Módulo para el análisis de la percepción que tienen los turistas de la ciudad de Santa Marta usando datos colaborativos

Jerson David Ferrer Marcony

Universidad Magdalena

Facultad de ingeniería

Programa de Ingeniería de Sistemas

Santa Marta, Colombia

2022

Módulo para el análisis de la percepción que tienen los turistas de la ciudad de Santa Marta usando datos colaborativos

Jerson David Ferrer Marcony

Trabajo presentado como requisito parcial para optar al título de:

Ingeniero de Sistemas

Director:

PhD Alexander Armando Bustamante Martínez

Línea de Investigación:

Análisis de datos

Grupo de Investigación:

Desarrollo y Gestión de Tecnologías para las Organizaciones y la Sociedad - TecnOS

Universidad del Magdalena

Facultad de ingeniería

Programa de Ingeniería de Sistemas

Santa Marta, Colombia

2022

Nota de aceptación:

Aprobado por el Consejo de Programa en cumplimiento de los requisitos exigidos por la Universidad del Magdalena para optar al título de Ingeniero de Sistemas.

_____

Jurado

_____

Jurado

Santa Marta, ____ de ____del _____

Dedico este trabajo a mi familia, que siempre me ha apoyado en todo lo que emprendo, a pesar de las adversidades.

AGRADECIMIENTOS

Quiero agradecer primeramente a mi madre Felisa Marcony, ya que con su apoyo incondicional me brindó la posibilidad de formarme profesionalmente, y de crecer cada día como una persona integra.

A toda mi familia que en mayor o menor medida han contribuido en mi desarrollo como profesional, y como persona.

A todas las personas que han alentado en mí el deseo de conocimiento, de saber más cada día, y de intentar hacer las cosas de la mejor forma posible siempre. Hago especial mención a Adrián Quiñonez y Daniel Fontalvo, que son amigos incondicionales.

Quiero extender un especial agradecimiento a la persona que fungió como director de este proyecto, el PhD Alexander Bustamante, que me guio durante todo el proceso compartiéndome los conocimientos necesarios para abordar de la mejor manera posible el desarrollo de este retador proyecto.

Resumen

Este trabajo presenta el resultado de investigación, diseño, y desarrollo de un módulo para el análisis de la percepción que tienen los turistas de la ciudad de Santa Marta usando datos colaborativos. Para lo cual se utilizaron datos extraídos de la red social *Twitter*, a los cuales se les realizaron procedimientos de preprocesamiento, desambiguación de sentidos, y clasificación en dominios conceptuales (turismo, comida, comercio, entre otros), con el uso de técnicas de Procesamiento del Lenguaje Natural (PLN). Además, una vez clasificados los *tweets* se realizó un componente de visualización, donde se ubicaron en un mapa los lugares desde los que fueron publicados los *tweets*, con el fin de determinar si fueron realizados cerca de algún lugar turístico, histórico, comercial (restaurantes, hoteles, etc.), entre otros.

Palabras Clave: Procesamiento del Lenguaje Natural, Clasificación de Textos, Datos Colaborativos, Turismo.

Abstract

This work presents the result of research, design, and development of a module for the analysis of the perception that tourists have of the city of Santa Marta using collaborative data. For which data extracted from the social network Twitter were used, to which procedures of preprocessing, disambiguation of senses, and classification into conceptual domains (tourism, food, commerce, among others) were conducted, with the use of Natural Language Processing (NLP) techniques. In addition, once the *tweets* were classified, a visualization component was made, where the places from which the tweets were published were located on a map, in order to determine if they were made near a tourist, historical, commercial place (restaurants, hotels, etc.), among others.

Keywords: Natural Language Processing, Text Classification, Collaborative Data, Tourism.

Contenido

Lista de figuras

Pág.

Lista de tablas

Pág.

Introducción

En este informe final de pasantía de investigación se recoge el proceso de desarrollo, y se presentan los resultados de índole académico, del trabajo “Módulo para el análisis de la percepción que tienen los turistas de la ciudad de Santa Marta usando datos colaborativos” realizado en el programa de Ingeniería de Sistemas en la Universidad del Magdalena.

