directamente por un borde, usando la ecuación sobre la centralidad de $Katz$ (ver Ecuación 2.1); que cuenta con un parámetro libre α, el cual gobierna el equilibrio entre el término del vector propio y el término constante en el cual puede observarse en la Ecuación 2.1 [91].

   Para aplicar la centralidad de $Katz$ primero se selecciona un valor para esta constante α. Por lo que es importante señalar que α no puede ser arbitrariamente grande. Por lo tanto, si $\alpha \to 0$, entonces sólo el término constante se mantiene en la ecuación. A medida que α aumenta desde cero, las centralidades aumentan y eventualmente llega un punto en el que divergen. Para que la expresión de la centralidad converja se debe seleccionar un valor de α menor que este [91]. Sin embargo, en la literatura existe poca información sobre el valor que debe tomar α. La mayoría de los investigadores han empleado valores cercanos al máximo de $\frac{1}{x_1}$, lo que coloca la cantidad máxima de peso en el término del vector propio y la cantidad más pequeña en el término constante. Esto devuelve una centralidad que es numéricamente cercana a la centralidad del vector propio ordinario, pero proporciona valores pequeños distintos de cero a los vértices que no están en los componentes fuertemente conectados a sus componentes externos [91]. El procedimiento realizado converge en la matriz M_G.

   $$M_G = (I - \alpha M)^{-1} \tag{2.1}$$

   dónde

   a) I es la matriz identidad.

   b) M es la matriz de relaciones

   c) α factor de decaimiento fijado en 0.75 que determina cómo dominan los caminos más cortos.

2. M_G está sujeto a la información mutua puntual (PMI) para reducir el posible sesgo introducido por la conversión a palabras con más sentido [110]. La información mutua puntual (PMI ver ecuación 2.2) es una medida de cuánto difiere la probabilidad real de una coocurrencia particular de eventos $p(x, y)$ de lo que esperaríamos que fuera sobre la base de las probabilidades de los eventos individuales y la hipótesis de independencia $p(x)p(y)$. Aún cuando PMI puede ser negativo o positivo, su resultado esperado en todos los eventos conjuntos es positivo [13].

$$pmi(x, y) = \log \frac{p(x, y)}{(p(x)p(y))} \tag{2.2}$$

3. Para una aplicación de conversión correcta: cada línea en M_G se normaliza usando la norma $L2$ para corresponder a un vector cuyas puntuaciones suman 1.

4. Los vectores que forman la matriz M_G se reducen a una dimensión de tamaño 300 aplicando el Análisis de Componentes Principales (PCA) [123]. El análisis de componentes principales (PCA), es un algoritmo de extracción de características y representación de datos utilizada en las áreas de procesamiento de texto, reconocimiento de patrones y visión por computadora [58]. Por lo que, PCA es un algoritmo estadístico utilizado para explorar series complejas de observaciones multivariadas mediante el reconocimiento de su contenido informativo más relevante. El PCA se comporta como una técnica de compresión que captura las principales características de los datos y al mismo tiempo revela su estructura [28]. Por lo que es un algoritmo de extracción de características y representación de datos utilizado en las áreas de procesamiento de texto, reconocimiento de patrones y visión por computadora [58]. PCA es utilizado para explorar series complejas de observaciones multivariadas mediante el reconocimiento de su contenido informativo mas relevante [28].

Hasta el momento solo existe un trabajo en donde se emplean las relaciones semánticas en un modelo de incrustación. Por lo que en esta tesis doctoral se parte

del desarrollo de un modelo de incrustación de relaciones semánticas como área de oportunidad para generar una aportación al conocimiento.

Una de las tareas de la inteligencia artificial (*IA*) más empleada es la clasificación de documentos. Esta tarea puede ser realizada con algoritmos que forman parte del aprendizaje automático o aprendizaje profundo. Estos tipos de aprendizaje así como algunas de sus características serán presentadas en las siguientes secciones.

2.5. Aprendizaje automático

El aprendizaje automático (*machine learning*) es una rama de la Inteligencia Artificial (*IA*), que investiga tareas que implican reconocimiento de patrones, diagnósticos, planificación, control y predicción [25, 39]. Esta tarea está formada por un conjunto de métodos que detectan automáticamente patrones en los datos y utilizan esos patrones para predecir nuevos datos o llevar a cabo una toma de decisiones para otra tarea en particular [100].

El aprendizaje automático se divide en supervisado y no supervisado, los cuales se describen a continuación.

2.5.1. Aprendizaje no supervisado

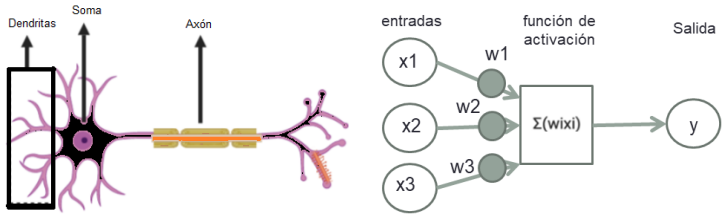
El aprendizaje no supervisado es usado para el descubrimiento de patrones en datos de los que se desconoce la respuesta correcta [97]. Los algoritmos de agrupamiento o el descubrimiento de tópicos son un ejemplo de aprendizaje no-supervisado.

La ventaja de este tipo de aprendizaje es que es posible aprender cosas acerca de los datos que antes eran desconocidas para encontrar estructuras ocultas en los mismos [97]. Por lo tanto, se puede concluir que el aprendizaje no supervisado produce modelos que se ajustan a las observaciones.

Este tipo de aprendizaje se caracteriza por el hecho de que no hay un conocimiento a priori, a diferencia del aprendizaje supervisado que se describirá en la siguiente sección.

2.5.2. Aprendizaje supervisado

La principal característica del aprendizaje supervisado es que necesita contar con un conjunto de datos de entrenamiento, es decir, datos que han sido correctamente



(a) Red neuronal biológica (b) Red neuronal artificial

Figura 2.6: Representación de una red neuronal biológica y artificial. Imagen tomada de [92].

etiquetados por un experto [97]. Además de contar con un conjunto de prueba que es usado para validar los resultados. El objetivo del aprendizaje supervisado es generar una función f que produce un valor y correspondiente a cualquier objeto de entrada x [12].

Uno de los modelos usados para el aprendizaje supervisado son las redes neuronales artificiales, que serán descritas a continuación.

2.5.3. Redes neuronales artificiales

Las redes neuronales artificiales *Artificial Neural Networks (ANN)* son un modelo computacional que simula el mecanismo de aprendizaje en organismos biológicos.

Las *ANN* se definen como técnicas de aprendizaje automático que simulan el mecanismo de aprendizaje de los organismos biológicos [1]. El sistema nervioso humano como organismo biológico contiene células conocidas como neuronas. Las neuronas están conectadas entre sí mediante el uso de axónes y dendritas y la conexión entre ellos se denomina sinápsis. La Figura 2.6(a) expone esas conexiones. La fuerza de las conexiones sinápticas van cambiando en respuesta a estímulos externos, es decir, es como se produce el aprendizaje en los organismos vivos [1]. Dicho procedimiento biológico es simulado en redes neuronales artificiales. El cual contienen unidades de cálculo denominadas neuronas. Las neuronas están conectadas entre sí a través de pesos, que cumplen el mismo papel que la fuerza de las conexiones sinápticas en los organismos biológicos [1]. Cada entrada a una neurona se escala con un peso, que afecta la función calculada en esa neurona como se muestra en la Figura 2.6(b).

Las redes neuronales son construidas como abstracciones de alto nivel de los modelos clásicos usados comúnmente en el aprendizaje automático. Las unidades

de cálculo básicas en la red neuronal están inspiradas en algoritmos tradicionales de aprendizaje automático como la regresión por mínimos cuadrados y la regresión logística [1].

Las redes neuronales obtienen su poder al agrupar estas unidades básicas y al aprender los pesos de las diferentes unidades de manera conjunta para minimizar el error de predicción. Una red neuronal puede verse como un grafo computacional de unidades elementales, usada en su forma más básica, sin unir varias unidades. Los algoritmos de aprendizaje se reducen a modelos clásicos de aprendizaje automático.

El poder de un modelo neuronal sobre los métodos clásicos se debe a que las unidades computacionales se combinan y los pesos de los modelos se entrenan usando sus dependencias entre sí. Al combinar varias unidades, el modelo tiene bases sólidas para aprender funciones de los datos más complicadas que las inherentes a los modelos elementales del aprendizaje automático básico. La forma en que estas unidades se combinan juega un papel en el poder de la arquitectura neuronal [1].

Además, los parámetros necesarios en las redes neuronales son aprendidos mediante el uso de programación dinámica llamada *backpropagation* ó retropropagación, que es actualmente la arquitectura de red neuronal más aplicada [47].

En los últimos años el aprendizaje profundo ha sido empleado como una técnica para la clasificación de documentos. Por lo que, los resultados obtenidos son más precisos que los proporcionados por un clasificador tradicional. Este concepto se define a continuación.

2.6. Aprendizaje profundo

El aprendizaje profundo (en inglés *Deep Learning*) es una forma de aprendizaje automático. El cual cuenta con algoritmos que permiten a una computadora recibir información en lenguaje natural y por lo tanto aprender de la experiencia y comprender el mundo en términos de una jerarquía de conceptos [12].

Una computadora obtiene el conocimiento de la experiencia, es decir, de la información proporcionada por un usuario en lenguaje natural. La jerarquía de conceptos permite que la computadora aprenda conceptos complejos construyéndolos a partir de conceptos simples.

El aprendizaje profundo es un método para recolectar y extraer conocimiento de sistemas complejos. Por lo que, es un campo de investigación activo que ha demos-

trado su utilidad en visión por computadora, procesamiento de audio, procesamiento del lenguaje natural y procesamiento de voz [12].

El aprendizaje profundo ha contribuido en el campo de investigación de inteligencia artificial obteniendo buenos resultados. Lo que ha originado que empresas como *Twitter* y *Facebook* incorporen aprendizaje profundo para la mejora de sus algoritmos [12].

En el estado del arte existen trabajos dentro del *PLN* que emplean aprendizaje profundo enfocando sus estudios a análisis de sentimientos, reconocimiento biométrico, descubrimiento de tópicos, detección de eventos, por mencionar algunos [12].

La tarea de descubrimiento de tópicos a partir de redes sociales con un enfoque de aprendizaje profundo ha sido investigada para generar modelos que comprendan los intereses de los usuarios de internet, hacer recomendaciones adecuadas en el futuro, informar fallas o detectar los temas más populares [12].

2.6.1. Redes neuronales convolucionales

Una red neuronal puede estar compuesta de tres capas: la capa de entrada, la capa oculta, y la capa de salida. Las neuronas se comunican entre capas adyacentes mediante un conjunto de conexiones. Una arquitectura de red es identificada como un patrón de conectividad de una red, donde cada neurona (denominada x_i) tiene como entrada los valores recibidos de las neuronas de la capa anterior (las que alimentan a x_i) [1].

Las redes neuronales convolucionales (*CNN*, del inglés *Convolutional Neural Networks*) son una red neuronal artificial. Las cuales imitan a las neuronas del cerebro humano [1]. Por lo que la *CNN* puede ser vista como una red multicapa o una red neuronal jerárquica, puesto que las *CNN* se obtienen apilando múltiples capas de características. Una capa está formada por k filtros lineales seguida de una función de respuesta no lineal [1]. Las *CNN* toman su nombre de una operación matemática lineal entre matrices llamada convolución. La *CNN* está formada por múltiples capas incluida la capa convolucional, capa de no linealidad, capa de agrupación y la capa completamente conectada.

Las capas convolucionales y completamente conectadas cuentan con parámetros, pero las capas de agrupación y no linealidad no tienen parámetros adicionales. Las capas convolucionales son un conjunto de filtros llamados: campos receptivos, que se ajustan para la extracción de características de una señal. En las *CNN* se comparten

neuronas a través de filtros que permiten extraer información de los conjuntos de datos a analizar. Un bloque con tres variables está presente en cada capa de la que se compone la *CNN*: entrada, pesos y salida. La característica fundamental de esas capas es que la salida de una capa se convierte en la entrada de la próxima. La principal tarea de las *CNN* es extraer patrones [20].

Las CNN juegan un papel preponderante en tareas como clasificación y segmentación de imágenes, visión por computadora y *PLN* [1]. Las *CNN* están formadas por múltiples capas de filtros convolucionales de una o más dimensiones. En cada capa se incorpora una función para realizar un mapeo [1]. Las capas tienen una fase de extracción de características compuesta de neuronas convolucionales, posteriormente hay una reducción por muestreo y al final se tienen neuronas de perceptrón simples para realizar la clasificación final sobre las características extraídas [1]. La fase de extracción de características es similar al proceso de estimulación en las células de la corteza visual humana [1]. En la Figura 2.7 se presenta la arquitectura básica de una red neuronal convolucional. Algunas de las tareas principales que realizan las *CNN* son:

- Detección/categorización de objetos.
- Clasificación de escenas.
- Clasificación de imágenes en general.
- Clasificación de texto.

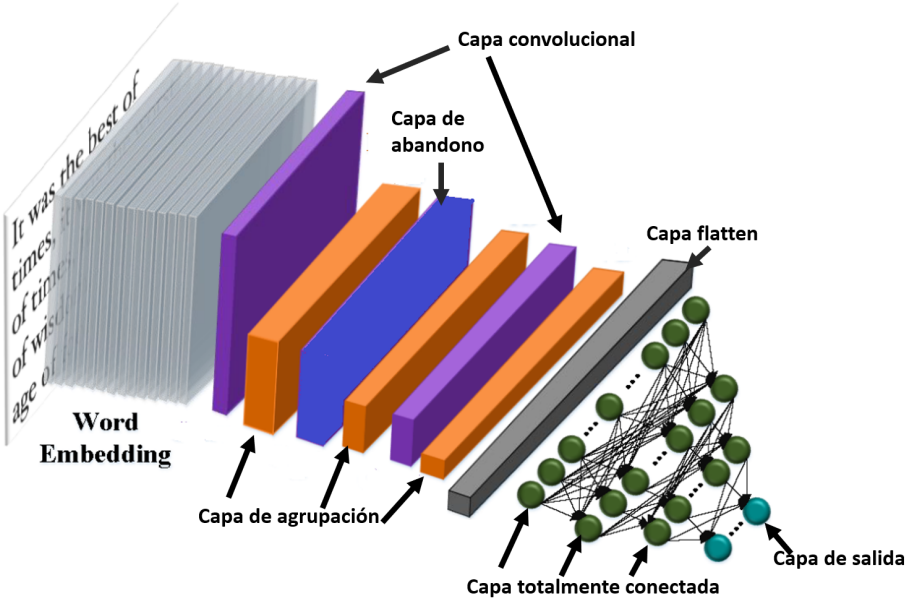


Figura 2.7: Representación de una red neuronal convolucional. Imagen tomada de [1].

La arquitectura de una *CNN* se compone de dos piezas clave:

- Modelo Convolucional: Es un modelo de extracción de características que aprende a extraer las características de cada documento.
- Modelo Completamente Conectado: Es la interpretación de las características previamente extraídas.

Las convoluciones en una *CNN* se realizan a través de una palabra de entrada usando núcleos de diferentes tamaños. Los mapas de características que se obtienen son procesados usando una capa de agrupación para resumir las características extraídas.

La implementación de una red neuronal convolucional es posible por medio de las herramientas proporcionadas por el lenguaje de programación Python, es decir Tensorflow [31] y Keras [130]. TensorFlow es una plataforma de código abierto para el aprendizaje automático, que proporciona herramientas, bibliotecas y recursos que permiten trabajar de manera innovadora el aprendizaje profundo [14].

Keras es una biblioteca diseñada para el desarrollo y tratamiento de redes neuronales de código abierto, se encuentra desarrollada en el lenguaje de programamción Python. La plataforma TensorFlow es la responsable de la ejecución de Keras [14].

La tarea clasificación de texto (ver sección 2.5.2), y el descubrimiento de tópicos (ver sección 2.3) requieren de un proceso de evaluación donde se podrá observar la calidad de los resultados obtenidos. Con este propósito en la siguiente sección se exponen las métricas *precisión, recall, accuracy* y medida-F_1 empleadas para evaluar el rendimiento de la tarea de clasificación de texto. Así como la métrica de coherencia del tópico normalizada para evaluar el redimiento de la tarea de descubrimiento de tópicos

2.7. Métricas de evaluación

El objetivo de las métricas de evaluación es medir el nivel de rendimiento de la tarea de clasificación de textos y el descubrimiento de tópicos

La tarea de clasificación normalmente es evaluada con las métricas de precisión (ver ecuación 2.3), exhaustividad (ver ecuación 2.4), exactitud (ver ecuación 2.5) y F_1 (ver ecuación 2.6).

La tarea de descubrimiento de tópicos es evaluada con la métrica de coherencia del tópico normalizada (ver ecuación 2.7), la cual mide el rendimiento de los tópicos descubiertos, es decir, que tan relacionadas están las palabras que integran a los tópicos.

A continuación se define cada métrica:

- Precisión. La métrica de precisión permite medir la calidad del modelo evaluado. La precisión se define como la proporción de *términos recuperados* realmente relevantes del total de los *términos relevantes* [111]. Un resultado de precisión de 1 se refiere a una recuperación perfecta donde únicamente se recuperan los términos relevantes [32]. En la Ecuación 2.3 se muestra la fórmula de precisión.

$$precision = \frac{Terminos\ relevantes\ recuperados}{Terminos\ recuperados} \tag{2.3}$$

- Exhaustividad. La métrica exhaustividad se define como la proporción de *términos relevantes recuperados* del total de los *términos relevantes* en el conjunto de datos a evaluar, independientemente de que éstos se recuperen o no

[32]. En la Ecuación 2.4 se muestra la fórmula de exhaustividad.

$$exhaustividad = \frac{Terminos\ relevantes\ recuperados}{Terminos\ relevantes} \tag{2.4}$$

– F_1. La métrica F_1 está formada por la combinación de las métricas de precisión y exhaustividad, es decir es la media armónica [32]. En la Ecuación 2.5 se muestra la fórmula de F_1.

$$F_1 = 2 * \frac{precision * exhaustividad}{precision + exhaustividad} \tag{2.5}$$

– Exactitud. La métrica de exactitud mide el porcentaje de casos que el modelo acertó, es decir, el promedio de las predicciones correctas entre el *total de casos* [32]. En la Ecuación 2.6 se muestra la fórmula de la exactitud.

$$exactitud = \frac{Cantidad\ de\ casos\ correctos}{Total\ de\ casos} \tag{2.6}$$

– Coherencia del tópico normalizada: El cálculo de esta métrica consiste en primero obtener la *Coherencia Normalizada (CohN)* de cada tópico (t_i) de la forma que se presenta en la Ecuación 2.7.

$$CohN(t_i) = \frac{2}{K(K-1)} \sum_{j=2}^{K} \sum_{i=1}^{j-1} \frac{\log \frac{p(w_i,w_j)}{p(w_i)p(w_j)}}{-\log p(w_i,w_j)} \tag{2.7}$$

La Coherencia del tópico normalizada se basa en obtener la información puntual mutua normalizada (*NPMI* por sus siglas en inglés) de cada par de palabras que pertenecen a las k palabras que representan a cada tópico. $P(w_i, w_j)$ es la probabilidad de que las palabras w_i y w_j co-ocurran en un mismo párrafo dentro del conjunto de textos externos, $P(w_i)$ y $P(w_j)$ es la probabilidad de que la palabra w_i y w_j, co-ocurran en un mismo párrafo. Posteriormente se obtiene una *Coherencia Global Normalizada (CohGloN)* de todos los tópicos con un promedio considerando la Coherencia del tópico normalizada de cada tópico, mediante la Ecuación 2.8. La cual proporciona valores en el rango [-1,1]

[39].

$$CohGloN = \frac{1}{T}\sum_{i=1}^{T} CohN(t_i) \tag{2.8}$$

Capítulo 3

Estado del arte

El descubrimiento de tópicos ha sido ampliamente estudiado durante varios años a través de la implementación de diferentes métodos. El cual permite extraer la idea central (tópico) de un documento. A su vez puede ser aplicado en colecciones de documentos permitiendo descubrir la idea central en múltiples documentos. Dichas ideas centrales pueden ser utilizadas en diferentes aplicaciones que enfrenten la necesidad de trabajar con los tópicos de un texto en particular.

Por lo que con el objetivo de conocer las investigaciones orientadas al descubrimiento de tópicos es llevar a cabo una revisión del estado del arte de los métodos existentes. Los cuales emplean modelos como *LDA, LSA* y *PLSA*, sin embargo, algunos autores incorporan como estrategia adicional modelos de incrustación de palabras como *word2vec, BERT, Sense2Vec* y *FastText* o una arquitectura de aprendizaje profundo como *variational autoencoding*, redes neuronales profundas, *Long Short-Term Memory* o *autoencoder*. Además, se encontró un modelo de incrustación diferente a los mencionados anteriormente, el cual se basa en el uso de relaciones semánticas previamente extraídas de una base de conocimiento y un procedimiento de factorización de matrices.

