Análisis sobre textos y algunas tareas

La expresión escrita es una de las denominadas destrezas lingüísticas –junto a la expresión oral, la comprensión auditiva y la comprensión lectora– y se refiere a la producción del lenguaje escrito. La expresión escrita se sirve del lenguaje verbal, las gráficas y los esquemas para manifestar ideas, pensamientos o sentimientos e igualmente consiste en construir un mensaje por medio del uso de signos de forma ordenada y de acuerdo con las normas de un idioma.

¿Qué es la **analítica de textos**? Podemos decir que es la disciplina que se ocupa de analizar y extraer información valiosa y significativa a partir de los textos comunicados mediante la expresión escrita. Utilizando técnicas del procesamiento de lenguaje natural y de machine learning, es posible identificar patrones, temas, sentimientos y relaciones presentes en grandes volúmenes de datos textuales no estructurados.

¿Qué tareas de aprendizaje están relacionadas con la analítica de textos? Veamos algunas (ver Fig. 1).

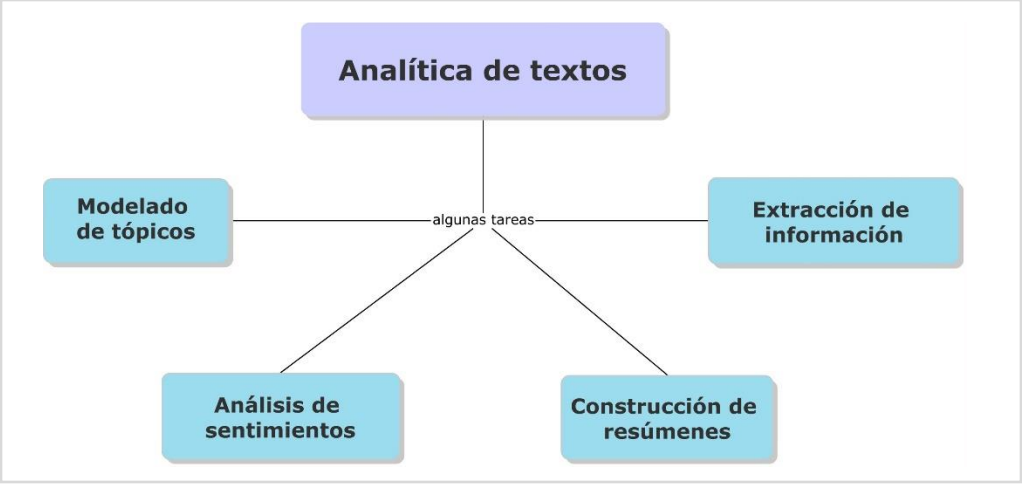

Fig. 1. Algunas tareas de la analítica de textos.

Modelado de tópicos. Tiene como objetivo identificar y extraer los temas predominantes en un conjunto de documentos, proporcionando una visión estructurada y resumida de la información textual. En este contexto, se asume que hay una estructura subyacente (latente) no observada en los

datos (los tópicos). La idea entonces es desglosar el corpus en términos de la distribución de tópicos que lo caracteriza.

Los tópicos están representados por patrones recurrentes, es decir, conjunto de palabras que tienden a aparecer juntas en los mismos contextos, por lo que se asumen que hacen referencia a los mismos temas. Así, el resultado del modelado de tópicos generalmente se presenta en forma de distribuciones de palabras y documentos asociadas a cada tópico, donde cada uno de ellos se representa como una lista de palabras con ponderaciones vinculadas. El tópico puede ser visualizado como una agrupación *soft* de palabras con reducción de la dimensionalidad, en la cual un documento pertenece a un grupo con una cierta probabilidad.

Análisis de sentimientos. También conocido como análisis de opiniones, se centra en determinar la actitud o la emoción expresada en un fragmento de texto, relacionado con entidades, individuos, servicios, eventos o temas. El objetivo principal es comprender si el texto tiene una connotación positiva, negativa o neutra y, en algunos casos, identificar emociones específicas (alegría, tristeza, enojo, etc.).

En general, hay dos enfoques para realizar esta tarea: semántico y basado en aprendizaje. El primero, utiliza diccionarios de términos (lexicones) con una orientación semántica que indica alguna polaridad u opinión. Incluye un tratamiento de parámetros modificadores (como "muy", "poco", "demasiado") que aumentan o reducen la polaridad de los términos, así como de parámetros inversores o negadores (como "no" y "tampoco"), que invierten la polaridad de los términos a los que afectan. El segundo, construye un clasificador a partir de una colección de textos anotados, con textos representados con base en vectores en combinación con otras características semánticas que modelan, por ejemplo, la intensificación, la negación y la subjetividad, entre otros.

Construcción automática de resúmenes. Es un proceso mediante el cual se genera, de manera automática, un resumen conciso y coherente de un documento. El objetivo es condensar la información esencial manteniendo la relevancia y coherencia del contenido original. Se emplean diversas técnicas en el campo del procesamiento de lenguaje natural para llevar a cabo esta tarea.

En general, se pueden realizar dos tipos de resúmenes: extractivo y abstractivo. El primero está constituido de unidades de información extraídas del texto original, como palabras y frases. También es posible, como en el modelado de tópicos, identificar los temas principales del texto. Al

destacar las secciones más relevantes relacionadas con estos tópicos, se construye un resumen que refleja la esencia del contenido. Por su parte, el abstractivo implica la creación de un texto nuevo que no está limitado a las frases originales. Se utilizan técnicas más avanzadas, como modelos de lenguaje generativos, que pueden entender y reformular el contenido en términos más breves y precisos.

Extracción de información. Se centra en la identificación y recuperación de información específica y estructurada a partir de documentos. El objetivo es transformar datos no estructurados en información organizada y relevante. Esta información pueden ser entidades, objetos o conceptos específicos (por ejemplo, nombres de personas, organizaciones, ubicaciones y fechas) que están presentes en el texto. También es posible extraer relaciones semánticas para determinar conexiones y asociaciones entre entidades (por ejemplo, sintoma_de, autor_de y CEO_de). Esta tarea está relacionada con el reconocimiento de entidades nombradas (*Named-entity recognition*, NER), técnica que se centra en identificar y clasificar entidades específicas en un texto.

¿Cómo preparar los textos para estas tareas?

Los datos textuales son altamente no estructurados, ya que no se adhieren a alguna sintaxis o patrón regular. Por lo tanto, no es posible que los algoritmos de machine learning los utilicen directamente. Será necesario entonces aplicar transformaciones a los textos para llevarlos a un formato estructurado y numérico, pero tratando de no perder el contexto y las relaciones semánticas. Actualmente, los modelos de lenguaje pre-entrenados basados en transformadores (*tranformers*) e incrustaciones de palabras (*word embedding*), como Word2Vec, Glove y BERT, han mejorado significativamente la capacidad de comprensión y extracción de información, permitiendo una representación semántica avanzada del texto.

Bibliografía

Albrecht, J., Ramachandran, S., Winkler, C. (2020) Blueprints for Text Analytics Using Python. O'Reilly Media.

Universidad de los Andes | Vigilada Mineducación Reconocimiento como
Universidad: Decreto 1297 del 30 de mayo de 1964.
Reconocimiento personería jurídica: Resolución 28 del 23 de febrero de 1949 Minjusticia.