La actividad turística es una de las que más relevancia tiene en departamentos como el Magdalena, siendo una de las actividades que más crecimiento ha tenido en los últimos años. Tal como se puede ver en [1], la llegada de visitantes extranjeros por puntos de control migratorio, principalmente aéreo, no paró de crecer desde el año 2012 hasta el año 2019 en el departamento del Magdalena. De igual forma, el número de visitantes a parques naturales tuvo un aumento del 112,11% entre los años 2020 y 2021, así como también ha crecido el número de prestadores de servicios turísticos activos en el Registro Nacional de Turismo (RNT) en un 38,79%, en el mismo periodo de tiempo [2].

Enfocando las estadísticas hacia la cuidad de Santa Marta se puede ver que, según cifras del año 2019, los motivos por los cuales un viajante llega a la cuidad son principalmente vacaciones, recreo y ocio, teniendo estos una proporción del 91,31% [3]. Como se concluye en [4], el turismo como sector es un tema que influye en diferentes aspectos del desarrollo de la ciudad de Santa Marta […], durante los años de 2010 en adelante, ha crecido considerablemente y tenemos un sector bastante fuerte en la economía Samaria. Además, como se observa en [5], el turismo competitivo hace parte de uno de los ejes estratégicos para el desarrollo del distrito, y en donde se plantea como meta incrementar en un 1,51% el puntaje del Índice de Competitividad Turística Regional de Colombia (ICTRC) para la ciudad de Santa Marta.

Las cifras presentadas anteriormente muestran la clara importancia que tiene el sector turismo en el departamento del Magdalena, por lo cual cobran relevancia las actividades que se lleven a cabo para tratar de mejorar la competitividad del sector. Una forma de mejorar esta competitividad es haciendo uso de las tecnologías de la información y la comunicación. Un ejemplo claro de esto es hacer uso del concepto de turismo 2.0 (el cual es la explotación de las ventajas de la Web 2.0 en el sector del turismo) buscando potenciarlo con todos los beneficios de la web. Además, se pueden hacer uso de datos extraídos de redes sociales con el fin de analizarlos y obtener información relevante acerca de los comportamientos y la percepción de los turistas.

Con todo lo anterior claro se hace evidente que, a nivel técnico, una de las cosas que se puede hacer para mejorar el sector turismo a través de la tecnología, es analizar los datos de las redes sociales, con el fin de brindarles a los turistas productos y servicios que se adapten mejor a sus necesidades, gustos, recursos, etc. Los datos extraídos de redes sociales pueden ayudar a analizar la percepción de los turistas hacia determinados sectores, lo cual se evidencia en [6], donde se utilizaron datos extraídos de publicaciones de *Facebook* para determinar el empoderamiento psicológico de los anfitriones de Airbnb en Brasil. También se puede observar cómo en [7] se utilizó la red social *Instagram* para recolectar y analizar datos de los prestadores de servicios turísticos, como número de seguidores, número de publicaciones, tipo de publicación y contenido de estas. Y luego de analizar la interacción de los usuarios de estos servicios se observó que:

> *[…] Las fotos más populares en los perfiles analizados son las de platos y menús promocionales. Los clientes que reciben respuestas positivas de sus comentarios se muestran satisfechos. También fue posible diferenciar los restaurantes que utilizan la red social para difusión y un contacto más fluido con el cliente y los que la utilizan sólo para difusión.* [7].

En [8] se utilizaron datos de *Facebook*, *Twitter*, *Instagram* y *YouTube* para analizar la estrategia de promoción en las redes sociales desarrollada por el Consejo de Promoción Turística de México, pudiendo llegar a conocerse cosas como las épocas del año en que la entidad intensifica su intervención en las redes sociales, o comprobar los temas más utilizados en las publicaciones, y también pudiendo determinar que la entidad hace continuas referencias a la historia, la cultura, la naturaleza, la tradición, la gastronomía, la

arqueología, la playa, etc. como principales referencias y valores de México como destino turístico.

En síntesis, este proyecto busca, construir un módulo para el análisis de la percepción que tienen los turistas de la ciudad de Santa Marta usando datos colaborativos y Procesamiento del Lenguaje Natural (PLN). Trabajar en el actual proyecto, teniendo en cuenta todo lo descrito con anterioridad, tiene una relevancia a nivel tanto académico como profesional de gran peso, puesto que usar la tecnología para mejorar sectores como el del turismo, es algo que tiene un impacto muy grande a nivel económico y social, en una ciudad turística como Santa Marta, así como también en todo el departamento del Magdalena.