Por lo tanto, es importante conocer los diferentes métodos existentes para la extracción de relaciones semánticas en diferentes bases de conocimiento. Otro tema abordado son los métodos para la extracción de relaciones semánticas. Algunos autores se basan en la extracción de características semánticas y sintácticas. Con el objetivo de mejorar los resultados obtenidos con las características anteriormente mencionadas los autores incorporan algún modelo de incrustación de palabras tales como *word2vec, glove, CW, HPCA*, *fastText* o aprendizaje profundo.

3.1. Métodos para el descubrimiento de tópicos

En esta sección se presentan tres métodos orientados al descubrimiento de tópicos tales como métodos sin aprendizaje profundo, métodos con aprendizaje profudo y métodos con incrustación de pálabras. En la sub-sección 3.1.1 se presentan métodos que incorporan modelos como *Non-Negative Matrix Factorization*, Biterm Topic Model (BTM), *LDA, LSA* y *PLSA*. La sub-sección 3.1.1 expone trabajos que incorporan junto con los modelos anteriormente mencionados un modelo de aprendizaje profundo como *variational autoencoding*, redes neuronales profundas, *Long Short-Term Memory* o *autoencoder*. La sub-sección 3.1.3 expone trabajos que descubren tópicos utilizando modelos de incrustación de palabras como *word2vec, BERT, Sense2Vec* y *FastText*.

3.1.1. Métodos sin aprendizaje profundo

Sha et al. [116] presentan un análisis de la toma de decisiones del gobierno de Estados Unidos y la pandemia COVID-19. Los autores aplicaron un modelado de tópicos para tener conocimiento sobre el manejo que el gobierno ha dado a la pandemia COVID-19. El análisis rastrea subtópicos en evolución en torno al riesgo, las pruebas y el tratamiento. Los autores realizan pruebas con *LDA* (Análisis de Dirichelet Latente) y *HBTM* (red binomial de *Hawkes*). Además construyen una red de influencia entre los funcionarios gubernamentales por medio de la causalidad de Granger y evalúan por medio de la métrica coherencia del tópico. Los resultados obtenidos mostraron que en el aspecto temporal *HBTM* tiende a mejorar la coherencia del tópico en relación con LDA.

Beguerisse-Díaz et al. [9] exponen un estudio en *tweets* relacionados con la enfermedad de diabetes. Los cuales incluyen grupos temáticos como: información de salud, noticias, interacción social y comerciales. Los *tweets* fueron agrupados en contenedores semanales por medio de un grafo de adyacencia de palabras. El texto fue procesado y posteriormente crearon una red de co-ocurrencia de palabras. Los nodos fueron representados por palabras y los bordes por dos palabras que aparecieron conjuntamente en los *tweets* con mayor probabilidad. El grafo de co-ocurrencia lo analizaron por medio de la estabilidad de Markov para la extracción de comunidades o grupos relevantes de palabras que aparecen en los *tweets*. Las comunidades de palabras las utilizaron como entrada al modelo *LDA* para el descubrimiento de los

tópicos. Los autores evaluaron con las puntuaciones *hub/authority*, lo que les permitió recuperar el puntaje a lo largo del tiempo para 10 tópicos principales clasificados.

Por otro lado, Ghenai y Mejova [38] exponen un método para el rastreo de información errónea sobre la enfermedad del Zika. El método emplea el modelo *LDA* y un enfoque de recuperación de información para identificar los *tweets* relevantes. Los autores aplicaron externalización abierta de tareas (*crowdsourcing*) para distinguir entre rumores y *tweets* de aclaración, que posteriormente fueron usados para construir clasificadores como árboles de decisión, *Naïve Bayes* y *random forest*. Las métricas precisión, exhaustividad y medida-F_1 fueron utilizadas para la evaluación de sus resultados. Los autores descubrieron un comportamiento explosivo de tópicos relacionados con rumores y demostraron que, es posible identificar *tweets* que contienen rumores utilizando técnicas automatizadas.

De igual manera Karami et al. [57] exponen un método para el descubrimiento de tópicos haciendo un análisis lingüístico y semántico. Los autores se apoyan de la herramienta de análisis lingüístico *LIWC* (La Investigación Lingüística y el Conteo de Palabras), que es capaz de encontrar sentimientos, personalidad y motivaciones en un corpus. Además, utilizan un diccionario relacionado con salud que sirve de soporte para determinar si un tópico contiene palabras asociadas con la salud. Para explorar las opiniones en los *tweets* usaron *LDA*. Los tópicos descubiertos fueron sobre dieta, ejercicio, obesidad y diabetes (*DDEO*). Por lo que fue posible el descubrimiento de subtemas incluídos términos tanto *DDEO* como no *DDEO*.

Por su parte Majdabadi et al. [79] exponen una versión modificada del algoritmo *RankClus*. La versión presentada extrae tendencias de *tweets* en inglés y persa basado en un grafo heterogéneo. Los *tweets* están representados con 3 tipos de nodos: *tweets*, palabras y *hashtags*. Además, emplean un algoritmo de agrupamiento basado en *ranking* para detectar nodos relevantes en el grafo que se consideran tendencias. El algoritmo detecta grupos de *tweets* que tienen palabras y *hashtags* en común. Así mismo, propusieron un algoritmo de puntuación que evalúa cada grupo y encuentra las tendencias con mayor coherencia. Los autores realizaron una comparación en términos de precisión el desempeño del algoritmo propuesto con el desempeño de los modelos *k-means* y *LDA*. Lo que les permitió verificar que su modelo supera a los otros dos modelos para ambos idiomas.

Yu et al. [150] descubren los tópicos en *tweets* obtenidos durante la pandemia de COVID-19. Los autores emplearon el modelo *LDA* para extraer 8 tópicos en dos

conjuntos de datos de los periódicos españoles el país y el mundo. Cada conjunto de datos fue dividido en 3 periodos diferentes. El primero previo a la crisis, el segundo durante el confinamiento y el tercero cuando el gobierno español adoptó medidas sanitarias. Los autores reasignaron los tópicos descubiertos en grupos, por lo que obtuvieron un nuevo conjunto de datos con los *tweets* de cada periódico clasificado. Finalmente generaron una red de relaciones entre los grupos generados a partir de su matriz de coocurrencia de palabras para cada diario durante un periodo de tiempo. Por lo que los autores señalan que su trabajo contribuye a comprender cómo los medios de comunicación españoles cubren las crisis de salud pública en las plataformas de redes sociales.

En la investigación realizada por Ordun et al. [95] exponen un análisis para evaluar el carácter distintivo de los tópicos, los términos, características clave, la velocidad de difusión de la información y el comportamiento de la red para los *tweets* de COVID-19. Los autores proponen utilizar la coincidencia de patrones y el descubrimiento de tópicos a través de *LDA*. Lo que generó veinte tópicos diferentes que discuten la propagación de casos, los trabajadores de la salud y el equipo de protección personal. El análisis propuesto es una contribución con métodos de aprendizaje automático no reportados anteriormente en la literatura de *twitter* de COVID-19. Los autores proporcionan un cálculo de los tiempos de retuiteo para comprender que tan rápido se propaga la información sobre COVID-19 en *Twitter*.

Qiang et al. [104] exponen una revisión exhaustiva de técnicas de modelado de tópicos en textos cortos. Los autores presentaron el desarrollo de la primera biblioteca integral de código abierto llamada *STTM (Short Text Topic Modeling)*. La biblioteca se encuentra desarrollada en Java e integra varios algoritmos para el descubrimiento de tópicos. Uno de los algoritmos desarrollados para el descubrimiento de tópicos fue *LDA* y como métrica de evaluación fue la coherencia del tópico. Los conjuntos de datos utilizados fueron en inglés en los dominios medicina y tecnología. El resultado fue una aplicación que incluye las tareas de clasificación, agrupamiento y descubrimiento de tópicos.

El trabajo de Buenaño Fernández et al. [15] expone un método para el descubrimiento de tópicos que comienza con una recopilación de información de encuestas sobre autoevaluación docente en una universidad ecuatoriana. Los autores propusieron un caso de estudio para la evaluación de la metodología genérica basada en el descubrimiento de tópicos, redes de texto y la clasificación manual de los tópicos en-

contrados. Para el descubrimiento de los tópicos los autores hicieron uso del modelo *LDA*. La evaluación fue hecha con la métrica de coherencia del tópico. Los autores señalan que su modelo está diseñado para el análisis de respuestas a preguntas abiertas, sin embargo, su modelo es lo suficientemente flexible para usarse con diferentes fuentes de datos.

Fuentes-Pineda y Meza-Ruiz [33] presentan un método para el descubrimiento de tópicos basado en *min-hasing*. El cual extrae y agrupa co-ocurrencias de palabras. El corpus empleado en los experimentos realizados pertenece al dominio de noticias y finanzas en inglés y español. El método no requiere que se especifique el número de tópicos de antemano. Además, es capaz de trabajar con corpus de texto masivo. El método consistió en múltiples particiones aleatorias del corpus para encontrar conjuntos de palabras recurrentes que fueron agrupadas para generar los tópicos finales. Los resultados obtenidos mostraron evidencia sobre la relevancia de las relaciones para el descubrimiento de patrones en datos discretos a gran escala. En general los autores señalan que su método es capaz de producir tópicos coherentes con amplio vocabulario, demostrando su solidez con palabras poco comunes.

Yang et al. [147] exponen un método de descubrimiento de tópicos representativos (novel topic representative term discovery por sus siglas en inglés *TRTD*) para agrupar textos cortos. El método descubre grupos de términos representativos de tópicos estrechamente vinculados y explota la cercanía y el significado de las palabras que forman los tópicos. La cercanía de los términos representativos del tópico fue evaluada por su co-ocurrencia. Los resultados experimentales demostraron que el método logró resultados competitivos en la agrupación de texto corto que los existentes en la literatura. El método fue evaluado con las métricas índice de *rand* ajustado, información mutua ajustada e información mutua normalizada. El conjunto de datos usado para sus experimentos fue el proporcionado por *Text REtrieval Conference* (*TREC*) 2011 sobre eventos históricos.

Otros autores como Núñez-Reyes et al. [93] exponen un método no supervisado para la identificación de tópicos en textos cortos. Los autores realizaron pruebas con datos proporcionados por los organizadores del *RepLab* (campaña de evaluación competitiva para sistemas de gestión de la reputación online) del 2013 en inglés y español con los dominios autos, bancos, universidades y música. Los documentos fueron preprocesados y representados como un vector de términos ponderados obteniendo una bolsa de palabras que se emplearon como datos de entrada y los pesos los

definen por medio del ponderado booleano, es decir, se asigna el peso de 1 si la palabra ocurre en el documento y 0 en otro caso. Además, aplican la medida del coseno como métrica de proximidad. Los autores desarrollaron una representación de baja dimensionalidad obtenida mediante el método de punto de transición. El método hace uso del algoritmo de agrupamiento *k-means* y estrella para tratar de eliminar las limitaciones de los esquemas tradicionales. Los resultados obtenidos mostraron que la representación propuesta permite obtener un comportamiento similar al que se logra cuando se evalúa el método en el mismo dominio en el que fue construida la representación. Los resultados experimentales mostraron que el método propuesto permite tener una representación de textos robusta.

Satu et al. [114] proponen un método de clasificación y descubrimiento de tópicos basado en agrupamiento denominado TClustVID. Los autores señalan que su modelo es el primer estudio que investiga en datos de *Twitter* relacionados con COVID-19. Los autores aplicaron el algoritmo k-mean para el agrupamiento de los textos. De los cuales seleccionaron los mejores aplicando algoritmos para la clasificación de los textos como árbol de decisión, k-vecinos, regresión logística, *Naïve Bayes, random forest* y máquinas de soporte vectorial de aprendizaje. Posteriormente los grupos seleccionados serán grupos positivos, negativos y neutros los cuales fueron el conjunto de datos sobre el cual descubrieron tópicos utilizando el modelo LDA. El enfoque propuesto logró identificar las opiniones públicas (tópicos) relacionadas con el COVID-19, así como actitudes hacia las estrategias de prevención de infecciones de personas de diferentes países con respecto a la situación pandémica actual. El método fue evaluado utilizando las métricas precisión, área bajo la curva (AUC), medida-F_1, media-g, sensibilidad y especificidad.

Wei and Guo, [144] presentan un método de descubrimiento semántico de tópicos basado en el grado de co-ocurrencia condicional. El método divide cada documento en múltiples subdocumentos de acuerdo con la estructura semántica del documento. Las palabras con una fuerte relevancia semántica se extraen en función del grado de co-ocurrencia dentro de cada subdocumento. En base a las palabras previamente extraídas forman nuevos documentos. El descubrimiento de tópicos de los nuevos subdocumentos fue obtenido aplicando el modelado de tópicos con muestreo de Gibbs. Los autores obtuvieron distribuciones de tipo documento-tópico de los documentos originales fusionando distribuciones de tipo documento-tópico de nuevos subdocumentos. Los experimentos fueron realizados utilizando diferentes corpus de

tipo público. Los resultados experimentales mostraron que el método propuesto es capaz de aportar tópicos con una mayor coherencia entre sí. Los córpora utilizados fueron cuatro corpus en idioma chino recopilados y clasificados por la *Universidad Fudan y Sougou Lab*. Además usan tres corpus en inglés con un gran conjunto de datos de reseñas de películas y *Reuters-21578*. Los autores evaluaron sus resultados por medio de las métricas micro y macro precisión, exhaustividad y medida-F_1. Los resultados obtenidos superaron a los obtenidos por modelos como los que se basan en el uso de modelos de incrustación de palabras o el proceso jerárquico de Dirichlet.

En el trabajo desarrollado por Cong et al. [21] exponen una representación alternativa de la asignación de Dirichlet Latente Profunda (*DLDA*). Los autores proponen el cambio de parámetros de los vectores básicos restringidos (*simplex*). Además, derivaron una matriz de información de Fisher y calcularon la inversa de la matriz de información, que resultó en un algoritmo de gradiente estocástico. La representación propuesta fue capaz de aprender diferentes parámetros. Los experimentos fueron realizados con los conjuntos de datos *20-Newsgroups, Reuters* y *Wikipedia*.

Torres-Rondón et al. [131] exponen un método para el descubrimiento de tópicos basado en el contexto de los documentos. El método obtiene de manera automatica una representación de la información basada en grafos. Los algoritmos *PageRank* y *HITS (Hyperlink Induced Topic Search)* fueron utilizados para determinar la relevancia de cada tópico recuperado. Por lo que es posible descubrir los tópicos con mayor relevancia. Para la evaluación del método propuesto los autores utilizaron las métricas de precisión, exhaustividad y medida-F_1. Los resultados obtenidos mostraron mejores resultados al utilizar el algoritmo *HITS* lo que permitió ejemplificar su uso en entornos como el análisis de noticias y en el resúmen de artículos científicos.

Jiang et al. [53] exponen un método dinámico paralelo para el descubrimiento de tópicos. Los autores desarrollaron un mecanismo de ajuste de tópicos en evolución y la reducción de las probabilidades de muestreo de palabras indiscriminadas por tópico. Lo que mostró un rendimiento competitivo en comparación con los existentes en la literatura. Los autores llevaron a cabo experimentos supervisados y no supervisados utilizando los conjuntos de datos *KOS* del Daily Kos, un blog político, *NIPS* proveniente de artículos de NIPS de 1987 a 2016, tercero $ENRON$ que proviene de los correos electrónicos de una empresa llamada *Enron* y $NYTIMES$ del periódico *The New York Times*. Los autores señalan que el método propuesto generó un desempeño competitivo en comparación con los modelos existentes en la literatura.

El modelo propuesto por Arutchelvan and Selvan [7] se centra en generar resúmenes abstractivos a partir de multidocumentos. Por lo que construyeron grafos semánticos entre oraciones con frases clave y entidades nombradas existentes en los corpus y lo enriquecieron con tópicos descubiertos con el modelo *LDA*. Por lo que aplicaron una técnica de clasificación de oraciones para encontrar oraciones importantes además de usar la similitud coseno para reducir la información redundante. Los autores emplearon la técnica de clasificación de nodos de centralidad además de una técnica de clasificación de grafos ponderados para obtener la secuencia de las oraciones. El conjunto de datos empleado fue *Daily Mail* [48]. La métrica de evaluación utilizada para medir el rendimiento del enfoque propuesto fue *Rouge* [75].

Las investigaciones hechas por Zech et al. [151] exponen un análisis de diferentes métodos para generar características a partir de informes de radiología. El conjunto de datos está formado por 96,303 informes de tomografía computarizada *(TC)* de cabeza. Para el análisis de los informes médicos construyeron un enfoque basado en bolsa de palabras, incrustación de palabras y el descubrimiento de tópicos con el modelo *LDA*. Un total de 1,004 informes de *TC* fueron etiquetados manualmente con información de interés para especialistas en el área. La métrica de evaluación fue área bajo la curva característica (AUC) [49]. El modelo *LDA* logró un AUC de 0.872.

Saqlain et al. [113] proponen un método para el etiquetado de grupos que asigna una etiqueta genérica. La técnica incorpora la frecuencia de término y la frecuencia de documento inverso, con el cálculo de frecuencia de término refinado mediante el uso de un diccionario de sinónimos. Los autores utilizaron *WordNet* como un recurso externo para la generación de hiperónimos de los términos que tienen el *tf-idf* más alto. Los hiperónimos con la frecuencia más alta fueron tomados como la etiqueta del grupo. Los conjuntos de datos utilizados fueron *Daily Jang newspaper, Open Directory Project (ODP), 20-Newsgroup* y *Reuters*. El desempeño del modelo propuesto fue evaluado con la métrica de exactitud.

El trabajo de Pratama et al. [103], exponen un análisis de sentimientos de reseñas de un hotel. Las reseñas fueron determinadas en base a 5 aspectos del hotel: comida, servicio, ubicación, comodidad y limpieza. Los autores aplicaron el modelo LDA para descubrir los tópicos existentes para obtener pares de opiniones de las reseñas con el fin de generar un sistema de análisis de opinión a nivel de aspecto. Posteriormente los documentos son clasificados por coincidencia de similitud. En cada documento

se realiza una expansión semántica con sinónimos. El propósito es extender el valor de similitud con los tópicos descubiertos usando similitud coseno. El algoritmo de máquinas de soporte vectorial fue usado para clasificar cada opinión etiquetada. El conjunto de datos fue extraído por los autores de *tripadvisor.com*. Las métricas para evaluar el desempeño del modelo expuesto fueron precisión, exactitud, exhaustividad y la medida-F_1. El modelo propuesto superó a varios autores en la literatura con una precisión de 0.956.

Gao et al. [34] exponen un método para la extracción de tópicos de alta calidad y correlaciones de tópicos. Los autores generan grafos evolutivos de tópicos a partir de textos breves, que no solo capturen la línea de tiempo del tópico principal, sino que también revelen las correlaciones entre subtemas relacionados. Primero propusieron un modelado de lenguaje transformador ($ETLM$) para cuantificar la relación entre las palabras. El modelo fue denominado modelo de tópico correlacionado regularizado de campo aleatorio condicional ponderado ($CCTM$). Los grafos evolutivos de tópico son generados por una versión en línea de $CCTM$ para capturar los patrones evolutivos de los tópicos principales y subtópicos relacionados. Los autores utillizaron los conjuntos de datos de noticias en inglés, StackOverflow, y Q&A. Los resultados obtenidos demostraron que el método superó la calidad de los tópicos y presenta patrones motivados para la extracción de la evolución de los tópicos. El rendimienrto del modelo fue evaluado con las métricas coherencia del tópico, UCI (medida extrínseca) [90] y UMass (medición intrínseca) [84].

Por su parte Kinariwala y Deshmukh [61] proponen un método llamado $Onto\_TML$ de etiquetado automático basado en ontologías. La ontología utilizada se llama $CEPS(Crime, Environment, PoliticsandSports)$ y contiene los dominios delincuencia, medio ambiente, política y deportes. Los modelos para el descubrimiento de tópicos utilizados fueron *LDA, NMF (Non-negative Matrix Factorization)* y *SeaNMF (semantics-assisted non-negative matrix factorization)*. El modelo $CEPS$ es utilizado para asignar una etiqueta genérica adecuada a los tópicos descubiertos. Los conjuntos de datos utilizados fueron *News Headline* y *News Category*. La evaluación del método propuesto es realizada con la puntuación de distancia de Google normalizada. Los resultados arrojados mostraron que el método generó etiquetas apropiadas para los tópicos descubiertos. El objetivo principal del algoritmo $Onto\_TML$ fue asignar automáticamente las etiquetas apropiadas a tópicos mediante el uso de $CEPS$. La métrica utilizada para evaluar el desempeño del modelo fue la exactitud y la

coherencia del tópico.