Este informe consta de ocho capítulos donde se abordan todas las partes del trabajo, empezando por el capítulo I donde se describe el marco conceptual para aclarar los conceptos y definiciones necesarios para entender el documento. Continua el capítulo II que es la presentación del proyecto, donde se presentan elementos como el título, los objeticos y la metodología. En el capítulo III, se muestra un inventario de las herramientas *Open Source* disponibles para el Procesamiento del Lenguaje Natural, así como sus ventajas y desventajas. El capítulo IV está conformado por la revisión a la literatura de proyectos donde se haya utilizado técnicas de Procesamiento de Lenguaje Natural para analizar datos procedentes de *Twitter*. El capítulo V es sobre el diseño del módulo software para el procesamiento de los datos extraídos de *Twitter*. En el capítulo VI, se presenta la implementación del módulo de clasificación de *tweets* en un dominio conceptual. Seguido, está el capítulo VII que contiene la parte de las visualizaciones de los resultados obtenidos. Por último, está el capítulo donde se consignan las conclusiones y aportes sobre los resultados de todo el trabajo.

Los alcances de la pasantía son:

- Desarrollar un módulo software que permita el preprocesamiento, extracción de entidades, desambiguación de sentidos, clasificación de *tweets*, y la visualización de los resultados obtenidos.
- Una ponencia enviada a un congreso nacional para divulgar todo lo desarrollado en la pasantía.

Las limitaciones presentes para esta pasantía son:

- Los resultados del análisis de los datos tienen una intención descriptiva y no prescriptiva.

1. Marco conceptual

Este capítulo tiene como objetivo aclarar los conceptos y definiciones necesarios para entender con claridad los posteriores capítulos del documento. También se presentan algunos trabajos donde se han extraído datos de las redes sociales, y se han utilizado para ayudar a mejorar diferentes partes del sector del turismo.

La Inteligencia Artificial (IA) definida en palabras de Bellman [9], citado en [10], es: "[La automatización de] actividades que vinculamos con procesos de pensamiento humano, actividades como la toma de decisiones, resolución de problemas, aprendizaje…". Es decir, la IA tiene como fin realizar procesos que normalmente haría un humano (selección y clasificación, por ejemplo), pero de una manera más eficiente con el fin de ayudar a solucionar problemas y tomar decisiones de una manera más eficiente.

Dentro de la IA se encuentra el Procesamiento del Lenguaje Natural (PLN) que, tal como se define en [11], es la habilidad de la máquina para procesar la información comunicada, no simplemente las letras o los sonidos del lenguaje. Con lo cual, el PLN tiene una gran variedad de aplicaciones prácticas, entre las que se encuentra la clasificación de textos en algún dominio conceptual, y la detección de sentimientos en los mismos.

El PLN es frecuentemente combinado con otros subtemas de la IA como lo son el Aprendizaje de Maquinas (AM), que se puede definir como el área de la IA que engloba un conjunto de técnicas que hacen posible el aprendizaje automático a través del entrenamiento con grandes volúmenes de datos [12].

El concepto de turismo 2.0 hace referencia a la explotación de las ventajas de la Web 2.0 en el sector del turismo, buscando potenciarlo con todos los beneficios de la web. Por lo que es importante resaltar que:

> *El turismo 2.0 es un complemento indispensable de los viajes en internet, surge con la web 2.0 la cual permite a sus usuarios la interacción y colaboración entre sí, es decir permite además de interactuar con los clientes crear, compartir e intercambiar información entre las empresas y los órganos públicos gestores del turismo […]. Las redes sociales, por otro lado, suponen un medio fundamental que contribuye de forma muy positiva al marketing online posibilitando analizar usuarios, posicionamiento de la empresa, abrir nuevas líneas de negocio, fidelización de clientes, etc.* [13].

En [7] se analizó cómo era el uso de la red social *Instagram* por parte de seis proveedores de servicios turísticos gastronómicos de la ciudad de Recife (Brasil). Se utilizó un abordaje cualitativo con el fin de estudiar y entender el objeto de estudio, para lo cual se realizó un trabajo de campo mediante una inserción directa en el lugar de estudio siguiendo una ruta de observación de permitiera percibir las publicaciones en el perfil de *Instagram* de los seis establecimientos durante 2 meses, entre diciembre de 2015 y enero de 2016. Los resultados obtenidos mostraron una multiplicidad de formas de comunicarse con el consumidor y una posibilidad real de estrechar la relación con la empresa. Además, se descubrió que las publicaciones sobre los platos clave, promociones y acciones exclusivas son importantes para los clientes que acceden al perfil, pues logran mantener al consumidor atento a lo que está sucediendo con la empresa, además de obtener información general sobre servicios sin tener que acceder al sitio oficial.

En [8] se analizó la estrategia seguida en las redes sociales por el Consejo de Promoción Turística de México, para lo cual se obtuvieron datos de las publicaciones realizadas en las redes sociales de *Facebook* y *Twitter*. Adicionalmente se tomaron datos como el número de seguidores, reacciones, comentarios y recomendaciones con el fin de tener un análisis completo. Con todos esos datos se determinó que la red social donde el Consejo de Promoción turística de México tiene más seguidores es *Facebook*, mientras que *YouTube* está en el último lugar. Por lo que, si se requiriera hacer una campaña de promoción, *Facebook* seria la red social que más impacto tendría a nivel de difusión. El estudio también concluyó que la palabra que tenía más frecuencia de uso era “viajemostodospormexico”, destacando también que los destinos turísticos más mencionados son, entre otros, Guanajuato, Jalisco, Nayarit, Chiapas, Tabasco, Oaxaca, Durango, Yucatán, Acapulco, Michoacán, Veracruz y Sinaloa. Así también, las principales

razones a las que se hace referencia para visitar México son, la historia, la cultura, la naturaleza, la tradición, la gastronomía, la arqueología, la playa, entre otras.

En [6] se analizaron datos obtenidos de *Facebook* sobre la actuación y capacitación de los anfitriones de la plataforma de alojamiento de economía colaborativa *Airbnb* en Brasil. Se utilizó un enfoque cualitativo para determinar los efectos de la participación en comunidades virtuales por parte de los anfitriones de *Airbnb*, además se hicieron entrevistas a 14 participantes para tener una mejor comprensión del grupo y las implicaciones que tendría en el mismo. Los resultados determinaron que el aprendizaje y la actualización sobre los temas concernientes a la plataforma de *Airbnb* son las principales finalidades de participación en el grupo, las cuales son alcanzadas gracias a la interacción con los demás o de las lecturas de las publicaciones y los comentarios hechos en la comunidad. Lo anterior sugiere la existencia de una posible influencia de ese contenido en el conocimiento de los anfitriones sobre la plataforma, y además puede indicar que esos miembros atribuyen a los más experimentados un grado de confianza suficiente como para seguir sus recomendaciones.

2. Presentación del proyecto

El título del trabajo es “Módulo para el análisis de la percepción que tienen los turistas de la ciudad de Santa Marta usando datos colaborativos” siendo lo más preciso posible para enmarcar todo el trabajo realizado en la pasantía.

El objetivo principal de la pasantía fue desarrollar un módulo software para el análisis de la percepción que tienen los turistas de la ciudad de Santa Marta usando datos colaborativos y Procesamiento del Lenguaje Natural (PLN). Para lo cual, se planteó el siguiente objetivo general:

- Desarrollar un módulo dentro de una plataforma de análisis de datos del sector turismo que permita analizar datos procedentes de *Twitter* utilizando Procesamiento del Lenguaje Natural (PLN).

Cumpliendo los siguientes objetivos específicos:

1. Hacer un inventario de las herramientas *Open Source* disponibles para el Procesamiento del Lenguaje Natural identificando las ventajas y las desventajas de cada una de estas, con el propósito de que se pueda decidir cuál utilizar y por qué.
2. Revisar en la literatura, proyectos donde se haya utilizado técnicas de Procesamiento de Lenguaje Natural para analizar datos procedentes de *Twitter* de manera que se puedan identificar buenas prácticas y recomendaciones que puedan ser incorporadas en el módulo a desarrollar.
3. Diseñar un módulo software que contemple el preprocesamiento de los tweets, el reconocimiento de entidades nombradas y estrategias para la desambiguación de sentidos de palabras dentro del *tweet*, para que se puedan utilizar en otras tareas de análisis.
4. Implementar rutinas para la clasificación de los *tweets* en un dominio conceptual, mediante el uso de recursos o herramientas de PLN que proporcionen estos dominios

conceptuales, para determinar si los *tweets* se encuentran dentro del dominio del turismo.