Otros autores como Li et al. [74] proponen un método para el descubrimiento de tópicos en documentos secuenciales basado en la dependencia híbrida de tópicos entre documentos. El modelo considera la dependencia consecutiva, dependencia de tendencia e independencia en documentos contextuales. Los autores proponen una secuencia de evolución de tópicos más compleja, considerando relaciones de dependencia local detalladas. Los conjuntos de datos utilizados fueron *Reuters, Multilingual Text, Annotations Dataset, THUCNews, PubMed, Weibo-84168* y *Twitter Event Detection*. La métrica de evaluación utilizada para medir el desempeño de los métodos expuestos es la de perplejidad. Los experimentos realizados mostraron que los métodos expuestos superan a los existentes en el estado del arte, en términos de precisión, calidad y efectividad de la detección de valores atípicos.

Por su parte Hu et al. [51], proponen un método llamado *SP-BTM* que elige las palabras con partes gramaticales específicas para formar bitérminos para el descubrimiento de tópicos. Los experimentos realizados verificaron que los sustantivos, verbos y adjetivos son benéficos para la representación de tópicos y el modelo expuesto puede mejorar la eficacia de la agrupación de servicios. El modeo *SP-BTM* se basa en aplicar el mapeo de Gibbs y el algoritmo de agrupamiento K-means. Las métricas empleadas fueron precisión, exactitud, exhaustividad, medida-F_1 y puridad.

De igual manera Kawamae [60] exponen un método llamado TAN una adaptación de modelos de lenguaje neuronal (NLM) previamente entrenados a la tarea de generación de texto incondicional. El modelo se centra en tópicos para cerrar la brecha semántica entre corpus de diferentes dominios e inyectando tópicos en modelos de incrustación de tipo $BERT$. El modelo introduce la alineación de tópicos (TA) como manipulación y modelado de distribución de tópicos (TEM) y modelado de incrustación de tópicos. Lo que permitió a TAN descubrir tópicos específicos en los dominios de cada conjunto de datos empleado. Los conjuntos de datos utilizados fueron *Amazon review* y $Yelp$. La métrica utilizada para evaluar el rendimiento del modelo propuesto es la coherencia del tópico. Los autores concluyen que TAN es capaz de adaptar NLM para generar texto válido y ademas supera a modelos como LDA, NMF y $Topic2Vec$.

En el trabajo de Shahbazi y Byun [119] presentan un método que está formado por una combinación de aprendizaje por refuerzo y un modelo de factorización de matrices no negativa. El objetivo de los autores fue descubrir tópicos significativos y

subyacentes en textos cortos. El método se basa en el enfoque *Seq2Seq (sequence to sequence)* utilizando el modelo SeaNMF (a semantics-assisted non-negative matrix factorization) y el algoritmo de descenso de coordenadas en bloque. Además, los autores hicieron una comparación de diferentes conjuntos de datos del mundo real mediante cálculos numéricos y seleccionaron dos modelos para obtener un mejor rendimiento en el descubrimiento de tópicos en textos cortos. Los autores señalaron que en base a los resultados obtenidos su método supera las técnicas más avanzadas en términos de métodos para el descubrimieto de tópicos en documentos de textos cortos.

3.1.2. Métodos con aprendizaje profundo

Chai and Li [16] proponen un método de minería de textos llamado *neural topic embedding* capaz de extraer representaciones útiles e interpretables de textos a través de redes neuronales profundas, *variational autoencoding,* y la Asignación supervisada de Dirichlet latente (*SLDA* por sus siglas en inglés). El método propuesto es capaz de resolver tareas de aprendizaje supervisado, semi-supervisado y de aprendizaje multitarea. El método fue evaluado con un banco de preguntas de reseñas de clientes. Las métricas usadas para la evaluación fueron precisión, exactitud, exhaustividad, medida-F_1. Los resultados obtenidos superaron a métodos con un accuracy de 0.8157.

En la investigación realizada por Jin et al. [54] proponen un método para la integración del aprendizaje profundo con el descubrimiento de tópicos para la extracción de información de contexto. Los autores implementaron una arquitectura de red neuronal artificial (memoria a largo plazo, en inglés *Long Short-Term Memory*). Para el descubrimiento de tópicos implementan el modelo *LDA* integrándolo con la red neuronal en un marco de factorización matricial. Los autores realizaron pruebas con un conjunto de datos de *amazon* y evaluaron su modelo por medio de la métrica *Mean Square Error (MSE).* Los autores compararon los resultados obtenidos con los resultados obtenidos al aplicar algunos métodos como *Probabilistic Matrix Factorization (PMF), LDA* o *Convolutional Matrix Factorization (ConvMF)* y observaron que su modelo supera a todos en términos de MSE

De igual manera Shafqat et al. [117] proponen un método de descubrimiento de tópicos con redes neuronales profundas, es decir, *LSTM.* El modelo utilizó datos no textuales y de reseñas. Los datos textuales usaron el modelo *LDA* para descubrir vectores latentes de tópicos y adoptan la arquitectura *LSTM* para generar vectores

latentes de documentos. Los autores proponen una recopilación de datos (comentarios hechos por usuarios relacionados con un producto). El método fue evaluado por medio de la métrica de exactitud y superó a algunos de los revisados en la literatura por los autores. El objetivo principal de los autores fue implementar un sistema de recomendación que utilice el método de descubrimiento de tópicos.

En el trabajo de Srivastava y Sutton [125] exponen un método de inferencia basado en codificación automática de Bayes variacional (*AEVB*) para el modelo *LDA*. Los autores realizaron evaluaciones por medio de la métrica de coherencia del tópico, además utilizaron *Variational Document Model* logrando ajustar su método en aproximadamente un millón de documentos. Los córporas utilizados durante los experimentos fueron *20-Newsgroup* y *Reuters*. Los resultados obtenidos obtienen en términos de coherencia del tópico son los más altos en comparación con los modelos que se basan en *LDA* y el muestreo de *Gibbs* y NVDM *(Neural Variational Document Model)*.

De acuerdo con Bougteb et al. [12] proponen un método que incorpora el uso de un *autoencoder* para el descubrimiento de los tópicos presentes en un conjunto de *tweets* sobre noticias y empresas como *Apple, Google* y *Microsoft*. Los autores aplicaron el algoritmo *k-means++* para obtener el número de grupos necesarios para la clasificación de los tópicos. Posteriormente, el conjunto de datos fue utilizado con los algoritmos de agrupamiento *CluStream, DenStream* y *Dstream* mejorando los resultados al incorporar el autoencoder. Los autores evaluaron con las métricas de precisión, exhaustividad y medida-F_1. Los resultados obtenidos con el método propuesto fueron de 0.81, 0.86 y 0.83 en términos de precisión, recall y medida-F_1 respectivamente.

Otros autores como Jelodar et al. [52] exponen un método que analiza la asociación entre el sentimiento de los comentarios extraídos de la red social *reddit* sobre COVID-19 basado en el descubrimiento de tópicos aplicando el modelo *LDA* y *Gibbs sampling* a partir de las opiniones públicas. El modelo emplea una red neuronal recurrente *LSTM* para la clasificación de sentimientos de los comentarios sobre COVID-19. El objetivo de los autores fue demostrar la importancia de utilizar opiniones públicas y técnicas computacionales para guiar la toma de decisiones y la clasificación de comentarios de sentimiento sobre temas relacionados con COVID-19 en foros de atención médica.

3.1.3. Métodos con incrustación de palabras

Pandey [98] expone un método para el descubrimiento de sentimientos en torno a COVID-19 en las redes sociales. Los autores utilizan *BERT, BIOBERT* y *SCIBERT* para la clasificación de sentimientos sobre los comentarios existentes sobre COVID-19 en la red social *Reddit* en inglés. El método emplea *LDA* y el mapeo de *Gibbs*. Además de la búsqueda de palabras clave relacionadas con la salud mental, la creación de un bigrama y una red de palabras relevantes para formar una opinión general sobre varios temas.

En la investigación reportada por Chang y Hwang [17] exponen un método para el descubrimiento de tópicos multilingües, llamado *Cb-CLTM (cross-lingual topic model)*. El cual incorpora un modelo de incrustación de palabras en varios idiomas basado en el algoritmo *skipgram* con muestreo negativo. El modelo asume que cada documento tiene una distribución de tópicos representada como la distribución multinomial de Dirichlet. Primero construyeron un espacio de palabras monolingüe a través de diferentes técnicas como extracción de tokens y etiquetado de partes del discurso, eliminaron las palabras cerradas y entrenan los vectores de palabras solo con los sustantivos y verbos. Los experimentos realizados mostraron que el espacio de palabras en varios idiomas exhibe isomorfismo. El modelo fue capaz de obtener tópicos coherentes con una mayor diversidad e incluye la representación de documentos en diferentes idiomas para otras tareas. Los tópicos fueron descubiertos aplicando *LDA*. Los corpus utilizados fueron *Reuters* y el corpus llamado *UM* que está compuesto por oraciones en inglés y chino. La evaluación de los tópicos resultantes fue realizada con la métrica de coherencia del tópico normalizada, diversidad de tópicos y calidad de la representación de documentos multilingües. Los resultados obtenidos superaron a otros modelos existentes en la literatura.

En el trabajo de Scarpino et al. [115] expone una comparación de diferentes técnicas de modelado de tópicos. Los autores llevaron a cabo una comparación de diferentes técnicas para el modelado de tópicos, como *LDA* y el modelado de tópicos basado en el transformador *BERT*. El propósito es extraer información significativa en la narración italiana de la pandemia de COVID-19. Los conjuntos de datos empleados fueron 187 narraciones relacionadas con la COVID-19 de dos proyectos italianos: *R-Esistere* y Síndrome post COVID-19. El modelo BERTopic es aplicado para genera representaciones de temas a través de un proceso de cuatro pasos: convertir documentos a su representación incrustada utilizando un modelo de lenguaje

previamente entrenado, además realiza una reducción de dimensionalidad a través de UMAP [81] y aplica algoritmo de agrupamiento HDBSCAN para agrupar textos en grupos que tienen un significado similar. Luego utiliza una variación basada en clases de $TF - IDF$ para dar una representación vectorial de cada documento que refleje la importancia de cada palabra para recuperar las palabras más representativas de cada grupo/tópico. El rendimiento de los modelos fue evaluado en términos de coherencia del tópico y perplejidad. Los resultados muestran que el enfoque basado en $BERT$ supera al enfoque basado en LDA.

Las investigaciones continúan con Pita et al. [102], donde exponen un método llamado *Vec2Graph Topic Model (VGTM)*. Los autores crearon una representación basada en grafos utilizando incrustación de palabras. El modelo $VGTM$ lleva a cabo el descubrimiento de patrones en grafos de palabras utilizando el concepto de comunidades de grafos. Los autores se basaron en la premisa que, dado un grafo de incrustación de palabras, las comunidades o grupos de palabras son buenos discriminadores de tópicos. El modelo *Skip-gram* y similitud coseno son usados como métrica de similitud para generar el grafo semántico. Los autores señalan que $VGTM$ detecta comunidades de palabras en los tópicos que pueden superponerse. La métrica usada para evaluar el rendimiento del modelo propuesto es la coherencia del tópico.

Por otro lado, las investigaciones realizadas por Eslami et al. [30] proponen un método para encontrar la relación entre un fármaco y sus efectos secundarios según lo informado por los usuarios habituales de un sitio web llamado pregunte a un paciente. Los autores incorporaron en su modelo un algoritmo basado en aprendizaje profundo donde los comentarios de los usuarios sobre los efectos secundarios reflejaron que los medicamentos se encuentran sesgados. Cada comentario lo clasificaron haciendo uso del modelado de tópicos para la identificar eventos farmacológicos. El modelo fue evaluado por medio de las métricas precisión, medida-F_1, exactitud y coeficiente *kappa* [94]. Los fármacos involucrados fueron: para tratar neurosis, problemas de digestión y anticonceptivos. Los modelos *HAN*, *FastText*, *NMF* y *Word2Vec* fueron usados en el desarrollo del método.

La investigación hecha por He et al. [44] expone un método denominado modelo temático de Asignación de Dirichlet Latente Representativo ($RLDA$) para revelar las relaciones cercanas y complejas entre diabetes, obesidad y otras enfermedades. Los autores realizaron pruebas con un corpus de más de 337,000 publicaciones sobre diabetes y obesidad. Para revelar las relaciones significativas entre diabetes mellitus,

obesidad y otras enfermedades, realizaron un análisis explícito de los resultados de su modelo. Los resultados señalaron que en los últimos 10 años las enfermedades con estrecha relación con obesidad se encuentran asma, enfermedad gástrica y enfermedad cardiaca. El modelo *RLDA* se basó en *LDA*, *Word2Vec* y agrupación de propagación por afinidad.

Por su parte Xun et al. [146] exponen un método para el descubrimiento de tópicos para un corpus de texto corto usando un modelo de incrustación de palabras basado en $Word2vec$ con un corpus de $Wikipedia$. Las incrustaciones de palabras fueron incorporadas al modelo para proporcionar semántica adicional. Por lo tanto, modelaron cada documento corto como un tema gaussiano sobre incrustaciones de palabras en el espacio vectorial. Los autores consideraron necesario introducir un modo de fondo discreto sobre los tipos de palabras para complementar los tópicos continuos de Gauss. El corpus utilizado fue sobre títulos de noticias de fuentes de datos como *abcnews*. El modelo fue evaluado con las métricas precisión, *recall* y medida-F_1. Los autores realizaron experimentos y compararon los resultados obtenidos por su modelo con los resultados obtenidos al aplicar el modelo LDA y *Gaussian-LDA* y llegaron a la conclusión que su modelo es capaz de superar significativamente a ambos.

Wahid et al. [141] proponen un modelo para etiquetar datos a través de un enfoque de modelado de tópicos. Para construir un vector de características para clasificar datos contextuales incorpora el uso de LDA e incrustaciones de $BERT$. Los autores proponen una metodología basada en tres capas. La primera capa genera tópicos con LDA y posteriormente son clasificados y mapean el tópico mejor clasificado en una etiqueta para anotar los datos. La segunda capa, transforma el texto previamente etiquetado como características utilizando $BERT$. La tercera capa con aprendizaje profundo y clasificadores existentes obtienen múltiples categorías. Los resultados experimentales mostraron que el estudio mejora el rendimiento en comparación con otros enfoques existentes en la literatura.

Costa and Ortale [22] exponen un nuevo método de aprendizaje estadístico para combinar el modelado de tópicos y la agrupación de documentos. En particular desarrollaron un modelo generativo Bayesiano de colecciones de textos. El cual consistió en incrustaciones de palabras *(Word2Vec)*, para capturar las regularidades semánticas y sintácticas entre palabras. El enfoque hace uso del muestreo de Gibbs colapsado junto con la estimación de parámetros, con el propósito de realizar el modelado de tópicos y la agrupación de documentos a través del razonamiento Bayesiano. El en-

foque fue probado con dos corpus en inglés, el primero sobre el dominio financiero y el segundo sobre noticias. Los autores evaluaron su método con las métricas de exactitud e información mutua normalizada.

Otros autores como Gao et al. [35] exponen un método para el descubrimiento de tópicos en textos cortos que utiliza correlaciones semánticas globales y locales para descubrir tópicos significativos. Los autores lo llamaron modelo de tópico regularizado de campo aleatorio condicional (*CRFTM*). Las palabras relacionadas semánticamente comparten la misma asignación del tópico. El modelo de incrustación de palabras utilizado fue *Word2Vec*. El modelo aplica el muestreo de *Gibbs* y el campo aleatorio condicional para alentar a las palabras relacionadas semánticamente a compartir la misma etiqueta de tópico. Los conjuntos de datos empleados por los autores para los experimentos sobre su modelo fueron *stackOverflow* y noticias. El modelo propuesto incluyó el desarrollo de una nueva métrica para la distancia entre textos cortos. El modelo fue evaluado por medio de las métricas coherencia del tópico y exactitud.

En el trabajo desarrollado por Rivera y Torres-Moreno [106] exponen un método para la detección de neologismos semánticos *(SN)*. Los autores implementaron una combinación de descubrimiento de tópicos, extracción de palabras clave y desambiguación del sentido de las palabras. Los autores examinaron los modelos de incrustación de palabras: *Word2Vec, Sense2Vec* y *FastText* en español. Además, presentan una comparación de estos resultados con las concordancias de cada palabra. El modelo propuesto utiliza un modelo basado en *tf-idf* y regresión logística para la detección del idioma como catalán, francés y español. Una vez que el texto tiene un tópico asignado el sistema extraerá automáticamente las palabras clave usando *TextRank* con filtrado *PoS*. El corpus usado fue compilado utilizando artículos de publicaciones especializadas en español: *PC World*, deportes y el Financiero. El descubrimiento de los tópicos lo realizan por medio de un sistema denominado *DENISE* que evalúa mediante regresión logística para la predicción de los tópicos. El sistema es evaluado con las métricas F_1, precisión y exhaustividad.

Por su parte Li et al. [73] exponen un método temático sobre la base del modelo *Dirichlet Multinomial Mixture (DMM)* llamado *PDMM*. Cada texto se puede asociar con una pequeña cantidad de temas relevantes. Los resultados experimentales proporcionaron evidencia que *PDMM* ofrece una mejor precisión de clasificación y coherencia del tópico que *DMM*. Además, proponen dos nuevos modelos de temas

sobre la base de *DMM* y *PDMM* respectivamente llamados *GPU-DMM* y *GPU-PDMM*. Para el desarrollo de sus modelos utilizan el modelo Pólya para mejorar la similitud de tópico para dos palabras relacionadas semánticamente. Para incorporar las relaciones semánticas para el aprendizaje del tópico latente sobre textos breves incorporaron modelos de incrustación de palabras, que los autores nombraron auxiliares. El conjunto de datos empleado para los experimentos realizados fue *BaiduQA*. Los métodos fueron evaluados con las métricas de coherencia del tópico y exactitud.

La investigación expuesta por Moody [87] presenta un método que aprende vectores de palabras densas junto con mezclas de vectores temáticos a nivel de documentos latentes. El objetivo de los autores fue modificar *Skipgram Negative-Sampling*. El propósito es utilizar vectores de características en todo el documento mientras se obtienen pesos de documentos continuos. El modelo se basa en una ventana que contiene cinco *tokens* antes y después del *token* de pivote. Para cada par de palabras pivote-objetivo la palabra pivote se usa para predecir la palabra objetivo. Cada palabra se representa con un vector de representación distribuida de longitud fija utilizando los mismos vectores de palabras tanto en la representación de pivote como en el objetivo. El modelo es evaluado por medio de la métrica de coherencia del tópico.

Zhang et al. [153] presenta un método para el descubrimiento de tópicos con incrustación de palabras. El modelo fue llamado *LDA* de incrustación de contenido múltiple (Multi-Content Embedding LDA *MCeLDA*). El objetivo de los autores fue capturar las relaciones de la forma síntoma-génesis-hierba, hierba-hierba y síntoma-síntoma que pueden ser usadas en el diagnóstico y tratamiento de enfermedades en la medicina China tradicional. Los autores proponen dos modelos de incrustación basados en *word2vec*, el primero fue nombrado *symptom2vec* y el segundo *herb2vec*. Los cuales fueron usados durante la implementación del trabajo expuesto. El modelo propuesto por los autores fue capaz de codificar la similitud semántica de los síntomas y la similitud semántica de las hierbas. Los resultados experimentales en dos conjuntos de datos de casos médicos de *Traditional Chinese medicine (TCM)* demostraron la eficacia del modelo propuesto para analizar la patogenia y ayudaron a realizar diagnósticos y tratamientos en la práctica clínica. Los autores compararon el modelo propuesto con el modelo *LDA, MC-LDA* y *G-LDA*. El modelo fue evaluado con la métrica de coherencia del tópico.

Las investigaciones continúan con Shi et al. [121] proponen un método de factori-

zación de matriz no negativa asistida por semántica (SeaNMF) para el descubrimiento de tópicos para textos cortos. Los autores incorporaron correlaciones semánticas de tipo palabra-contexto. El modelo utiliza un algoritmo de descenso de coordenadas de bloque. Además, incorpora información semántica utilizando incrustaciones de palabras. Lo que permitió a $SeaNMF$ recuperar la coincidencia de palabras a partir de las relaciones semánticas entre palabras clave y sus contextos. Los conjuntos de datos utilizados fueron *TagNews, Yahoo.Ans, Tweets, DBLP, Yahoo.CA* y *ACM.IS*. El modelo de incrustación utilizado fue *word2vec*. Posteriormente, se evaluó el desempeño del modelo con la métrica de coherencia del tópico y exactitud.