5. Construir visualizaciones de los resultados obtenidos, utilizando librerías especializadas, para facilitar su análisis e interpretación.

Para el desarrollo del proyecto se siguió un plan de actividades como el que se muestra en la . En esta se muestra a nivel general las cinco actividades generales que se detallarán a continuación:

- **Revisión de la literatura y de las herramientas**: referente a los algoritmos necesarios para el enriquecimiento de los *tweets*, para la desambiguación de sentidos, y para clasificación de *tweets*. Todos los algoritmos mencionados hacen que la clasificación por temas de los *tweets* se realice de una manera más efectiva, puesto que se eliminan ambigüedades dentro de los sentidos de los contextos de los *tweets*, y también se hace un enriquecimiento semántico de los *tweets*.

- **Diseño del módulo**: llevando a cabo el diseño cada uno de los diferentes componentes necesarios para la clasificación de los *tweets*, tales como el preprocesamiento, el reconocimiento de entidades, la desambiguación de sentidos, así como la clasificación en dominios.

- **Implementación y pruebas del módulo**: en sus diferentes componentes. Para la parte de preprocesamiento es necesario eliminar de los *tweets* los caracteres innecesarios para el procesamiento de los *tweets*. Para el componente de reconocimiento de entidades nombradas, se utilizará alguna librería de *Python* destinada a ese fin. El componente de desambiguación de sentidos se implementará haciendo uso de algoritmos para la desambiguación de sentidos que se encuentra en la literatura científica. Para la clasificación de los tweets se usará una herramienta que contenga dominios conceptuales, la cual contiene múltiples sentidos asociados a uno o varios de esos dominios.

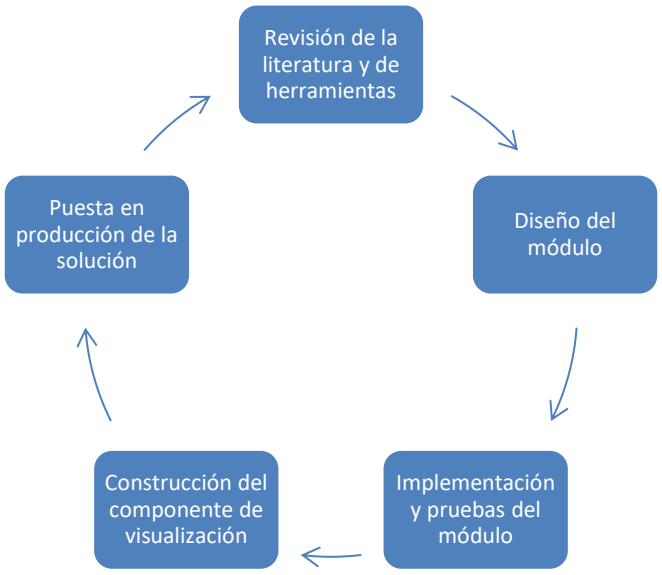
  Se realizarán pruebas de todos los componentes anteriormente descritos haciendo uso de herramientas automatizadas para este fin. Por último, se hará la unificación de todos

los componentes, haciendo uso de una metodología incremental, donde en cada revisión de la implementación se detectan y corrigen errores.

- **Construcción del componente de visualización:** de los resultados obtenidos, utilizando librerías especializadas, para facilitar su análisis e interpretación. Este componente estará completamente orientado a una visualización de calidad en entornos web.

- **Puesta en producción de la solución:** con todos los componentes unificados y libres de errores.

Figura 1. Actividades para realizar.



Fuente: Creación propia.

Adicionalmente, la ***Figura 2*** muestra el cronograma de actividades que se siguió durante toda la realización del proyecto. En el cronograma se observa cada una de las actividades ubicadas en el tiempo, para este caso la unidad de tiempo es de semanas.

Figura 2. Cronograma de actividades.