Las investigaciones continúan en Yang y Tang [148] presentan un método de descubrimiento de tópicos de noticias basado en grafos semánticos de cápsula (CSG). Los autores aplicaron *TextRank* para extraer frases clave del texto y cada una fue modelada como un grafo de palabras clave. Cada grafo está dividido en múltiples subgrafos a través de la detección de comunidades. Los autores emplearon un método de agrupamiento incremental para agrupar textos de noticias. Cada texto está representado por CSG, y la similitud entre los textos se calcula mediante el kernel del grafo. El descubrimiento de tópicos se lleva a cabo procesando cada documento y posteriormente generan grupos. El algoritmo devuelve un conjunto de grupos de tópicos.

Por otro lado, Yu et al. [149] muestran un método de tópicos denominado asignación jerárquica de Dirichlet latente de *Twitter* ($thLDA$). El modelo tiene como objetivo extraer automáticamente los tópicos existentes en un conjunto de tweets. El modelo $thLDA$ utiliza $word2vec$ para analizar relaciones semánticas de las palabras en cada tweet analizado. Para el descubrimiento de tópicos identifican la relación entre los usuarios y los tweets. Posteriormente construyeron un modelo jerárquico en función de la distribución de probabilidad y del modelo LDA. La evaluación del desempeño del modelo es realizada con la métrica de coherencia del tópico y perplejidad.

En las investigaciones realizadas por Pipanmekaporn et al. [101] presentan un método para el modelado de textos cortos. El modelo se basa en la factorización matricial no negativa ($NNMF$) y descenso de gradiente. El modelo incorpora relaciones semánticas entre palabras generadas por métodos de aprendizaje no supervisado. El descubrimiento de tópicos se realiza mediante el uso de una matriz de palabras generada por incrustación de palabras e incorporando relaciones semánticas entre

palabras capturadas por algoritmos de aprendizaje no supervisados. Los autores realizaron experimentos en tareas como el análisis de sentimientos y la clasificación de títulos de noticias. Los conjuntos de datos empleados fueron *Sentiment140* y *20-Newsgroup*. Los autores extrajeron de *twitter* un conjunto de datos compuesto de 1,600,000 documentos. Además, llevaron a cabo una comparación con los modelos del estado del arte como *LDA, Non-Negative Matrix Factorization (NNMF)* y *Biterm Topic Model (BTM)*. De esta manera los autores comprobaron que su método es capaz de obtener resultados mayores. Las métricas de evaluación para medir el rendimiento del modelo propuesto fueron precisión, exactitud y medida F_1.

Liu et al. [76] exponen un método sobre el problema de la incrustación de *hashtags* combinando el contenido de los textos cortos con las diversas relaciones heterogéneas en las redes sociales. Los autores establecieron una red con *hashtags* como sus nodos. Cada nodo de *hashtag* está asociado con un conjunto de *tweets* y cada *tweet* contiene un conjunto de palabras. Posteriormente diseñaron un modelo de incrustación denominado $Hashtag2Vec$. El modelo explota múltiples relaciones de *hashtag-hashtag, hashtag-tweet, tweet-palabra* y palabra-palabra basadas en la red heterogénea jerárquica. El conjunto de datos utilizado fue extraído de la red social *twitter*. Los *tweets* datan de los años 2011 y 2015. Para la construcción del modelo de incrustación aplican un algoritmo de agrupamiento y de descubrimiento de tópicos. Para el descubrimiento de tópicos aplicaron el modelo *LDA* obteniendo resultados altos comparados con la literatura en textos cortos. El desempeño del modelo lo evaluaron con la métrica de coherencia del tópico.

Por otro lado, Qin et al. [105] proponen un método colaborativo basado en la factorización de matrices no negativas para determinar de forma precisa los tópicos e incrustaciones de palabras específicas del dominio. Los autores desarrollaron un grafo de conocimiento basado en relaciones semánticas entre palabras, para acumular información de contexto global descubierta por modelos de descubrimiento de tópicos e información de contexto local reflejada por incrustaciones de palabras de contexto. Además, desarrollaron un grafo de subpalabras basado en la codificación de relaciones de palabras por pares para explotar la información de subpalabras de las palabras en el corpus actual del dominio. El conjunto de datos utilizado es *Amazon Review*. La métrica de evaluación utilizada para medir el desempeño del modelo expuesto es la coherencia del tópico normalizada.

Guo et al. [41] proponen un método de descubrimiento de tópicos basado en el

modelo LDA, denominado *CV-LDA (LDA* basado en vectores de palabras sensibles al contexto). El método se basa en la incrustación de palabras que genera un vector sensible al contexto. El propósito es agrupar las palabras para disminuir la dimensionalidad. Además, de algoritmos de similitud para calcular la similitud entre los vectores de palabras. El algoritmo K-means es utilizado para agrupar vectores por similitud. Posteriormente, los grupos resultantes son los datos de entrada al modelo LDA para generar tópicos por grupos. La evaluación del desempeño del método la llevaron a cabo con la métrica de perplejidad. Los experimentos realizados mostraron que su método tiene una menor perplejidad y una capacidad de respuesta satisfactoria.

Las investigaciones realizadas por Gupta y Katarya [42] proponen un enfoque basado en incrustaciones de palabras de $Word2vec$. Los autores recopilaron información morfológica y aplican el análisis de componentes principales (PCA) calculada sobre la similitud de palabras. Posteriormente aplicaron el algoritmo de agrupamiento K-means con un centroide que representó el tópico de cada grupo. Las palabras clave relacionadas se enriquecen con información morfológica. La evaluación se realizó mediante la clasificación de documentos en varios conjuntos de datos. Los conjuntos de datos utilizados fueron *20-Newsgroup* y *Reuters*. El rendimiento del enfoque propuesto fue comparado con los métodos tradicionales de representación de documentos, como el promedio de *word2vec, doc2vec, LSA, fastText* y bolsa de palabras.

De igual manera Kawamae [59] expone un método que aprende incrustaciones de tópicos y de palabras para mejorar el desempeño de un modelo de lenguaje probabilístico natural. El método recibe como entrada incrustaciones que representan la semántica latente de cada palabra y por lo tanto capturan las propiedades sintácticas y semánticas. El enfoque lo nombraron $Wat2vec$, es decir, incrustación de palabras y tópicos que aprende representaciones de palabras distribuidas junto con representaciones de tópicos distribuidos. El método *Wat2vec* emplea la propiedad de *skip-gram* donde predice palabras dada una palabra objetivo en una ventana deslizante y a su vez predice palabras circundantes. Por lo que $Wat2vec$ aprende incrustaciones simétricamente para explotar la información semántica y contextual. El enfoque permite que cada par de palabras-tópico posean sus propios parámetros de incrustaciones y comparta dichos parámetros entre las mismas palabras y tópicos. Los conjuntos de datos utilizados fueron *20-Newsgroup* y *Amazon* review.

En la investigación realizada por Navarro-Colorado [89] exponen el desarrollo de

dos pruebas una con *LDA* estándar implementado en *MALLET* y otra con *LF-LDA* (versión específica de *LDA*) desarrollada para documentos cortos y basada en incrustaciones de palabras. Los vectores de características fueron generados con *word2Vec* con una ventana de contexto de 5 palabras. La diferencia con *LDA* estándar fue que cada palabra se representa a través de un vector de características (basado en *word2Vec*) en un contexto de 5 palabras en todo el corpus. Por lo tanto, *LF-LDA* extrajo tópicos de *LDA* buscando relaciones de palabras dentro de cada poema individual o a lo largo de todo el corpus. Los autores miden la coherencia de los tópicos usando dos técnicas comunes: la primera especifica la coherencia de los tópicos calculando el *nPMI* de los tópicos de palabras; la segunda lo hace en base al descubrimiento de una palabra intrusa entre los tópicos de palabras. Los resultados muestran, por un lado, que cuando se trata de un corpus poético no es aconsejable la lematización del texto porque en el proceso se pierden rasgos poéticos; y, por otro lado, que un algoritmo *LDA* estándar es mejor que una versión específica de *LDA* para textos cortos *(LF-LDA)*. Los tópicos descubiertos con *LDA* los analizaron manualmente para definir la relación entre tópicos de palabras y poemas. El análisis mostró que existen principalmente dos tipos de relaciones semánticas: un tópico *LDA* podría representar el sujeto o tópico del poema, pero también podría representar un motivo poético. Los experimentos son llevados a cabo en un corpus de poemas del siglo de oro.

En el trabajo de García-Díaz et al. [36] proponen el uso de incrustaciones de palabras para capturar propiedades significativas de las palabras junto con sus relaciones semánticas que se utilizan para alimentar una red neuronal con el fin de aprender a distinguir entre textos positivos, negativos o neutrales. Los autores compilan y etiquetan un corpus compuesto por 7,435 *tweets* de economistas y sitios de noticias financieras. Los resultados indican que el modelo *fastText* entrenado con los corpora en español no anotados y junto con las características lingüísticas logró obtener una precisión de 58.036 %.

3.2. Métodos para la extracción de relaciones semánticas

A continuación, se presentan cuatro enfoques orientados a la extracción de relaciones semánticas. El primero son métodos que solo incorporan la extracción de características semánticas y sintácticas. El segundo expone trabajos que incorporan junto con las características anteriormente mencionadas un modelo de aprendizaje profundo como *variational autoencoding*, redes neuronales profundas, *Long Short-Term Memory* o *autoencoder*. El tercer enfoque expone trabajos que extraen relaciones semánticas utilizando modelos de incrustación de palabras como *word2vec, BERT, Sense2Vec* y *FastText*. El cuarto enfoque expone trabajos que incorporan modelos de incrustación de palabras y una arquitectura de aprendizaje profundo.

3.2.1. Métodos basados en características semánticas y sintácticas

Las investigaciones hechas por Zhang et al. [152] exponen un método de minería de relaciones semánticas entre genes, trastornos y fármacos provenientes de diferentes conjuntos de datos biomédicos. Los autores usan documentos que están formados por información correspondiente a la enfermedad de Parkinson como un caso de estudio y se enfocan en extraer las relaciones entre el trastorno, gen y el fármaco de la enfermedad a partir de cuatro conjuntos de datos biomédicos. Los conjuntos de datos utilizados en las pruebas realizadas fue *SemMedDB, KEGG, Uniprot* y *PharmGKB*. Los corpora contienen patrones de dominio para trastornos, sustancias químicas y fármacos, genes y secuencias moleculares. Cada conjunto de datos fue convertido en formato *RDF* y posteriormente propusieron un algoritmo para extraer relaciones semánticas de cada conjunto de datos. Las variables que el algoritmo almacena son predicados, sujetos y objetos. El método fue evaluado con la métrica de precisión obteniendo desde el 82 % hasta el 100 %.

Por otro lado, León-Araúz et al. [63] proponen un método para el desarrollo de gramáticas con el objetivo de extraer relaciones semánticas usadas en el campo genérico-específico, parte-todo, ubicación, causa y función. El objetivo de los autores es proporcionar un método accesible y de fácil uso para encontrar contextos ricos en conocimiento. El método propuesto busca cada relación en la que participa cada pa-

labra. El método genera "bocetos" de palabras que representan diferentes relaciones como verbo-objeto, modificadores o frases preposicionales. Los resultados experimentales fueron hechos con el corpus *EcoLexicon* en inglés de dominio medioambiental. Los autores generan una formalización de patrones gramaticales en forma de expresiones regulares combinadas con etiquetas *PoS*. En total generaron 56 diferentes gramáticas considerando diferentes aspectos específicos en cada relación. El método fue evaluado por medio de las métricas precisión y exhaustividad para evaluar la calidad de las gramáticas generadas y encontrar el equilibrio adecuado.

Por su parte Bentrcia et al. [10] proponen un método híbrido que tiene como objetivo enriquecer la construcción automática de una ontología del Corán. Los autores explotaron los patrones conjuntivos árabes que existen en la gramática árabe tradicional. El método extrae frases conjuntivas y relaciones semánticas apoyándose de la regla de la conjunción propia de la gramática árabe. El método utiliza un conjunto de patrones para la extracción de frases conjuntivas. Los autores llevan a cabo una combinación de pruebas estadísticas y resultados previos obtenidos por expertos en el dominio. Además, obtienen tres categorías diferentes de relaciones semánticas del Corán que son antonimia, género y clase. Posteriormente extraen un conjunto de palabras que forman términos de una ontología. Los autores utilizan un método de filtrado basado en un coeficiente de correlación para seleccionar relaciones sólidas. El método propuesto cuenta con una fase de extracción de términos que incluyó la extracción de sustantivos, nombres propios y adjetivos en su forma raíz. Los autores proponen 10 diferentes tipos de patrones compuestos de sustantivos, determinantes, adjetivos y la conjunción "y". Las métricas de evaluación utilizadas fueron precisión y exhaustividad obteniendo un 84 % y 92 % respectivamente.

En la investigación realizada por Ta and Thi [127] exponen un método que combina el *PLN* y técnicas estadísticas para la extracción de relaciones semánticas de documentos extraídos de la librería digital *ACM*. El objetivo es el enriquecimiento de una ontología de dominio. Las relaciones semánticas extraídas fueron sinónimos, hipónimos, hiperónimos, parte de, hecho de, atributo de, delimitado por, tiene lugar en, resultado de y afecta. La base de datos léxica *WordNet* es utilizada para la construcción de relaciones de sinónimos, hipónimos e hiperónimos. Para la construcción de las demás relaciones proponen un algoritmo diferente que refina la oración, es decir, elimina palabras innecesarias en la oración en función del árbol de dependencias generado. La evaluación es llevada a cabo por las métricas de precisión, exhaustividad

y medida-F_1.

En la investigación realizada por Shanidze y Petrasova [120] proponen un método para la extracción de relaciones semánticas con un enfoque basado en reglas. Los autores sugieren identificar verbos entre un sujeto y un objeto para obtener una secuencia de relaciones semánticas de *Wikipedia*. Por lo que emplearon el etiquetador de partes del discurso *treetagger*. Además, emplean *synsets* presentes en *WordNet* para la extracción de relaciones semánticas entre conceptos y sus sinónimos del corpus de texto. El algoritmo procesa 200 artículos de *Wikipedia* del dominio de tecnologías de la información. Por lo cual identifica verbos entre sujeto y objeto de expresiones que son asumidas como relaciones semánticas. La búsqueda se amplía a sujetos y objetos de los predicados y de esta manera obtienen los sinónimos desde *WordNet* de los sujetos y objetos identificados previamente.

Al-Zaidy and Giles [2] exponen un método para la extracción de relaciones semánticas entre entidades en artículos académicos haciendo uso de patrones sintácticos extraídos de la literatura de hipónimo-hiperónimo centrándose en el resumen y palabras clave de los documentos a analizar. El sistema extrae entidades semánticas como conceptos e instancias con sus atributos del texto completo. Las entidades extraídas fueron sustantivos y frases nominales. Las fuentes de datos externas de la web, como el grafo de conceptos de *Microsoft* fueron usados para la evaluación de la calidad de los conceptos y relaciones extraídas. Los conceptos fueron usados para construir una taxonomía científica que cubra el contenido de investigación de los documentos. Un conjunto de diez mil documentos académicos fue utilizado para pruebas y evaluación. Las métricas utilizadas para evaluar el desempeño del sistema fueron precisión y exhaustividad

3.2.2. Métodos con un enfoque de aprendizaje profundo

En el trabajo de Suárez-Paniagua et al. [126] exponen un método para la extracción de relaciones semánticas entre dos frases clave identificadas por medio de la extracción de patrones de sinónimos. Los patrones extraídos fueron los fragmentos de texto entre pares de sinónimos. Los autores usaron una red neuronal convolucional (*CNN*) entrenada con el conjunto de datos de entrenamiento proporcionado por *SemEval*. El corpus usado contiene 500 artículos de revistas sobre informática, ciencias de los materiales y física de *ScienceDirect*. El modelo de aprendizaje profundo reportó resultados exitosos para la extracción de relaciones semánticas. El

corpus completo contiene 500 artículos de revistas sobre informática, ciencias de los materiales y física de *ScienceDirect.*

Por su parte Lee et al. [62] exponen un método para la extracción de relaciones semánticas con aprendizaje profundo. Los autores proponen el uso de una red neuronal convolucional (*CNN* por sus siglas en inglés) para la extracción de relaciones presentes en textos científicos. El conjunto de datos empleado es proporcionado por *SemEval-2017*. Los autores proponen un pre-procesamiento que toma como entrada un texto y la ubicación de todas las entidades presentes en el. Cada texto es representado de 3 maneras diferentes, primero como una lista de fichas con cuatro características, es decir, las posiciones relativas de las dos entidades y sus tipos de entidad y etiqueta correspondiente a la parte del discurso. Los datos de entrada que alimenta la *CNN* fueron preprocesados. La *CNN* fue formada por 4 capas principales. Las relaciones extraídas por los autores con el modelo propuesto fueron hiperonimia, hiponimia y sinonimia. Los resultados obtenidos mostraron que la *CNN* fue capaz de identificar las relaciones de manera eficiente. Las métricas de evaluación utilizadas son precisión, exhaustividad y medida-F_1. Los resultados obtenidos en precisión para relaciones de sinonimia son de 0.820 y para relaciones de hiponimia de 0.455.

3.2.3. Modelos con incrustación de palabras

Vintar et al. [140] exponen un método para la extracción de palabras que expresan una relación semántica por medio de incrustación de palabras. Los autores utilizan un conjunto de semillas (adjetivos) para recuperar palabras cercanas por medio de incrustaciones de palabras y posteriormente intersecciones de los vecindarios resultantes. El uso de incrustaciones de palabras permitió identificar palabras que pertenecen a la misma relación. El dominio estudiado fue accidentes geográficos, que son descritos a través de su forma, ubicación, causa, función y composición. El idioma empleado fue inglés y croata. Los autores utilizaron el modelo de incrustación *fastTex*. La métrica usada para evaluar los resultados obtenidos fue precisión obteniendo resultados que van de 0.28 a 0.80.

Shah et al. [118] proponen un método para la extracción de las limitaciones presentes en las relaciones semánticas en un corpus de dominio. Los autores establecen un modelo de restricciones de desigualdad de relaciones. Si la entidad a está relacionada con la entidad b con el tipo de relación w y no está relacionada con la entidad c por la misma relación entonces está semánticamente más cerca de b que dé c. El

modelo presentó ventajas como: no limita sólo a sinonimia y antonimia, sino que también se pueden generar a partir de relaciones léxicas como hiperonimia y holonimia. También de las relaciones no léxicas y grafos de conocimiento. Posteriormente las restricciones de desigualdad son usadas como datos de entrenamiento para el mapeo de incrustaciones de palabras previamente entrenadas en un espacio vectorial. Los autores incorporaron tres modelos de incrustación previamente entrenados, que son *glove, word2vec* y *fastText*. La evaluación fue llevada a cabo con dos conjuntos de datos *SimLex-999* y *Sim-Verb* y con el coeficiente de correlación de *Spearman*. Para *glove* el mayor puntaje obtenido fue de 0.63, para *fastText* 0.59 y para *word2vec* de 0.58.

En la investigación realizada por Liu et al. [77] exponen los resultados obtenidos al participar en *SemEval 2017*. Los autores se enfocaron en la tarea de clasificación de frases clave. Además, exploraron patrones de frases clave en publicaciones científicas utilizando modelos de incrustación de palabras previamente entrenados. El modelo propuesto utiliza como método de aprendizaje automático máquinas de soporte vectorial para la clasificación de frases clave. El sistema obtuvo una puntuación F_1 de 0.67 para la clasificación de frases clave y 0.64 para la detección de relaciones de sinonimia. El modelo usa un sistema basado en reglas que considera pares de entidades como candidatos de relación extrayendo los textos de contexto entre dos entidades. Las relaciones que compartían al menos una entidad los agruparon como una relación de acuerdo con el requisito del formato de salida proporcionado por la competencia.

3.2.4. Modelos con incrustación de palabras y aprendizaje profundo

Dima y Hinrichs [26] proponen la interpretación automática de compuestos de nombres en inglés por medio de la extracción de relaciones semánticas, un modelo redes neuronales profundas y un modelo de incrustación de palabras. Los modelos de incrustación utilizados fueron *word2vec, glove, CW* y *HPCA*. Las relaciones extraídas fueron *whole-part_or_member_of* y *location*. El conjunto de datos utilizado pertenece al dominio de la salud. Las métricas utilizadas para la evaluación fueron medida-F_1 y exactitud. Los autores reportan que la interpretación propuesta obtiene resultados significativos en comparación de los trabajos reportados donde solo

aplican un modelo de incrustación de palabras.

Patel et al. [99] emplean una red neuronal convolucional y un modelo de incrustación de palabras *word2vec* para clasificar relaciones semánticas del dominio médico. La red utilizada cuenta con una capa de incrustación de palabras utilizando el conjunto de datos clínicos proporcionados por la competencia *i2b2* de 2010 el cual consiste en 9,070 relaciones. Las relaciones extraídas son de tipo problema-tratamiento, problema-pruebas y problema-problema. Las métricas utilizadas en la evaluación de la propuesta de los autores fueron precisión y exhaustividad. El sistema logró una precisión de 75 % y 72 % de exhaustividad. El rendimiento del sistema fue comparado con diferentes sistemas mostrando mejores resultados con el método propuesto.

En el estado del arte se han revisado diferentes métodos para el descubrimiento de tópicos. Los autores proponen métodos que descubren tópicos en textos en diferentes idiomas entre ellos español e inglés y de diferentes longitudes. Tales como los córpora formados por mensajes extraídos de redes sociales, periódicos, opiniones, encuestas entre otros. Algunos autores involucran el uso de un enfoque de aprendizaje profundo y/o modelos de incrustación de palabras, pero sin incorporar la semántica de los textos.

Por otro lado, cabe recalcar que se encontró un trabajo que desarrolla un modelo de incrustación de relaciones semánticas extraídas de la base de datos léxica *WordNet*, aplicando la factorización de matrices, pero orientado a la tarea de similitud semántica.

Por lo que, se revisaron trabajos que extraen relaciones semánticas de diferentes corpus en diferentes idiomas. Los trabajos revisados emplean características semánticas y/o sintácticas, modelos de aprendizaje profundo o modelos de incrustación de palabras. Sin embargo, solo dos autores incorporan un modelo de aprendizaje profundo y de incrustación de palabras en su metodología.

En este trabajo se propone el desarrollo de un modelo de incrustación de relaciones semánticas para el descubrimiento de tópicos. El cual involucra la integración de la extracción de relaciones semánticas de sinonimia, hiponimia e hiperonimia, el desarrollo de un modelo de incrustación de relaciones semánticas, clasificación de textos empleando el modelo de incrustación de relaciones semánticas y el descubrimiento de tópicos.

Capítulo 4

Metodología de solución

En esta sección se presenta la metodología propuesta para el descubrimiento de tópicos. La cual es la tarea principal de esta tesis doctoral.

Los procesos contenidos en esta metodología son: la extracción de relaciones semánticas existentes en un corpus de *Wikipedia* en inglés. Las cuales se extraen por medio de patrones léxico-sintácticos. Las relaciones semánticas obtenidas son la base fundamental para generar un modelo de incrustación basado en relaciones semánticas. Posteriormente se aplica un proceso de clasificación de texto empleando el modelo de incrustación de relaciones previamente desarrollado. Finalmente, de las clases obtenidas en la tarea de clasificación se descubren los tópicos específicos existentes en ellas.

En la Figura 4.1 se expone gráficamente la metodología general propuesta para el desarrollo de la metodología de solución.

En las siguientes secciones se describe detalladamente cada proceso expuesto en la metodología general.

4.1. Pre-procesamiento de textos

Los córpora para el descubrimiento de tópicos es decir, *20-Newsgroup* y *Reuters* están formados por documentos del dominio de las noticias en lenguaje natural.

Dado que son textos escritos en lenguaje natural y con la estructura necesaria para transmitir una idea, cuentan con texto que aún cuando tienen la capacidad de dar sentido a una oración, no aportan información importante en tareas de PLN. Por lo tanto, es necesario eliminar algunas partes de cada oración que forma al corpus.

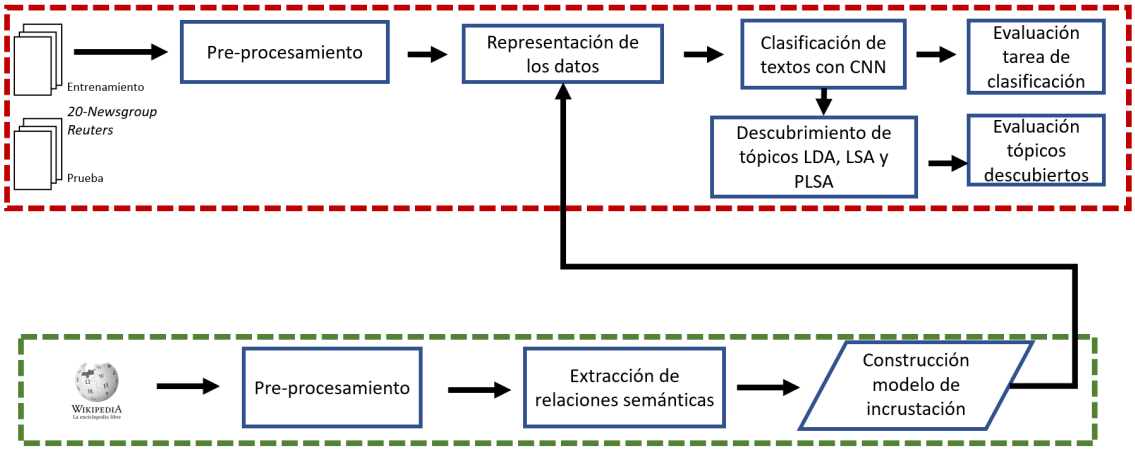


Figura 4.1: Metodología general propuesta.

Por lo que es necesario aplicar pasos para "homogenizar" los datos. Los pasos aplicados son:

- Retirar símbolos no *ascii*
- Convertir mayúsculas a minúsculas: con la finalidad de normalizar el vocabulario.
- Eliminar elementos: XML, palabras cerradas (a, an, with, another, etc).

Por otro lado, el corpus a utilizar para la extracción de las relaciones semánticas es *Wikipedia* en inglés. Por lo que se le aplicará el siguiente pre-procesamiento:

- Extracción de contenido existente en <text>< /text>
- Retirar símbolos no *ascii*.
- Segmentación: separar un texto en segmentos más cortos como palabras u oraciones.

4.2. Extracción de relaciones semánticas

La extracción de relaciones semánticas se realiza a partir de un corpus de *Wikipedia* en inglés. Este corpus es empleado debido a que representa un amplio vocabulario y por lo tanto puede ser considerado como una base de conocimiento general. Por

lo que las relaciones semánticas resultantes son la base para construir un modelo de incrustación de relaciones semánticas.

Sin embargo, *Wikipedia* es un corpus que carece de conjuntos de datos etiquetados con relaciones semánticas. Por lo tanto, para extraer las relaciones semánticas de sinonimia, hiponimia e hiperonimia se utilizarán patrones léxico sintácticos.

Los patrones se convirtieron en expresiones regulares en el lenguaje de programación Python y se almacenaron en un repositorio. El resultado son conjuntos de patrones para sinonimia de [50, 132, 122] y para hiponimia-hiperonimia de [46, 96, 85, 135, 134, 133, 18].

Al corpus de *Wikipedia* se le aplicó un pre-procesamiento previo, como eliminar caracteres no *ascii* y su conversión a minúsculas. Los patrones se aplican a los textos de *Wikipedia* para obtener conjuntos de pares de palabras para cada relación semántica.

Algunos patrones obtenidos aplicados en este trabajo se muestran en las Tablas 4.1 y 4.2.

Concepto 1	Relación	Concepto 2
X	also called	Y
X	called as	Y
X	also known as	Y
X	usually called	Y
X	is called	Y
X	are called	Y
X	sometimes called	Y
X	know as	Y
X	also referred to as	Y
X	often described	Y
X	commonly known as	Y
X	also named as	Y
X	abbreviated as	Y
X	commonly called as	Y
X	is often referred to as	Y
X	is referred to as	Y
X	alias	Y
X	aka	Y
X	as known as	Y
X	frequently abbreviated as	Y
X	called as	Y
X	commonly known as	Y
X	anciently named as	Y

Tabla 4.1: Patrones léxico sintácticos para relaciones de sinonimia [71]

Concepto 1	Relación	Concepto 2
X	such as	*Y*
X	as	*Y*
X	other	*Y*
X	include	*Y*
X	especially	*Y*
X	be	*Y*
X	like	*Y*
X	like other	*Y*
X	one of the	*Y*
X	one of these	*Y*
X	one of those	*Y*
X	be example of	*Y*
X	for example	*Y*
X	which be call	*Y*
X	which be name	*Y*
X	mainly	*Y*
X	mostly	*Y*
X	notably	*Y*
X	particularly	*Y*
X	principally	*Y*
X	in particular	*Y*
X	except	*Y*
X	other than	*Y*

Tabla 4.2: Patrones léxico sintácticos para relaciones de hipónimo-hiperónimo [71]

El número de relaciones para sinonimia e hiponimia-hiperonimia extraídas de *Wikipedia* se muestra en la Tabla 4.3.

Tabla 4.3: Relaciones semánticas extraídas [71]

Relación	Total
Sinónimo	1,200,000
Hipónimo-hiperónimo	6,966,042

Los conjuntos de pares de palabras para la relación semántica descubierta se utilizan para representarlos en tres modelos de incrustación de relaciones semánticas. A cada palabra que compone la relación semántica se le asigna un *id* único. Los modelos son utilizados para la clasificación de los córpora *Reuters* y *20-Newsgroup*.

4.3. Construcción de modelo de incrustación

Un modelo de incrustación de palabras incorpora un corpus para generar un vector numérico correspondiente a cada palabra que lo forma. Lo que genera un modelo de representación del lenguaje que se puede utilizar para una amplia gama de procesos entre ellos la clasificación de documentos.

En la literatura existe evidencia de un modelo de incrustación de relaciones semánticas que está formado por las relaciones semánticas existentes en la base de datos léxica *WordNet* y posteriormente proponen aplicar un proceso de factorización de matrices. Dicho modelo de incrustación ha sido parte de experimentos en tareas de similitud semántica, lo que permitió a los autores proporcionar resultados tangibles sobre la importancia de agregar semántica en los modelos de representación de datos.

Por lo que, en esta investigación se propone generar un modelo de incrustación de relaciones semánticas de sinonimia, hiponimia e hiperonimia existentes en un corpus de *Wikipedia* en inglés. Una vez extraídas las relaciones semánticas se aplica el procedimiento de factorización de matrices propuesto en [109].

Por lo que dado un conjunto de relaciones $F(x, y)$ se le asigna un identificador de tipo $F(i, j)$ para construir tres modelos de incrustación denominados M aplicando el sistema presentado en 4.1. En consecuencia por cada par en $F(x, y)$:

$$M(x,y) = \begin{cases} 1 & si\ x\ esta\ relacionado\ con\ y \\ 0 & en\ caso\ contrario \end{cases} \tag{4.1}$$

Donde $F(x, y)$ es la existencia de la relación (x, y) en el corpus de relaciones semánticas. Por cada par de palabras que forman una relación semántica se asigna un 1 al valor almacenado en la matriz en la posición (i, j).

Por ejemplo:
Los identificadores en función de la posición de cada palabra que forma una relación semántica se asignan de la siguiente manera:

$$\begin{aligned} &F(cat, feline) = F(0,1) \\ &F(cat, mammal) = F(0,2) \\ &F(mammal, feline) = F(2,1) \\ &F(mammal, dog) = F(2,3) \\ &F(elephant, mammal) = F(4,2) \\ &F(cat, animal) = F(0,5) \\ &... = ... \end{aligned} \tag{4.2}$$

Por lo que, la matriz (ver Tabla 4.4) representa las relaciones mostradas en 4.2.

El desarrollo de tres modelos de incrustación de relaciones semánticas son pro-

Tabla 4.4: Ejemplo de la representación de la matriz de relaciones semánticas $M(x, y)$

	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**...**
0	0	**1**	**1**	0	0	**1**	0	0	0
1	**1**	0	**1**	0	0	0	0	0	0
2	**1**	**1**	0	**1**	**1**	0	0	0	0
3	0	0	**1**	0	0	0	0	0	0
4	0	0	**1**	0	0	0	0	0	0
5	**1**	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0
...	0	0	0	0	0	0	0	0	0

puestos, es decir, se generaron tres matrices $M(x, y)$. Sin embargo, los modelos incluyen las relaciones más frecuentes del vocabulario, por lo que se seleccionaron 40,000 relaciones utilizando *tf-idf*.

Por ejemplo, en la primera matriz $M(x, y)$ las palabras que forman la relación semántica son de sinonimia como: *cat* y *feline*. La relación semántica es simétrica, es decir se agrega la posición $M(x, y)$ y $M(y, x)$.

La segunda matriz $M(x, y)$ las palabras que forman la relación semántica son de hiponimia-hiperonimia como *cat* y *mammal*. La relación semántica también es simétrica es decir se agrega la posición $M(x, y)$ y $M(y, x)$.

Las relaciones semánticas de sinonimia, la hiponimia y la hiperonimia generan un importante aporte semántico, por lo que se propone generar un modelo con las tres relaciones semánticas. El resultado es una matriz $M(x, y)$ con las tres relaciones semánticas representadas como en el ejemplo 4.4.

El número de relaciones usadas en este modelo fue solo el 50 % de las usadas en el modelo que solo incluye sinónimos y el 50 % de las usadas en el modelo que solo incluye hiponimia e hiperonimia.

Para cada modelo de incrustación se generará la matriz de relaciones $M(x, y)$, es decir, se generan 3 matrices de relaciones semánticas. Posteriormente se aplica el siguiente procedimiento sobre cada matriz propuesta [109]:

1. Enriquecimiento de M: Aplicando la fórmula de centralidad de $Katz$ (ver Ecuación 4.3) se representa la semejanza entre cada relación identificada, pero que no están relacionadas directamente. El resultado es almacenado en una matriz

llamada M_G.

$$M_G = (I - \alpha M)^{-1} \tag{4.3}$$

dónde

a) I es la matriz identidad como factor neutro.

b) $M(x, y)$ es el matriz previamente creada.

c) α factor de decaimiento fijado en 0.75.

2. Frecuencia de las relaciones: M_G está sujeto a *Pointwise Mutual Information* (PMI ver ecuación 4.5). Dónde un PMI positivo significa que las relaciones ocurren con más frecuencia. Un PMI negativo significa que ocurren con menos frecuencia de lo esperado. Una matriz F con w filas y c columnas f_{ij} es el número de veces que w_i (fila i) ocurre en c_j (columna j).

$$\begin{aligned} P_{ij} &= \frac{f_{ij}}{\sum_{i=1}^{w}\sum_{i=1}^{c} f_{ij}} \\ P_{i*} &= \frac{\sum_{i=1}^{c} f_{ij}}{\sum_{i=1}^{w}\sum_{i=1}^{c} f_{ij}} \\ P_{*j} &= \frac{\sum_{i=1}^{w} f_{ij}}{\sum_{i=1}^{w}\sum_{i=1}^{c} f_{ij}} \end{aligned} \tag{4.4}$$

donde:

P_{ij} es la probabilidad de la fila i y la columna j

P_{i*} es la probabilidad de la fila i

P_{*j} es la probabilidad de la columna j

$$PMI_{ij} = \begin{cases} \log \frac{P_{ij}}{P_i P_j} & si\ PMI_{ij} > 0 \\ 0 & en\ caso\ contrario \end{cases} \tag{4.5}$$

3. Normalización: escala los vectores de M_G individualmente a una norma unitaria para que el vector tenga una longitud de uno. La norma aplicada es $L2$, también conocida como norma euclidiana (ver ecuación 4.6). Norma que permite considerar la frecuencia de las relaciones (palabras) entre sí. El proceso de normalización de un vector implica cambiar su longitud manteniendo su

dirección y su sentido.

$$\|x_1 ... x_n\| = \sqrt{x^2{}_1 + ... + x^2{}_n} \tag{4.6}$$

4. Reducción de componentes: Se aplica sobre la matriz M_G un análisis de componentes principales (PCA) como lo proponen en Martín et al. [80] el objetivo de esta tarea es reducir el tamaño de los vectores y establecer la dimensión del espacio semántico codificado en vectores de tamaño 300.

La Figura 4.2 expone gráficamente un ejemplo con la metodología descrita en los párrafos anteriores.

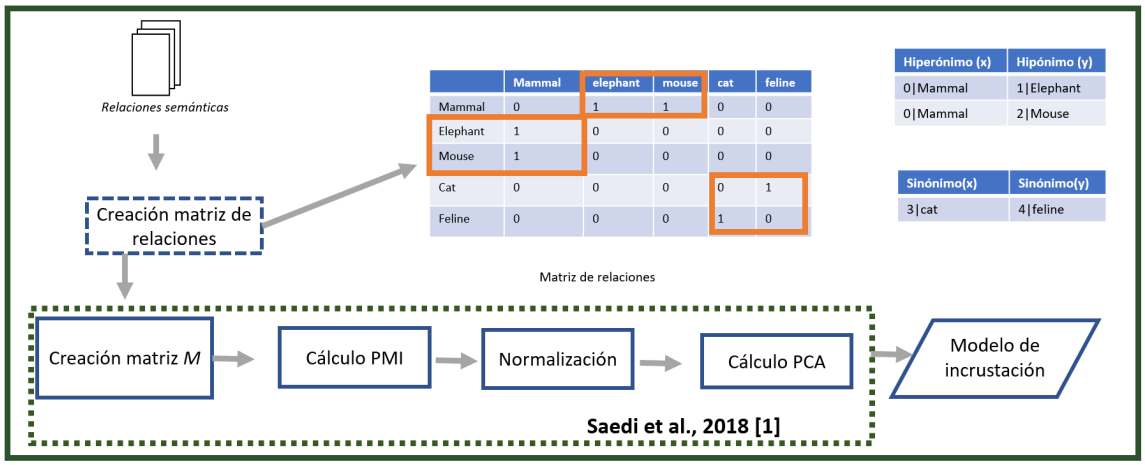


Figura 4.2: Metodología de creación de modelo de incrustación de relaciones semánticas. Elaboración propia [71]

El modelo de incrustación obtenido y que se representa en la Figura 4.2 se aplicará sobre la tarea de clasificación de los conjuntos de datos *20-Newsgroup* y *Reuters* obtenidos de la literatura. Por lo que la tarea de clasificación propuesta se presentará en la siguiente sección.

4.4. Clasificación de texto usando el modelo de incrustación y redes neuronales convolucionales

La tarea de clasificación de texto se aplica con el objetivo de obtener textos agrupados de acuerdo a su contenido. Dos córporas previamente pre-procesados son

clasificados utilizando una red neuronal convolucional descrita en la sección 2.6.1 y los modelos de incrustación de relaciones previamente obtenidos en la sección 4.3.

La tarea de clasificación de textos propuesta en este trabajo consiste de 5 fases expuestas a continuación:

- **Representación de datos:** en esta fase se utiliza el modelo de incrustación de relaciones semánticas desarrollados en la sección 4.3 para la representación de los datos. El texto original es representado en forma vectorial aplicando el modelo de incrustación de relaciones semánticas.

- **Modelo de clasificación:** esta fase emplea una red neuronal convolucional expuesta en la sección 2.6.1. La Figura 2.7 ilustra la arquitectura CNN empleada para la clasificación de texto. La cual está formada por una capa de incrustación de relaciones semánticas como capa de entrada, capas convolucionales 1D, capa de agrupación, capas completamente conectadas y capa de salida que son descritas a continuación:

  · La capa de incrustación: capa para incorporar un modelo de incrustación previamente entrenado con el propósito de llevar los textos a una representación vectorial. La CNN recibe el corpus de entrada e incrusta los vectores numéricos con los existentes en el modelo de incrustación con el que se esté trabajando.

  · La capa 1D: crea un núcleo que convoluciona con la entrada de la capa en una única dimensión para producir un tensor de salida. En este caso con 128 filtros, tamaño de kernel 5 y función de activación $Relu$.

  · La capa de concatenación: toma una lista de tensores como entrada y devuelve un solo tensor.

  · La capa *dropout*: o de abandono evita el sobreajuste al darle a cada neurona un 50 % de probabilidad de no activarse durante la fase de entrenamiento.

  · La capa $maxpooling1D$: reduce la muestra de la representación de entrada tomando el valor máximo sobre una ventana espacial de tamaño 5.

  · La capa $flatten$: llamada también capa de aplanamiento transforma un vector de entrada en un vector unidimensional.

· La capa *dense*: capa completamente conectada con una dimensionalidad de salida de 512 y función de activación *ReLu*.

– **Clasificación:** en esta fase se obtienen como resultado 20 clases pertenecientes al corpus *20-Newsgroup* y 90 clases para el corpus *Reuters*.

– **Evaluación de la tarea de clasificación:** en esta fase los resultados obtenidos al aplicar la arquitectura CNN son sometidos a una evaluación con las métricas de precisión, exhaustividad, medida-F_1 y exactitud descritas en la sección 2.7.

La Tabla 4.5 expone un ejemplo de las clases del corpus *Reuters* y *20-Newsgroup* respectivamente obtenidas al aplicar la *CNN* con el modelo de incrustación de relaciones semánticas previamente desarrollado.

Tabla 4.5: Ejemplo de las clases de los córpora *20-Newsgroups* y *Reuters*.