ACTIVIDADES	Mes 1				Mes 2				Mes 3				Mes 4				Mes 5			
Tiempo (Semanas)	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
ETAPA DE REVISIÓN LITERARIA																				
Búsqueda de artículos sobre PLN																				
Lectura de los artículos encontrados																				
Selección de los artículos de mayor interés																				
Síntesis de la información obtenida de los artículos																				
ETAPA DE DISEÑO																				
Diseño del componente de preprocesamiento																				
Diseño del componente de reconocimiento de entidades																				
Diseño del componente de desambiguación de sentidos																				
Diseño del componente de clasificación de tweets																				
ETAPA DE IMPLEMENTACIÓN Y PRUEBAS																				
Implementación del componente de preprocesamiento																				
Implementación del componente de reconocimiento de entidades																				
Implementación del componente de desambiguación de sentidos																				
Implementación del componente de clasificación de tweets																				
Corrección de errores																				
Integración de los componentes																				
Validación de pruebas de funcionalidad																				
ETAPA DE CONTRUCCIÓN DEL COMPONENTE DE VISUALIZACIÓN																				
Búsqueda de herramientas para la visualización																				
Selección de las herramientas de mayor utilidad																				
Construcción del componente																				
ETAPA DE PUESTA EN PRODUCCIÓN																				
Integración con los demás módulos de la plataforma																				
Puesta en producción del módulo completo																				

Fuente: Creación propia.

3. Herramientas para el Procesamiento del Lenguaje Natural (PLN)

Para este capítulo se hará un recuento de algunas de las herramientas *Open Source* que hay disponibles en internet para el PLN, de tal forma que se puedan conocer las características más relevantes de cada una de ellas, y de esta manera poder elegir las más adecuadas para el proceso de construcción del módulo software.

3.1 Preprocesamiento

El PLN consta de varias etapas de estructuración del texto (preprocesamiento de la información para extraer los datos que puedan ser analizados), por tanto, existen diversidad de herramientas que se enfocan en una de estas etapas específicamente, así como también hay algunas otras herramientas que abarcan varias de estas etapas. Estas etapas de las que se hablan son descritas en [14] como:

- *Tokenization*: dividir una secuencia de caracteres en *tokens* (“pedazos”) y despreciar los caracteres sin información como signos de puntuación.
- *Named Entity Recognition (NER)*: localizar en el texto y clasificar nombres de entidades relevantes conocidas o establecidas previamente.
- Normalización: relacionar cada entidad con un identificador.
- *Sentence segmentation*: dividir el texto en frases.
- *POS (Part Of Speech) Tagging* (Etiquetado): clasificar las palabras el texto (etiquetarlas) acorde a su función.
- *Relation Identification* (RI) (Identificación de relaciones - IR): reconocer las relaciones existentes entre las entidades detectadas.
- Clasificación: clasificar el [texto] en base al tema.

Teniendo en cuenta lo anteriormente mencionado, se hizo una investigación de herramientas *open source* y en la ***Tabla 1*** se muestra el resultado de esa investigación, mostrando la descripción de cada herramienta, así como algunas ventajas y desventajas. Con esa información se tomó la decisión sobre las herramientas para realizar el trabajo.

Tabla 1. Inventario de herramientas *open source* para PLN.

Herramienta	Descripción	Etapas	Ventajas	Desventajas
Natural Language Toolkit (NLTK)	Plataforma para crear programas de *Python* para trabajar con datos en lenguaje humano.	▪ *Tagging.* ▪ *Tokenization.* ▪ *Sentence segmentation.* ▪ Clasificación.	▪ Multiplataforma. ▪ Documentación amplia. ▪ Foro de discusión activo. ▪ Proporciona interfaces fáciles de usar para más de 50 corpus y recursos léxicos.	▪ Alto consumo de memoria. ▪ Requiere mayor espacio de almacenamiento.
General Architecture for Text Engineering (GATE)	Plataforma para crear programas de *Java* para todo tipo de tareas computacionales que involucran lenguaje humano	▪ *Tagging.* ▪ *Tokenization.* ▪ *Sentence segmentation.* ▪ Clasificación. ▪ *NER.*	▪ Documentación amplia. ▪ Gran comunidad. ▪ Tiene un marco de trabajo.	▪ Alta curva de aprendizaje. ▪ Muchas herramientas para integrar.
spaCy	Librería para procesamiento de lenguaje natural desarrollada en *Python*.	▪ *Tagging.* ▪ *Tokenization.* ▪ *Sentence segmentation.* ▪ *NER.* ▪ *RI.* ▪ Clasificación.	▪ Documentación amplia. ▪ Soporte para más de 64 idiomas. ▪ Componentes de canalización entrenables. ▪ Precisión robusta y rigurosamente evaluada.	▪ No tiene una comunidad tan grande. ▪ Curva de aprendizaje media.