Corpus	*Clase*
20-Newsgroups	*...Atheism, Sport, Politics, Computing, Cars...*
Reuters	*...Aluminium, Barley, Bop, coffee, Cocoa...*

Posteriormente por cada clase extraída se integra la tarea de descubrimiento de tópicos. El objetivo es obtener tópicos específicos. En la siguiente sección, se expone el proceso propuesto para el descubrimiento de los tópicos.

4.5. Descubrimiento de tópicos

Las clases generadas en la fase anterior son los documentos de entrada en esta etapa. El objetivo es integrar el resultado de la clasificación de textos en el descubrimiento de los tópicos existentes y de esta manera tener información específica. Por lo que la tarea de descubrimiento de tópicos se lleva a cabo de la siguiente manera:

– **Corpus de clases:** las clases obtenidas en 4.4 de cada corpus son sometidas como texto de entrada a los modelos para el descubrimiento de tópicos.

– **Descubrimiento de tópicos:** los modelos *LDA, LSA* y *PLSA* previamente descritos en la sección 2.3 son empleados para la extracción de los tópicos presentes en el corpus de clases con los siguientes parámetros:

- El corpus *20-Newsgroup* se extraen 20, 50 y 100 tópicos con 10 palabras representativas.

- El corpus *Reuters* se extraen 20, 50 y 100 tópicos con 10 palabras representativas.

En la literatura los autores Qiang et al. [104] proponen como parámetro α (el parámetro de Dirichlet) con un valor predeterminado de 0.1 y β con un valor predeterminado de 0.01.

Las clases obtenidas después de clasificar los córpora *20-Newsgroup* y *Reuters* están formadas por palabras de un dominio particular, por lo que descubrir los tópicos existentes en los documentos que forman una clase proporcionan tópicos específicos según el dominio de la clase . A continuación, se presentan ejemplos de palabras representativas de los tópicos descubiertos. Estos ejemplos se muestran en las Tablas 4.6 y 4.7. Las cuales exponen un ejemplo de clases del corpus *Reuters* y *20-Newsgroups* respectivamente. Cada Tabla expone tres clases con 2 tópicos descubiertos en cada una. De cada tópico se exponen solo 3 palabras representativas de 10 obtenidas con los modelos *LDA, LSA* y *PLSA*.

Tabla 4.6: Ejemplo de palabras representativas del corpus *Reuters* con LDA, LSA y PLSA con 20 tópicos [70]

	Tópicos					
	LDA		**LSA**		**PLSA**	
Clase	**Company**	**Disasters**	**Company**	**Disasters**	**Company**	**Disasters**
Aluminum	*...operations, dump, ton...*	*...debris, Panamá, toll...*	*...dlrs, bank, group...*	*...debris, Portugal, industries,*	*...godless, jewish, sabbath...,*	*...group, ...law, love...,*
	Cultivation	**Grain**	**Cultivation**	**Grain**	**Cultivation**	**Grain**
Barley	*...acreage, wheat, department...*	*...acreage, corn, farm...*	*...maize, wheat, department...*	*...corn, corn, drum...*	*...production, system, acres...,*	*...tonnes, february, export...,*
	Finance	**Duty**	**Finance**	**Duty**	**Finance**	**Duty**
Bop	*...pressure, country, finance...*	*...dollar, oil, price...*	*...singapore, britoil, finance...*	*...japan, oil, price...*	*...gas, models, pay...*	*...battery, wheel, oil...*

Tabla 4.7: Ejemplo de palabras representativas del corpus *20-Newsgroups* con LDA, LSA y PLSA con 20 tópicos [70]

	Tópicos					
	LDA		**LSA**		**PLSA**	
Clase	**Religion**	**Rituals**	**Religion**	**Rituals**	**Religion**	**Rituals**
Atheism	*...town, big, life...*	*...bible, ceremonial, love...*	*...goodles, sabbath, ceremonial...*	*...course, started, hostage,*	*...godless, jewish, sabbath...,*	*...ceremonial, ...law, love...,*
	Software	**Hardware**	**Software**	**Hardware**	**Software**	**Hardware**
Computing	*...virtual, video, file...*	*...cpu, harddisk, machine...*	*...software, driver, email...*	*...circuits, video, pc...*	*...windows, system, copies...,*	*...mail, acceso, location...,*
	Elements	**Others**	**Elements**	**Others**	**Elements**	**Other**
Cars	*...battery, bad, street...*	*...video, computer, design...*	*...price, batteries, street...*	*...video, computer, concrete...*	*...gas, models, pay...*	*...software, alert, safety...*

– **Evaluación de la tarea de descubrimiento de tópicos:** los resultados obtenidos se evalúan con la métrica de coherencia del tópico normalizada (ver sección 2.7). La cual se basa en obtener la *NPMI* de cada par de palabras que pertenecen a las k palabras que representan a cada tópico. Por ejemplo, de las palabras representativas de la Tabla 4.7 tomando como palabras a evaluar *town* y *big*; la probabilidad de que las palabras $P(w_i, w_j)$ (P(*town*,*big*)) es la probabilidad de que las palabras *town* y *big* co-ocurran en un mismo párrafo dentro del conjunto de textos externos, en nuestro caso *Wikipedia* en inglés. El mismo procedimiento es realizado para las k ($k = 10$) palabras que forman un tópico. Este procedimiento es aplicado para cada tópico descubierto. Posteriormente se obtiene una *Coherencia Global Normalizada* de todos los tópicos con un promedio considerando la Coherencia del tópico normalizada de cada tópico, es decir se realiza la suma de las coherencias obtenidas por cada tópico. La cual proporciona valores del rango [-1,1].

Capítulo 5

Experimentación y resultados

En esta sección se presenta una experimentación la cual utilizará un corpus de 5,881,00 documentos. También se expone una evaluación de la integración de la clasificación de documentos dentro de la tarea de descubrimiento de tópicos.

Finalmente se expone una comparación de los resultados obtenidos con los reportados en el estado del arte.

5.1. Descripción de conjuntos de datos

En esta sección se presentan cuatro conjuntos de datos empleados en esta tesis doctoral. Los cuales se exponen en la Tabla 5.1 mostrando el número de documentos, clases y número de palabras. Los córpora empleados en la tarea de clasificación de documentos son *Reuters* (https://trec.nist.gov/data/reuters/reuters.html, consultado el 1 de mayo de 2020) y *20-Newsgroup* (http://qwone.com/~jason/20-Newsgroups/, consultado el 1 de mayo de 2020) que pertenecen al dominio de noticias.

Para la extracción de relaciones semánticas se utilizó la *Wikipedia* en inglés con 5,881,000 documentos, porque representa un vocabulario amplio y por lo tanto es una base de conocimiento general. Por lo que las relaciones semánticas resultantes son parte fundamental para construir los modelos de incrustación de relaciones semánticas; que a su vez se utilizan para representar los córpora empleados en la tarea de clasificacion de textos. El desempeño de los modelos de incrustación de relaciones semánticas propuestos se compara con el de los modelos de incrustación de palabras *glove* y *fastText* y el modelo de incrustación de relaciones semánticas *wnet2vec*. Los

cuales se exponen en la Tabla 5.2 mostrando por cada modelo el tamaño de los vectores, tipo y que datos los forman. Además, se muestran los modelos de incrustación de relaciones propuestos en este trabajo: sinonimia e hiponimia-hiperonimia; y una combinación de ambos. Como se observa, las relaciones semánticas que forman los tres modelos propuestos contienen menos información que la que se muestra en la Tabla 4.3. Esto se debió a que se utilizaron al máximo los recursos de cómputo disponibles en supercómputo BUAP, lo que permitió utilizar 40 mil pares de relaciones semánticas como número máximo de recursos.

Para la evaluación de la tarea de clasificación de documentos se aplicaron las métricas descritas en la sección 2.7. Por lo que en base a los resultados obtenidos en la evaluación de la tarea de clasificación de textos se evalúa el rendimiento de los modelos de incrustación de relaciones semánticas propuestos y el proceso de extracción de relaciones semánticas. Debido a que se integra la tarea de clasificación de textos con la tarea de descubrimiento de tópicos para obtener tópicos específicos se utilizan los mismos conjuntos de datos; esta última es evaluada empleando un corpus de referencia de un millón de documentos provenientes de *Wikipedia* en inglés.

Tabla 5.1: Descripción de conjuntos de datos

Corpus	**Documentos**	**Palabras**	**Clases**
20-Newsgroup	20,000	1,800,385	20
Reuters	18,456	3,435,808	90
Wikipedia para relaciones semánticas	5,881,000	3,380,578,354	0
Wikipedia para evaluación de los tópicos	1,000,000	1,560,478,211	0

Tabla 5.2: Modelos de incrustación [72]

Modelos de incrustación	**Datos**	**Tamaño vector**	**Tipo modelo**
GloVe	6 mil millones de tokens y cuentan con representaciones para 400 mil palabras	300	basado en palabras
fastText	1 millón de vectores de palabras y 16 mil millones de tokens	300	basado en palabras
WordNet	60 mil tokens	300	basado en relaciones de sinonimia, hiponimia-hiperonimia
Sinónimos	40 mil tokens	300	basado en relaciones de sinonimia
Hipónimos-hiperónimos	40 mil tokens	300	basado en relaciones de hiponimia-hiperonimia
Combinación	40 mil tokens	300	basado en relaciones de sinonimia, hiponimia-hiperonimia

5.2. Resultados experimentales

En esta sección se presentan los resultados obtenidos de aplicar los modelos de incrustación de relaciones semánticas en la tarea de clasificación de textos. Además, se exponen los resultados de la integración de la tarea de clasificación de documentos en la tarea de descubrimiento de tópicos. Con esto se afirma la hipótesis de este trabajo de tesis en donde se observa que al integrar la tarea de clasificación de textos dentro de la tarea de descubrimiento de tópicos se mejora la coherencia. Por último, se expone una comparativa de los resultados obtenidos en esta tesis doctoral con algunos de los resultados reportados en la literatura donde se observó que el método propuesto mejoró en algunos casos los resultados obtenidos utilizando los mismos

córpora.

5.2.1. Resultados de la tarea de clasificación de textos

En esta sección se presentan los resultados obtenidos de la evaluación de la tarea de clasificación de textos; la cual se evalúa con el objetivo de que las clases sean proporcionadas a la etapa de descubrimiento de tópicos con una buena calidad. Los resultados obtenidos de la evaluación de la tarea de clasificación de textos proporcionan una visión del desempeño de los tres modelos de incrustación de relaciones semánticas propuestos. Además de los resultados obtenidos aplicando los modelos *glove, fastText* y *wn2vec* lo que permite llevar a cabo una comparación del rendimiento de los 3 modelos de incrustación de relaciones propuestos con los 3 obtenidos de la literatura. Los cuales 2 son modelos de incrustación de palabras y uno de incrustación de relaciones semánticas.

La Tabla 5.3 muestra los resultados obtenidos al clasificar los corpus *20-Newsgroup* y *Reuters*. La métrica de precisión se identifica con la etiqueta P, la exactitud con R, la exhaustividad con A y la medida de F_1 con la etiqueta F_1. Los resultados obtenidos al aplicar el modelo de incrustación de relaciones basado en *WordNet (wn2vec)* permiten observar que no superan a los obtenidos con los modelos *glove* y *fastText*.

Además, los resultados obtenidos al clasificar el corpus *20-Newsgroup* superan a los obtenidos al aplicar el modelo de *WordNet* con el modelo que incorpora relaciones de hipónimo-hiperónimo con una precisión, exactitud y exhaustividad de 0.75,0.78 y 0.79 respectivamente.

Por otro lado para el corpus *Reuters* el modelo que incorporó la relación de sinonimia superó al modelo basado en *WordNet* con una exactitud y una exhaustividad de 0.74 y 0.84 respectivamente. Sin embargo, el modelo que incorporó los 3 tipos de relaciones semánticas fue capaz de superarlo con una precisión de 0.80, exactitud de 0.87, exhaustividad de 0.77 y una medida F_1 de 0.87.

Se estima que los resultados superaron a los obtenidos con el modelo basado en *WordNet* debido a que las relaciones incluidas en cada modelo propuesto fueron las más frecuentes en el total de relaciones obtenidas. En algunos casos, los modelos expuestos superaron a *glove* y *fastText*. Sin embargo, estos resultados aún son superficiales, por lo que se espera que incluir un mayor número de relaciones semánticas en cada modelo supere tanto el modelo expuesto por [109] como *glove* y *fastText*.

Tabla 5.3: Resultados obtenidos de la tarea de clasificación de documentos con la CNN y los modelos de incrustación de relaciones propuestos [72]

Conjunto de datos	*20-Newsgroup*				*Reuters*			
	P	**R**	**A**	$\boldsymbol{F_1}$	**P**	**R**	**A**	$\boldsymbol{F_1}$
fastText	0.76	0.74	0.75	0.75	0.72	0.71	0.71	0.71
GloVe	0.79	0.79	0.79	0.79	0.72	0.66	0.66	0.67
WordNet	0.66	0.64	0.64	0.64	0.71	0.68	0.68	0.68
Hipónimos-hiperónimos	**0.75**	**0.78**	**0.79**	0.66	0.72	0.67	0.67	0.68
Sinónimos	0.66	0.64	0.64	0.64	0.70	**0.74**	**0.84**	**0.70**
Combinación	0.67	0.59	0.59	0.60	**0.80**	**0.87**	**0.77**	**0.87**

5.2.2. Resultados del descubrimiento de tópicos

En esta sección se presenta la integración de la clasificación de documentos en el proceso de descubrimiento de tópicos para obtener tópicos específicos para cada clase.

La integración propuesta proporciona tópicos latentes y específicos representados por palabras representativas con alta coherencia de cada clase obtenida.

Partiendo de la tarea de clasificación de documentos, los resultados son integrados en la tarea de descubrimiento de tópicos. Es decir, los resultados obtenidos en la tarea de clasificación de documentos son la entrada en un modelo de descubrimiento de tópicos con el propósito de descubrir tópicos específicos de acuerdo con la clase a la que pertenezcan.

Los modelos LDA, LSA y $PLSA$ fueron aplicados en la fase del descubrimiento de tópicos sobre el corpus de clases. Un total de 20, 50 y 100 tópicos fueron descubiertos para los corpus *20-Newsgroup* y *Reuters*. En ambos casos se obtienen 10 palabras representativas en cada tópico descubierto.

De los resultados obtenidos se extrajeron la media y las desviaciones estándar. La media con el objetivo de identificar tendencias en los corpus *20-Newsgroups* y *Reuters*. Por lo que una desviación estándar baja indica que las palabras que forman los tópicos encontrados se encuentran con una fuerte relación semántica entre ellos. De esta manera, fue posible analizar los resultados de cada corpus.

El número de tópicos (n), el promedio de la coherencia del tópico normalizada (media (Avg) y la desviación estándar (std)) de los tópicos descubiertos de las clases de los córpora *20-Newsgroup* y *Reuters* con los modelos *LDA, LSA* y *PLSA* se

exponen en las Tablas 5.4, 5.5 y 5.6 respectivamente.

Tabla 5.4: Promedio de la coherencia de tópico normalizada obtenida con el modelo *LDA* con 20, 50 y 100 tópicos descubiertos para el corpus *20-Newsgroup* y *Reuters* [70]

n	***20-Newsgroup***		***Reuters***	
	Avg	***std***	***Avg***	***std***
LDA_20	**0.1723**	0.0104	**0.1441**	0.0472
LDA_50	0.1572	0.0116	0.1394	0.0165
LDA_100	0.1453	0.0097	0.1370	0.0192

Tabla 5.5: Promedio de la coherencia de tópico normalizada obtenida con el modelo *LSA* con 20, 50 y 100 tópicos descubiertos para el corpus *20-Newsgroup* y *Reuters* [70]

n	***20-Newsgroup***		***Reuters***	
	Avg	***std***	***Avg***	***std***
LSA_20	**0.1622**	0.0158	**0.1360**	0.0176
LSA_50	0.1556	0.0095	0.1342	0.0170
LSA_100	0.1462	0.0098	0.139	0.0487

Tabla 5.6: Promedio de la coherencia normalizada del tópico con el modelo *PLSA* con 20, 50 y 100 tópicos para el corpus *20-Newsgroup* y *Reuters* [70]

n	***20-Newsgroup***		***Reuters***	
	Avg	***std***	***Avg***	***std***
PLSA_20	**0.1716**	0.0099	**0.1436**	0.0160
PLSA_50	0.1559	0.0095	0.1409	0.0531
PLSA_100	0.1457	0.0095	0.1457	0.01480

Como se mencionó anteriormente el descubrimiento de tópicos con los modelos LDA, LSA y $PLSA$ reportó los mejores resultados al descubrir 20 tópicos con 10 palabras representativas cada uno, por lo que las Tablas 5.7 y 5.8 exponen a manera de ejemplo las palabras representativas de solo dos tópicos descubiertos. Para cada uno de estos dos tópicos se exponen solo 3 palabras representativas de 10 que forman a cada tópico de esos dos tópicos ejemplo.

Los resultados mostraron que los modelos de incrustación de relaciones semánticas propuestos mejoran a los resultados obtenidos con el modelo de incrustación de relaciones extraídas de *WordNet* [109].

Tabla 5.7: Ejemplo de palabras representativas del corpus *20-Newsgroup* con *LDA, LSA* y *PLSA* con 20 tópicos [70]

	Tópicos					
	LDA		**LSA**		**PLSA**	
Clase	**Season**	**Games**	**Season**	**Games**	**Season**	**Games**
Futbol	*...game, player, penalties...*	*...team, baseball, league...*	*...hockey, player, fans...*	*...canada, league, rangers,*	*...players, games, sports...,*	*...baseball, ...season, team...,*
	Hardware	**Software**	**Hardware**	**Software**	**Hardware**	**Software**
Computer	*...machine, mouse, floopy...*	*...,email octave, driver...*	*...desktop, problem, department...*	*...norton, ftp, zip...*	*...mouse, machine, desktop...,*	*...windows, microsoft, norton...,*
	Religion	**Rituals**	**Religion**	**Rituals**	**Religion**	**Rituals**
Atheism	*...god, jewish, israel...*	*...love, beliefs, fatwa...*	*...hope, sabbath, day...*	*...ceremonial, law, god...*	*...bible, jewish, love...*	*...god, jewish, love...*

Los resultados obtenidos brindaron una visión sobre la integración de la tarea de clasificación de documentos en la tarea de descubrimiento de tópicos. En los modelos *LDA, LSA* y *PLSA* los datos de entrada son estrictamente un corpus de clases previamente extraído con una red neuronal convolucional y un modelo de incrustación de relaciones semánticas. Para cada modelo de descubrimiento de tópicos el corpus de clases aporta la presencia de información más específica.

El modelo que obtuvo un promedio de la coherencia del tópico normalizada alta para el corpus *20-Newsgroup* fue al descubrir 20 tópicos con 10 palabras representativas con el modelo *LDA* con 0.1723. Para el corpus *Reuters* el resultado más alto fue al descubrir 20 tópicos con 10 palabras representativas con el modelo *LDA* con un promedio de la coherencia del tópico normalizada de 0.1441.

Los resultados obtenidos en el corpus *20-Newsgroup* permitieron observar que el descubrimiento de 20 tópicos fue suficiente y coherente ya que se está trabajando con 20 clases de las cuales se extraen 20 tópicos en cada clase. Sin embargo, para el corpus *Reuters*, los resultados obtenidos mostraron que descubrir 20 tópicos fue suficiente y coherente ya que descubrir un número mayor de tópicos no aporta un aumento en la coherencia del tópico de los resultados. Además, se obtienen 90 clases diferentes y la extracción de 20 tópicos por clase proporciona resultados coherentes.

Tabla 5.8: Ejemplo de palabras representativas del corpus *Reuters* con *LDA, LSA* y *PLSA* con 20 tópicos [70]

	Tópicos					
	LDA		**LSA**		**PLSA**	
Clase	**Grain**	**Corn**	**Grain**	**Corn**	**Grain**	**Corn**
Oil	*...nil, dlrs, freight...*	*...potato, mln, soymeal...*	*...protein, vegetable, soybean...*	*...labour, italy, agricultural,*	*...godless, protein, rise...,*	*...tonnes, ...oilcake, meals...,*
	Consumer	**Sales**	**Consumer**	**Sales**	**Consumer**	**Sales**
Cocoa	*...malaysia, market, international...*	*...tonnes, crop, coffee...*	*...kilo, france, european...*	*...bean, tonnes, flower...*	*...malaysia, agriculture, shipment...,*	*...flower, mln, location...,*
	Elements	**Others**	**Elements**	**Others**	**Elements**	**Other**
Sugar	*...price, farm, zinc...*	*...cottom, floor, crop...*	*...quarter, zinc, cotton...*	*...agriculture, cocoa, crop...*	*...cost, floor, rice...*	*...farm, crop, agriculture...*

5.2.3. Comparación de resultados

En esta sección se expone una comparación de los resultados obtenidos en este trabajo de tesis doctoral y los resultados reportados por algunos autores en la literatura.

Algunos autores exponen diferentes métodos que son capaces de descubrir tópicos con diferentes metodologías y los conjuntos de datos empleados son en su mayoría el corpus *20-Newsgroup* y en algunos casos *Reuters*. Los autores utilizaron la métrica de evaluación de coherencia de tópico normalizada para evaluar el rendimiento de sus métodos.

La Tabla 5.9 muestra los resultados de los autores Fuentes-Pineda y Meza-Ruiz [33], Wang et al. [143], Xu et al. [145], Ding et al. [27], Bianchi et al. [11] y Jin et al.[55] y los obtenidos en este trabajo de tesis doctoral.

Los resultados obtenidos en esta investigación son superiores a los resultados obtenidos por los autores anteriormente mencionados. Por otro lado, en Jin et al.[55] obtienen mayores valores de coherencia para el corpus *20-Newsgroup*.

Los autores Wang et al. [142], y Austin et al. [8] obtienen valores de coherencia superiores a los resultados obtenidos en este trabajo para el corpus *Reuters*. En Jin et al. [55] y Wang et al. [142], los resultados son significativos a los obtenidos en este trabajo, ya que aplican algoritmos y métodos como codificador automático variacional, detección y minería comunitaria. Lo que permite concluir que los métodos

mencionados anteriormente benefician los resultados de los autores.

Tabla 5.9: Comparación de los resultados obtenidos con coherencia del tópico normalizada para ambos corpus [70]

Autor	**20-Newsgroup**	**Reuters**
Fuentes-Pinea et al. [33]	0.100	0.040
Wang et al. [143]	0.103	0.152
Xu et al. [145]	0.390	-
Ding et al. [27]	0.280	-
Bianchi et al. [11]	0.102	-
Jin et al. [55]	0.042	-
Jin et al. [55]	0.279	-
Austin et al. [8]	0.044	0.182
Wang et al. [142]	0.170	0.180
Este trabajo	0.172	0.147

Capítulo 6

Conclusiones

En esta sección se exponen las conclusiones obtenidas con la metodología que se encarga de descubrir tópicos específicos con el método de descubrimiento de tópicos propuesto en esta tesis doctoral. La cual consiste en aplicar técnicas de Procesamiento de Lenguaje Natural para generar córporas homogéneos, extracción de relaciones semánticas para crear el modelo de incrustación que a su vez será utilizado en la clasificación de documentos aplicando un modelo de aprendizaje profundo. Posteriormente la tarea de clasificación de textos es integrada en la tarea de descubrimiento de tópicos.

Los resultados obtenidos han permitido llegar a la conclusión de que los tres modelos de incrustación de relaciones semánticas propuestos mostraron la importancia de que las relaciones semánticas proporcionan coherencia al texto que forma una oración, lo cual es útil para las tareas de clasificación de textos al enriquecer los vectores de los documentos. El método propuesto puede ser útil para los analistas de datos porque los modelos de incrustación de relaciones semánticas continúan siendo una herramienta que mejora los resultados para tareas automáticas que involucran el tratamiento de información textual. Los resultados obtenidos son variables debido a que cada modelo de incrustación propuesto tiene información semántica diferente. El comportamiento de cada modelo presentado fue evaluado a través de la clasificación de textos. Además, se comparó su desempeño con los resultados obtenidos al evaluar los modelos de la literatura: *fasText, GloVe* y el modelo basado en las relaciones semánticas de *WordNet*. Los resultados mostraron que los modelos de incrustación de relaciones semánticas propuestas superan al modelo propuesto en [109]. El resultado de clasificar los córpora *20-Newsgroup* y *Reuters* fueron dos corpus de clases.

Los cuáles serán los datos de entrada en la tarea de descubrimiento de tópicos. Los resultados se comparan con la literatura, demostrando que el descubrimiento de los tópicos en el corpus de clases genera resultados con una relación más significativa entre los textos que lo forman. Los tópicos descubiertos se evaluaron con la métrica de coherencia de tópico normalizada (*NPMI*). Los resultados mostraron que la integración de la clasificación de textos proporcionó en su mayoría textos relacionados en cada clase recuperada; por lo tanto, los tópicos con las palabras principales estaban más relacionados.

Las principales contribuciones de la metodología propuesta en esta tesis doctoral son:

- La extracción de relaciones semánticas de sinonimia, hiponimia e hiperonimia de *Wikipedia* en inglés utilizando patrones léxico sintácticos. Las cuáles serán la base de conocimiento para el desarrollo de los modelos de incrustación de relaciones semánticas.

- Un enfoque para el desarrollo de un modelo de incrustación de relaciones semánticas basado en relaciones de sinonimia, hiponimia e hiperonimia previamente extraídas y validadas en la tarea de clasificación de textos.

- El modelo de incrustación generado a partir de relaciones semánticas de sinonimia, hiponimia e hiperonimia como vectores de baja dimensión se utilizaron para representar los córpora en una red neuronal convolucional que realiza la tarea de clasificación de documentos.

- Una comparación del rendimiento de los modelos de incrustación de relaciones semánticas con los modelos existentes en la literatura y el propuesto por Saedi et al. [109].

- Los modelos de incrustación de relaciones semánticas previamente propuestos se utilizaron para representar los conjuntos de datos, para integrar la tarea de clasificación de textos en el descubrimiento de tópicos.

- Una comparación del desempeño de la integración de la clasificación de documentos con el descubrimiento de tópicos con los existentes en la literatura.

- Los resultados expresan la importancia de contar con un texto clasificado para descubrir tópicos específicos.

– La metodología propuesta se convierte en un recurso útil en el procesamiento del lenguaje natural, demostrando que agregar semántica a un proceso de clasificación de textos traerá resultados positivos.

Los mejores resultados obtenidos en la clasificación del corpus *20-Newsgroup* se obtuvieron con el modelo de incrustación de hiponimia-hiperonimia, logrando una precisión de 0.79. Para el corpus de *Reuters*, se obtuvo una medida- F_1 y un recall de 0.87 utilizando el modelo de incrustación que involucra las relaciones de sinonimia, hiponimia e hiperonimia.
Por lo que el método para el descubrimiento de tópicos con aprendizaje profundo obtuvo una coherencia del tópico de 0.172 para el corpus *20-Newsgroup* con el modelo LDA con 20 tópicos y 0.147 para el corpus *Reuters* con el modelo LSA con 20 tópicos, en ambos casos con 10 palabras representativas.

Los trabajos a futuro que se proponen como líneas de investigación en el descubrimiento de tópicos son: Etiquetar los tópicos descubiertos con la ayuda de una ontología del mismo dominio y posteriormente comparar los resultados obtenidos en experimentos previos. Además, proponer un nuevo método para la extracción de relaciones semánticas como relaciones de tipo parte-todo o roles causales y semánticos. Aplicar la metodología presentada en esta tesis doctoral, pero con conjuntos de datos en español. Realizar pruebas con diferentes arquitecturas neuronales y comparar el rendimiento de cada nuevo modelo de clasificación con aprendizaje profundo aplicado con el utilizado en este trabajo.

Apéndice A

Publicaciones

En éste anexo se presentan los artículos publicados de esta tesis doctoral. Los artículos se encuentran clasificados en publicaciones JCR, capítulos de libro y revistas indizadas.

Publicaciones JCR:

1. Lezama-Sánchez, A.L.; Tovar Vidal, M.; Reyes-Ortiz, J.A. An Approach Based on Semantic Relationship Embeddings for Text Classification. Mathematics 2022, 10, 4161. https://doi.org/10.3390/math10214161 Q1 [72].

2. Lezama-Sánchez, A.L.; Tovar Vidal, M.; Reyes-Ortiz, J.A. Integrating Text Classification into Topic Discovery Using Semantic Embedding Models. Appl. Sci. 2023, 13, 9857. https://doi.org/10.3390/app13179857 Q2 [70].

Revistas indexadas (SCOPUS)

1. Lezama Sánchez, A. L., Tovar Vidal, M., y Reyes-Ortiz, J. A. (2022, October). Topic Discovery About Economy During COVID-19 Pandemic from Spanish Tweets. Vol 3 (pp. 521-533). Lecture Notes in Networks and Systems, vol 561. Springer, Cham. SCORPUS Q2 [71].

2. Lezama Sánchez, A. L., Tovar Vidal, M., y Reyes Ortiz, J. A. (2022). A Behavior Analysis of the Impact of Semantic Relationships on Topic Discovery. Computación y Sistemas, Indexada por CONACHYT 26(1), 149-160 [67].

3. Lezama Sánchez, A. L., Tovar Vidal, M., y Reyes-Ortiz, J. A. (2021, October). Hypernyms-based topic discovery using LDA. Lecture Notes in Computer Science(), vol 13068. Springer, Cham [66].

Revistas

1. Lezama Sánchez. A. L., Tovar Vidal, M., y Reyes-Ortiz, J. A. Enfoque para el descubrimiento de tópicos a través de WordNet. V. M. Orizaba, editor, Coloquio de Investigación Multidisciplinaria 2021, 2021 [64].

Capítulos de libro:

1. Lezama Sánchez, A. L., Tovar Vidal, M.; Reyes-Ortiz, J. A. (2021). Estado del Arte de Métodos de descubrimiento de tópicos. En J. M. González Calleros, J. Guerrero García, C. Zepeda Cortés, D. Villarino Ayala. (eds) Avances de ingeniería del lenguaje, del conocimiento y la interacción humano máquina. Volumen 1, pp 89-97 [65].

2. Lezama Sánchez, A. L., Tovar Vidal, M.; Reyes-Ortiz, J. A. (2022). Descubrimiento de tópicos con aprendizaje profundo: una revisión preliminar sistemática del estado del arte. En M. Tovar Vidal, C. Zepeda Cortés, D. Villarino Ayala, J. M. González Calleros y J. Guerrero García. (eds) Lenguaje, conocimiento y tecnología educativa: nuevos enfoques de aplicación, pp. 54-62. Publicaciones BUAP, ISBN: 978-607-525-846-1 [72].

3. Lezama Sánchez, A. L., Tovar Vidal, M.; Reyes-Ortiz, J. A. (2022). Una aproximación basada en n-gramas para la detección de palabras clave y relaciones semánticas. En M. Tovar Vidal, G. de Ita Luna, P. Bello López, M. Contreras González, F. Zacarias Flores, Y. Moyao Martínez y L. C. Altamirano Robles. (eds) Procesamiento de lenguaje natural y métodos basados en grafos, pp. 1-10. Publicaciones BUAP, ISBN BUAP: 978-607-525-845-4 [68].

4. Lezama-Sánchez, A. L., Tovar Vidal, M., y Reyes-Ortiz, J. A. (2023). Descubrimiento de tópicos para la detección de depresión utilizando una red neuronal convolucional. En M. Tovar Vidal, G. de Ita Luna, P. Bello (eds) Aplicaciones en procesamiento de lenguaje natural y teoría de grafos, pp. 35-48. Publicaciones BUAP, ISBN: 978-607-8957-18-7 [69].

Bibliografía

[1] C. C. Aggarwal et al. Neural networks and deep learning. *Springer*, 10:978–3, 2018.

[2] R. A. Al-Zaidy and C. L. Giles. Extracting semantic relations for scholarly knowledge base construction. In *2018 IEEE 12th international conference on semantic computing (ICSC)*, pages 56–63. IEEE, 2018.

[3] M. G. Al Zamil and A. B. Can. Rolex-sp: Rules of lexical syntactic patterns for free text categorization. *Knowledge-Based Systems*, 24(1):58–65, 2011.

[4] R. Alfaro-Flores. Evaluación del efecto en el algoritmo de análisis semántico latente al utilizar colecciones de datos cada vez más grandes para la detección y extracción de sinónimos y su independencia respecto al lenguaje, por medio de su implementación distribuida. Tesis de maestría, Instituto Tecnológico de Costa Rica, 2014.

[5] F. Almeida and G. Xexéo. Word embeddings: A survey. *arXiv preprint arXiv:1901.09069*, 2019.

[6] J. A. B. Andrades. Estado del arte en probabilistic latent semantic analysis aplicado a problemas de acceso a la información en la web. 2011.

[7] K. Arutchelvan and R. S. Selvan. Tmrsg: Topic model based rich semantic graph method for abstractive multi-document summarization. *Journal of Theoretical and Applied Information Technology*, 100(12), 2022.

[8] E. Austin, A. Trabelsi, C. Largeron, and O. R. Zaïane. Hierarchical topic model inference by community discovery on word co-occurrence networks. In *Data Mining: 20th Australasian Conference, AusDM 2022, Western Sydney, Australia, December 12–15, 2022, Proceedings*, pages 148–162. Springer, 2022.

[9] M. Beguerisse-Díaz, A. K. McLennan, G. Garduño Hernández, M. Barahona, and S. J. Ulijaszek. The who and what of #diabetes on twitter. *Digital health*, 3:1–29, 2017.

[10] R. Bentrcia, S. Zidat, and F. Marir. Extracting semantic relations from the quranic arabic based on arabic conjunctive patterns. *Journal of King Saud University-Computer and Information Sciences*, 30(3):382–390, 2018.

[11] F. Bianchi, S. Terragni, and D. Hovy. Pre-training is a hot topic: Contextualized document embeddings improve topic coherence. *arXiv preprint arXiv:2004.03974*, 2020.

[12] Y. Bougteb, B. Ouhbi, B. Frikh, et al. Deep learning based topics detection. In *2019 Third International Conference on Intelligent Computing in Data Sciences (ICDS)*, pages 1–7. IEEE, 2019.

[13] G. Bouma. Normalized (pointwise) mutual information in collocation extraction. *Proceedings of GSCL*, 30:31–40, 2009.

[14] J. Brownlee. *Deep learning with Python: develop deep learning models on Theano and TensorFlow using Keras*. Machine Learning Mastery, 2016.

[15] D. Buenaño Fernández, M. González, D. Gil, and S. Luján-Mora. Text mining of open-ended questions in self-assessment of university teachers: An LDA topic modeling approach. *IEEE Access*, 8:35318–35330, 2020.

[16] Y. Chai and W. Li. Towards deep learning interpretability: A topic modeling approach. *ICIS*, 26:1–10, 2019.

[17] C.-H. Chang and S.-Y. Hwang. A word embedding-based approach to cross-lingual topic modeling. *Knowledge and Information Systems*, 63(6):1529–1555, 2021.

[18] H. R. L. Chavez and V. M. Tovar. Proposal for automatic extraction of taxonomic relations in domain corpus. *Res. Comput. Sci.*, 133:29–39, 2017.

[19] K. W. Church. Word2vec. *Natural Language Engineering*, 23(1):155–162, 2017.

[20] A. Cifuentes, E. Mendoza, M. Lizcano, A. Santrich, and S. Moreno-Trillos. Desarrollo de una red neuronal convolucional para reconocer patrones en imágenes. *Investigación y desarrollo en TIC*, 10(2):7–17, 2019.

[21] Y. Cong, B. Chen, H. Liu, and M. Zhou. Deep latent dirichlet allocation with topic-layer-adaptive stochastic gradient riemannian mcmc. In *International Conference on Machine Learning*, pages 864–873. PMLR, 2017.

[22] G. Costa and R. Ortale. Document clustering meets topic modeling with word embeddings. In *Proceedings of the 2020 SIAM International Conference on Data Mining*, pages 244–252. SIAM, 2020.

[23] H. Costa. *Automatic Extraction and Validation of Lexical Ontologies from text*. PhD thesis, Master's thesis, University of Coimbra, Faculty of Sciences and Technology, 2010.

[24] M. del Carmen Ugalde. El lenguaje caracterización de sus formas fundamentales. *Letras*, (20-21):15–34, 1989.

[25] T. G. Dietterich. Machine-learning research. *AI magazine*, 18(4):97–97, 1997.

[26] C. Dima and E. Hinrichs. Automatic noun compound interpretation using deep neural networks and word embeddings. In *Proceedings of the 11th International Conference on Computational Semantics*, pages 173–183, 2015.

[27] R. Ding, R. Nallapati, and B. Xiang. Coherence-aware neural topic modeling. *arXiv preprint arXiv:1809.02687*, 2018.

[28] W. Dubitzky, O. Wolkenhauer, K.-H. Cho, and H. Yokota. *Encyclopedia of systems biology*, volume 402. Springer New York, NY, USA:, 2013.

[29] Á. Escobar. Hacia una definición lingüística del tópico literario. *Myrtia*, 15:123–160, 2000.

[30] B. Eslami, Z. Rezaei, M. Habibzadeh, M. Fouladian, and H. Ebrahimpour-Komleh. Using deep learning methods for discovering associations between drugs and side effects based on topic modeling in social network. *Social Network Analysis and Mining*, 10:1–17, 2020.

[31] M. A. et. al. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[32] J. D. G. Fierros. Tesis de maestría en ciencias. 2012.

[33] G. Fuentes-Pineda and I. V. Meza-Ruiz. Topic discovery in massive text corpora based on min-hashing. *Expert Systems with Applications*, 136:62–72, 2019.

[34] W. Gao, M. Peng, H. Wang, Y. Zhang, W. Han, G. Hu, and Q. Xie. Generation of topic evolution graphs from short text streams. *Neurocomputing*, 383:282–294, 2020.

[35] W. Gao, M. Peng, H. Wang, Y. Zhang, Q. Xie, and G. Tian. Incorporating word embeddings into topic modeling of short text. *Knowledge and Information Systems*, 61(2):1123–1145, 2019.

[36] J. A. García-Díaz, O. Apolinario-Arzube, and R. Valencia-García. Evaluating pre-trained word embeddings and neural network architectures for sentiment analysis in spanish financial tweets. In *Mexican International Conference on Artificial Intelligence*, pages 167–178. Springer, 2020.

[37] S. Ghannay, B. Favre, Y. Esteve, and N. Camelin. Word embedding evaluation and combination. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 300–305, 2016.

[38] A. Ghenai and Y. Mejova. Catching zika fever: Application of crowdsourcing and machine learning for tracking health misinformation on twitter. *arXiv preprint arXiv:1707.03778*, 1:518–528, 2017.

[39] I. E. E. M. González. Modelos de aprendizaje automático para el apoyo en la clasificación de tipos de cáncer a partir de datos estructurados y no estructurados de expedientes clínicos. Tesis de maestría, Universidad Autónoma Metropolitana Unidad Azcapotzalco División de Ciencias Básicas e Ingeniería, 2020.

[40] G. Grigonyte. *Building and evaluating domain ontologies: NLP contributions*. Logos Verlag Berlin GmbH, 2010.

[41] L. Guo, Z. Li, T. Yang, H. Zhang, D. Mu, and Y. Li. An improved latent dirichlet allocation method for service topic detection. In *2016 35th Chinese control conference (CCC)*, pages 7045–7049. IEEE, 2016.

[42] A. Gupta and R. Katarya. Improving document representation using kpca and clustered word embeddings. In *2021 5th International Conference on Electrical, Electronics, Communication, Computer Technologies and Optimization Techniques (ICEECCOT)*, pages 514–517. IEEE, 2021.

[43] R. M. Gutiérrez. Análisis semántico latente:¿ teoría psicológica del significado? *Revista signos*, 38(59):303–323, 2005.

[44] G. He, Y. Liang, Y. Chen, W. Yang, J. S. Liu, M. Q. Yang, and R. Guan. A hotspots analysis-relation discovery representation model for revealing diabetes mellitus and obesity. *BMC systems biology*, 12(7):116, 2018.

[45] M. A. Hearst. Automatic acquisition of hyponyms from large text corpora. In *Coling 1992 volume 2: The 15th international conference on computational linguistics*, 1992.

[46] M. A. Hearst. Automatic acquisition of hyponyms from large text corpora. In *COLING 1992 Volume 2: The 14th International Conference on Computational Linguistics*, pages 539–545, 1992.

[47] R. Hecht-Nielsen. Theory of the backpropagation neural network. In *Neural networks for perception*, pages 65–93. Elsevier, 1992.

[48] K. M. Hermann, T. Kocisky, E. Grefenstette, L. Espeholt, W. Kay, M. Suleyman, and P. Blunsom. Teaching machines to read and comprehend. *Advances in neural information processing systems*, 28, 2015.

[49] D. Hosmer, S. Lemeshow, and R. Sturdivant. Area under the receiver operating characteristic curve. *Applied Logistic Regression. Third ed: Wiley*, pages 173–182, 2013.

[50] F. Hu, Z. Shao, and T. Ruan. Self-supervised synonym extraction from the web. *J. Inf. Sci. Eng.*, 31(3):1133–1148, 2015.

[51] R. Hu, J. Liu, and Y. Wen. SP-BTM: A specific part-of-speech btm for service clustering. In *2020 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking (ISPA/BDCloud/SocialCom/SustainCom)*, pages 1050–1057. IEEE, 2020.

[52] H. Jelodar, Y. Wang, R. Orji, and H. Huang. Deep sentiment classification and topic discovery on novel coronavirus or covid-19 online discussions: Nlp using lstm recurrent neural network approach. *arXiv preprint arXiv:2004.11695*, pages 1–23, 2020.

[53] H. Jiang, Z. Lei, Y. Rao, H. Xie, and F. L. Wang. Parallel dynamic topic modeling via evolving topic adjustment and term weighting scheme. *Information Sciences*, 585:176–193, 2022.

[54] M. Jin, X. Luo, H. Zhu, and H. H. Zhuo. Combining deep learning and topic modeling for review understanding in context-aware recommendation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1605–1614, 2018.

[55] Y. Jin, H. Zhao, M. Liu, L. Du, and W. Buntine. Neural attention-aware hierarchical topic model. *arXiv preprint arXiv:2110.07161*, 2021.

[56] A. Joulin, E. Grave, P. Bojanowski, M. Douze, H. Jégou, and T. Mikolov. Fasttext. zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*, 2016.

[57] A. Karami, A. A. Dahl, G. Turner-McGrievy, H. Kharrazi, and G. Shaw Jr. Characterizing diabetes, diet, exercise, and obesity comments on twitter. *International Journal of Information Management*, 38(1):1–6, 2018.

[58] S. Karamizadeh, S. M. Abdullah, A. A. Manaf, M. Zamani, and A. Hooman. An overview of principal component analysis. *Journal of Signal and Information Processing*, 4(3B):173, 2013.

[59] N. Kawamae. Topic structure-aware neural language model: Unified language model that maintains word and topic ordering by their embedded representations. In *The World Wide Web Conference*, pages 2900–2906, 2019.

[60] N. Kawamae. Topic aware neural language model: Domain adaptation of unconditional text generation models. 2021.

[61] S. A. Kinariwala and S. Deshmukh. Onto_tml: Auto-labeling of topic models. *Journal of Integrated Science and Technology*, 9(2):85–91, 2021.

[62] J. Y. Lee, F. Dernoncourt, and P. Szolovits. Mit at semeval-2017 task 10: Relation extraction with convolutional neural networks. *arXiv preprint arXiv:1704.01523*, 2017.

[63] P. León-Araúz, A. San Martín, and P. Faber. Pattern-based word sketches for the extraction of semantic relations. In *Proceedings of the 5th international workshop on computational terminology (Computerm2016)*, pages 73–82, 2016.

[64] A. L. Lezama Sánchez, M. Tovar Vidal, and J. A. Reyes Ortiz. Enfoque para el descubrimiento de tópicos a través de Wordnet. In V. M. Orizaba, editor, *Coloquio de Investigación Multidisciplinaria 2021*, 2021.

[65] A. L. Lezama Sánchez, M. Tovar Vidal, and J. A. Reyes Ortiz. Estado del arte de métodos de descubrimiento de tópicos. In d. c. y. l. i. h. m. Avances de ingeniería del lenguaje, editor, *Procesamiento de lenguaje natural y métodos basados en grafos*, volume 1, pages 89–97, 2021.

[66] A. L. Lezama Sánchez, M. Tovar Vidal, and J. A. Reyes Ortiz. Hypernyms-based topic discovery using LDA. In I. Z. Batyrshin, A. F. Gelbukh, and G. Sidorov, editors, *Advances in Soft Computing - 20th Mexican International Conference on Artificial Intelligence, MICAI 2021, Mexico City, Mexico, October 25-30, 2021, Proceedings, Part II*, volume 13068 of *Lecture Notes in Computer Science*, pages 70–80. Springer, 2021.

[67] A. L. Lezama Sánchez, M. Tovar Vidal, and J. A. Reyes Ortiz. A behavior analysis of the impact of semantic relationships on topic discovery. *Computación y Sistemas*, 26(1):149–160, 2022.

[68] A. L. Lezama Sánchez, M. Tovar Vidal, and J. A. Reyes Ortiz. Una aproximación basada en n-gramas para la detección de palabras clave y relaciones semánticas. In E. BUAP, editor, *Procesamiento de lenguaje natural y métodos basados en grafos*, 2022.

[69] A. L. Lezama Sánchez, M. Tovar Vidal, and J. A. Reyes Ortiz. Descubrimiento de tópicos para la detección de depresión utilizando una red neuronal convolucional. In E. BUAP, editor, *Editorial BUAP*, 2023.

[70] A. L. Lezama-Sánchez, M. Tovar Vidal, and J. A. Reyes-Ortiz. Integrating text classification into topic discovery using semantic embedding models. *Applied Sciences*, 13(17):9857, 2023.

[71] A. L. Lezama Sánchez, M. Tovar Vidal, and J. A. Reyes-Ortiz. Topic discovery about economy during covid-19 pandemic from spanish tweets. In *Proceedings of the Future Technologies Conference*, pages 521–533. Springer, 2023.

[72] A. L. Lezama-Sánchez, M. Tovar Vidal, and J. A. Reyes-Ortiz. An approach based on semantic relationship embeddings for text classification. *Mathematics*, 10(21), 2022.

[73] C. Li, Y. Duan, H. Wang, Z. Zhang, A. Sun, and Z. Ma. Enhancing topic modeling for short texts with auxiliary word embeddings. *ACM Transactions on Information Systems (TOIS)*, 36(2):1–30, 2017.

[74] W. Li, H. Saigo, B. Tong, and E. Suzuki. Topic modeling for sequential documents based on hybrid inter-document topic dependency. *Journal of Intelligent Information Systems*, 56(3):435–458, 2021.

[75] C.-Y. Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

[76] J. Liu, Z. He, and Y. Huang. Hashtag2vec: Learning hashtag representation with relational hierarchical embedding model. In *IJCAI*, pages 3456–3462, 2018.

[77] S. Liu, F. Shen, V. Chaudhary, and H. Liu. Mayonlp at semeval 2017 task 10: Word embedding distance pattern for keyphrase classification in scientific

publications. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 956–960, 2017.

[78] R. R. López Barbosa. Análisis de sentimientos en textos de opinión: una evaluación práctica. *Plaza y Valdés, S.A. de C.V.*, 2019.

[79] Z. Majdabadi, B. Sabeti, P. Golazizian, S. A. A. Asli, O. Momenzadeh, et al. Twitter trend extraction: A graph-based approach for tweet and hashtag ranking, utilizing no-hashtag tweets. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 6213–6219, 2020.

[80] P. G. Martín, A. D. de Pascual, E. T. Lezama, and E. G. Olmos. Una aplicación del análisis de componentes principales en el área educativa. *Economía*, 19(9):55–72, 1994.

[81] L. McInnes, J. Healy, and J. Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[82] R. M. O. Mendoza. Descubrimiento automático de hipónimos a partir de texto no estructurado. *Tesis de lic. Mexico: Instituto Nacional de Astrofisica, Optica y Electronica*, 2007.

[83] G. A. Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

[84] D. Mimno, H. Wallach, E. Talley, M. Leenders, and A. McCallum. Optimizing semantic coherence in topic models. In *Proceedings of the 2011 conference on empirical methods in natural language processing*, pages 262–272, 2011.

[85] V. B. Mititelu. Hyponymy patterns in romanian. *Memoirs of the Scientific Sections of the Romanian Academy*, 34:31–40, 2011.

[86] O. P. Montoya. Redes neuronales convolucionales profundas para el reconocimiento de emociones en imágenes. Tesis de maestría, Escuela técnica supeior de ingenieros informáticos, 2018.

[87] C. E. Moody. Mixing dirichlet topic models and word embeddings to make lda2vec. *arXiv preprint arXiv:1605.02019*, 2016.

[88] D. Moreira, I. Cruz, K. Gonzalez, A. Quirumbay, C. Magallan, T. Guarda, A. Andrade, and C. Castillo. Análisis del estado actual de procesamiento de lenguaje natural. *Revista Ibérica de Sistemas e Tecnologias de Informação*, (E42):126–136, 2021.

[89] B. Navarro-Colorado. On poetic topic modeling: extracting themes and motifs from a corpus of spanish poetry. *Frontiers in Digital Humanities*, 5:15, 2018.

[90] D. Newman, J. H. Lau, K. Grieser, and T. Baldwin. Automatic evaluation of topic coherence. In *Human language technologies: The 2010 annual conference of the North American chapter of the association for computational linguistics*, pages 100–108, 2010.

[91] M. Newman. Networks: An introduction. 2010: Oxford university press. *Artif. Life*, 18:241–242, 2012.

[92] M. A. Nielsen. *Neural networks and deep learning*, volume 2018. Determination press San Francisco, CA, 2015.

[93] A. Núnez-Reyes, E. M. Cuevas, E. Villatoro-Tello, G. Ramírez-de-la Rosa, and C. Sánchez-Sánchez. Agrupamiento de textos cortos en dominios cruzados. *Research in Computing Science*, 115:133–145, 2016.

[94] M. d. C. Olvera Porcel et al. Coeficiente kappa promedio: un nuevo parametro para evaluar y comparar el rendimiento de tests diagnosticos binarios. 2016.

[95] C. Ordun, S. Purushotham, and E. Raff. Exploratory analysis of covid-19 tweets using topic modeling, umap, and digraphs. *arXiv preprint arXiv:2005.03082*, 2020.

[96] R. M. Ortega-Mendoza, L. Villaseñor-Pineda, and M. Montes-y Gómez. Using lexical patterns for extracting hyponyms from the web. In *Mexican International Conference on Artificial Intelligence*, pages 904–911. Springer, 2007.

[97] M. Paluszek and S. Thomas. *MATLAB machine learning*, volume 1. Apress, 2016.

[98] C. Pandey. redbert: A topic discovery and deep sentiment classification model on covid-19 online discussions using bert nlp model. *International Journal of Open Source Software and Processes (IJOSSP)*, 12(3):32–47, 2021.

[99] R. Patel, S. Tanwani, and C. Patidar. Relation extraction between medical entities using deep learning approach. *Informatica*, 45(3), 2021.

[100] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.

[101] L. Pipanmekaporn, S. Kamonsantiroj, and E. Suriyachay. Learning short text representation using non-negative matrix factorization and word semantic correlations. In *2019 8th International Congress on Advanced Applied Informatics (IIAI-AAI)*, pages 617–622. IEEE, 2019.

[102] M. Pita, M. Nunes, and G. L. Pappa. Probabilistic topic modeling for short text based on word embedding networks. *Applied Intelligence*, pages 1–16, 2022.

[103] M. D. Pratama, R. Sarno, and R. Abdullah. Sentiment analysis user regarding hotel reviews by aspect based using latent dirichlet allocation, semantic similarity, and support vector machine method.

[104] J. Qiang, Z. Qian, Y. Li, Y. Yuan, and X. Wu. Short text topic modeling techniques, applications, and performance: a survey. *IEEE Transactions on Knowledge and Data Engineering*, 2020.

[105] X. Qin, Y. Lu, Y. Chen, and Y. Rao. Lifelong learning of topics and domain-specific word embeddings. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2294–2309, 2021.

[106] A. T. Rivera and J.-M. Torres-Moreno. Detecting new word meanings: a comparison of word embedding models in spanish. In *CORIA*, 2019.

[107] J. J. Rodríguez and A. M. M. Santana. Adquisición y desarrollo del lenguaje. *A. Muñóz García, Psicología del desarrollo en la etapa de educación infantil*, pages 101–120, 2010.

[108] T. Roger and O. Santillan. *Minería de opiniones basado en aprendizaje supervisado en la evaluación de destinos turísticos de la región de Puno.* PhD thesis, 07 2019.

[109] C. Saedi, A. Branco, J. Rodrigues, and J. Silva. Wordnet embeddings. In *Proceedings of the third workshop on representation learning for NLP*, pages 122–131, 2018.

[110] A. Salle and A. Villavicencio. Understanding the effects of negative (and positive) pointwise mutual information on word vectors. *Journal of Experimental and Theoretical Artificial Intelligence*, 2022.

[111] G. Salton and M. J. McGill. *Introduction to modern information retrieval*. mcgraw-hill, 1983.

[112] S. Sánchez-Cuadrado, J. Lloréns, J. Morato, and J. A. Hurtado. Extracción automática de relaciones semánticas. In *2da. Conferencia Iberoamericana en Sistemas, Cibernética e Informática. CISCI 2003*, pages 265–268, 2003.

[113] S. M. Saqlain, A. Nawaz, I. Khan, F. A. Shah, and M. U. Ashraf. Text clusters labeling using wordnet and term frequency-inverse document frequency. In *Proceedings of the Pakistan Academy of Sciences*, volume 53, pages 281–291.

[114] M. S. Satu, M. I. Khan, M. Mahmud, S. Uddin, M. A. Summers, J. M. Quinn, and M. A. Moni. Tclustvid: A novel machine learning classification model to investigate topics and sentiment in covid-19 tweets. *medRxiv*, page 31, 2020.

[115] I. Scarpino, C. Zucco, R. Vallelunga, F. Luzza, and M. Cannataro. Investigating topic modeling techniques to extract meaningful insights in italian long covid narration. *BioTech*, 11(3):41, 2022.

[116] H. Sha, M. A. Hasan, G. Mohler, and P. J. Brantingham. Dynamic topic modeling of the covid-19 twitter narrative among u.s. governors and cabinet executives. *arXiv*, abs/2004.11692:1–6, 2020.

[117] W. Shafqat et al. *A Hybrid Approach for Topic Discovery and Recommendations based on Topic Modeling and Deep Learning*. PhD thesis, Jeju National University, 2020.

[118] S. Shah, S. Reddy, and P. Bhattacharyya. A retrofitting model for incorporating semantic relations into word embeddings. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1292–1298, 2020.

[119] Z. Shahbazi and Y.-C. Byun. Topic modeling in short-text using non-negative matrix factorization based on deep reinforcement learning. *Journal of Intelligent & Fuzzy Systems*, 39(1):753–770, 2020.

[120] O. Shanidze and S. Petrasova. Extraction of semantic relations from wikipedia text corpus. *Computational Linguistics and Intelligent Systems*, 2:74–75, 2019.

[121] T. Shi, K. Kang, J. Choo, and C. K. Reddy. Short-text topic modeling via non-negative matrix factorization enriched with local word-context correlations. In *Proceedings of the 2018 World Wide Web Conference*, pages 1105–1114, 2018.

[122] A. Simanovsky and A. Ulanov. Mining text patterns for synonyms extraction. In *2011 22nd International Workshop on Database and Expert Systems Applications*, pages 473–477. IEEE, 2011.

[123] K. N. Singh, S. D. Devi, H. M. Devi, and A. K. Mahanta. A novel approach for dimension reduction using word embedding: An enhanced text classification approach. *International Journal of Information Management Data Insights*, 2(1):100061, 2022.

[124] A. Søgaard, I. Vulić, S. Ruder, and M. Faruqui. Cross-lingual word embeddings. *Synthesis Lectures on Human Language Technologies*, 12(2):1–132, 2019.

[125] A. Srivastava and C. Sutton. Autoencoding variational inference for topic models. *arXiv preprint arXiv:1703.01488*, 2017.

[126] V. Suárez-Paniagua, I. Segura-Bedmar, and P. Martínez. Labda at semeval-2017 task 10: Relation classification between keyphrases via convolutional neural network. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 969–972, 2017.

[127] C. D. Ta and T. P. Thi. Automatic extraction of semantic relations from text documents. In *International Conference on Future Data and Security Engineering*, pages 344–351. Springer, 2016.

[128] T. H. Ta, A. B. S. Rahman, G. Sidorov, and A. Gelbukh. Mining hidden topics from newspaper quotations: The covid-19 pandemic. In *Mexican International Conference on Artificial Intelligence*, pages 51–64. Springer, 2020.

[129] V. Teller. Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition. *Computational Linguistics*, 26(4):638–641, 2000.

[130] J. Torres. *Deep Learning Introducción práctica con Keras*. Lulu. com, 2018.

[131] A. Torres-Rondón, W. Hojas-Mazo, and A. J. Simón-Cuevas. Método de detección de tópicos en documentos basado en análisis contextual del contenido. *Informática*, 2018.

[132] M. Tovar, G. Flores, J. A. Reyes-Ortiz, and M. Contreras. Validation of semantic relation of synonymy in domain ontologies using lexico-syntactic patterns and acronyms. In *Mexican Conference on Pattern Recognition*, pages 199–208. Springer, 2018.

[133] M. Tovar, D. Pinto, A. Montes, G. González, and D. Vilarino. Identification of ontological relations in domain corpus using formal concept analysis. *Engineering Letters*, 23(2), 2015.

[134] M. Tovar, D. Pinto, A. Montes, and G. González-Serna. A metric for the evaluation of restricted domain ontologies. *Computación y Sistemas*, 22(1):147–162, 2018.

[135] M. Tovar, D. Pinto, A. Montes, G. González-Serna, and D. Vilariño. Evaluación de relaciones ontológicas en corpora de dominio restringido. *Computacion y sistemas*, 19(1):135–149, 2015.

[136] A. C. Vásquez, J. P. Quispe, A. M. Huayna, et al. Procesamiento de lenguaje natural. *Revista de investigación de Sistemas e Informática*, 6(2):45–54, 2009.

[137] S. Vázquez. *Resolución de la ambigüedad semántica mediante métodos basados en conocimiento y su aportación a tareas de PLN*. Universidad de Alicante, 2009.

[138] J. Vilares. *Aplicaciones del procesamiento del lenguaje natural en la recuperación de información en español*. PhD thesis, Universidade da Coruña, 2005. Departamento de Computación.

[139] G. C. V. Vilca. *Generación automática de resúmenes abstractivos mono documento utilizando análisis semántico y del discurso*. PhD thesis, Pontificia Universidad Católica del Peru-CENTRUM Catolica (Peru), 2017.

[140] Š. Vintar, L. G. Simeunović, M. Martinc, S. Pollak, and U. Stepišnik. Mining semantic relations from comparable corpora through intersections of word embeddings. In *Proceedings of the 13th Workshop on Building and Using Comparable Corpora*, pages 29–34, 2020.

[141] J. A. Wahid, L. Shi, Y. Gao, B. Yang, L. Wei, Y. Tao, S. Hussain, M. Ayoub, and I. Yagoub. Topic2labels: A framework to annotate and classify the social media data through lda topics and deep learning models for crisis response. *Expert Systems with Applications*, 195:116562, 2022.

[142] D. Wang, Y. Xu, M. Li, Z. Duan, C. Wang, B. Chen, M. Zhou, et al. Knowledge-aware bayesian deep topic model. *Advances in Neural Information Processing Systems*, 35:14331–14344, 2022.

[143] D. Wang, H. Zhao, D. D. Guo, X. Liu, M. Li, B. Chen, and M. Zhou. Bat-chain: Bayesian-aware transport chain for topic hierarchies discovery.

[144] W. Wei and C. Guo. A text semantic topic discovery method based on the conditional co-occurrence degree. *Neurocomputing*, 368:11–24, 2019.

[145] Y. Xu, D. Wang, B. Chen, R. Lu, Z. Duan, and M. Zhou. Hyperminer: Topic taxonomy mining with hyperbolic embedding. *arXiv preprint arXiv:2210.10625*, 2022.

[146] G. Xun, V. Gopalakrishnan, F. Ma, Y. Li, J. Gao, and A. Zhang. Topic discovery for short texts using word embeddings. In *2016 IEEE 16th international conference on data mining (ICDM)*, pages 1299–1304. IEEE, 2016.

[147] S. Yang, G. Huang, and B. Cai. Discovering topic representative terms for short text clustering. *IEEE Access*, 7:92037–92047, 2019.

[148] S. Yang and Y. Tang. News topic detection based on capsule semantic graph. *Big Data Mining and Analytics*, 5(2):98–109, 2022.

[149] D. Yu, D. Xu, D. Wang, and Z. Ni. Hierarchical topic modeling of twitter data for online analytical processing. *IEEE access*, 7:12373–12385, 2019.

[150] J. Yu, Y. Lu, and J. Muñoz-Justicia. Analyzing spanish news frames on twitter during covid-19—a network study of el país and el mundo. *International journal of environmental research and public health*, 17(15):5414, 2020.

[151] J. Zech, M. Pain, J. Titano, M. Badgeley, J. Schefflein, A. Su, A. Costa, J. Bederson, J. Lehar, and E. K. Oermann. Natural language–based machine learning models for the annotation of clinical radiology reports. *Radiology*, 287(2):570–580, 2018.

[152] L. Zhang, J. Hu, Q. Xu, F. Li, G. Rao, and C. Tao. A semantic relationship mining method among disorders, genes, and drugs from different biomedical datasets. *BMC Medical Informatics and Decision Making*, 20(4):1–11, 2020.

[153] Y. Zhang, W. Ji, H. Wang, X. Wang, and J. Chen. Mc-elda: Towards pathogenesis analysis in traditional chinese medicine by multi-content embedding lda. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 489–500. Springer, 2019.