Fuente: Creación propia.

Tabla 1. (Continuación)

Herramienta	Descripción	Etapas	Ventajas	Desventajas
Ellogon	Software de interfaz gráfica para realizar diferentes tareas de procesamiento del lenguaje humano.	▪ *Tagging.* ▪ *NER.* ▪ *RI.* ▪ Clasificación.	▪ Arquitectura extensible basada en componentes: todos los aspectos de Ellogon se pueden ampliar con componentes proporcionados por el usuario. ▪ Los componentes se pueden escribir en: *C*, *C++*, *Tcl*, *Java*, *Python*, *Perl*. ▪ Arquitectura descomponible: las piezas de Ellogon se pueden utilizar para formar aplicaciones especializadas. ▪ Multiplataforma.	▪ Hay que crear los componentes que no existan. ▪ No tiene una comunidad grande.
Standford CoreNLP	Librería para diversas tareas de procesamiento del lenguaje natural desarrollada en *Java*.	▪ *Tagging* ▪ *Tokenization.* ▪ *NER* ▪ *RI*	▪ Soporta más de 8 idiomas. ▪ Buena documentación. ▪ Se puede descargar la librería para un idioma especifico ▪ Multiplataforma.	▪ No tiene una comunidad grande. ▪ Proceso de instalación tedioso.

Fuente: Creación propia.

3.2 Diccionarios de sentidos

Los diccionarios de sentidos son bases de datos que recopilan de forma organizada las glosas (definición de sentido) para las palabras de un idioma. Por lo tanto, para el desarrollo de este trabajo es necesario el uso de este tipo de herramientas que serán útiles a la hora de procesar y analizar los textos. Actualmente existen dos diccionarios de sentidos organizados en forma de árbol (ontología léxica), los cuales son los más populares por su robustes, documentación y facilidad de implementación. Estos recursos se presentan y describen a continuación:

- **WordNet:** es una gran base de datos léxica del inglés. Los sustantivos, verbos, adjetivos y adverbios se agrupan en conjuntos de sinónimos cognitivos (*synsets*), cada uno de los cuales expresa un concepto distinto. Los *synsets* están interconectados por medio de relaciones conceptuales, semánticas y léxicas. *WordNet* interconecta no solo las formas de las palabras (cadenas de letras), sino también los sentidos específicos de las palabras. Como resultado, las palabras que se encuentran muy cerca unas de otras en la red se desambiguan semánticamente. En segundo lugar, *WordNet* etiqueta las relaciones semánticas entre palabras, mientras que las agrupaciones de palabras en un diccionario de sinónimos no siguen ningún patrón explícito que no sea la similitud de significado. Cada uno de los 117 000 *synsets* de *WordNet* está vinculado a otros *synsets* por medio de un pequeño número de "relaciones conceptuales". Además, un *synset* contiene una breve definición (glosa) y, en la mayoría de los casos, una o más oraciones cortas que ilustran el uso de los miembros del *synset*. Las formas de palabras con varios significados distintos se representan en tantos *synsets* distintos. Por lo tanto, cada par forma-significado en *WordNet* es único [15].

- **EuroWordNet:** es una base de datos multilingüe con redes de palabras para varios idiomas europeos (holandés, italiano, español, alemán, francés, checo y estonio). Las redes de palabras están estructuradas de la misma manera que *WordNet*, en términos de *synsets* con relaciones semánticas básicas entre ellos. Cada red de palabras representa un sistema interno de lexicalización único en el idioma. Además, las redes de palabras están vinculadas a un índice interlingüístico, basado *WordNet*. A través de este índice, los idiomas están interconectados de manera que es posible pasar de las palabras de un idioma a palabras similares en cualquier otro idioma. El índice también

da acceso a una ontología superior compartida de 63 distinciones semánticas. Esta ontología superior proporciona un marco semántico común para todos los idiomas, mientras que las propiedades específicas del idioma se mantienen en las redes de palabras individuales. La base de datos se puede utilizar, entre otros, para la recuperación de información monolingüe y multilingüe [16].

3.3 Herramientas seleccionadas

Para la realización de este trabajo, se decidió hacer uso de las siguientes herramientas: ***NTLK***, en vista de que es una herramienta robusta y con amplia documentación. ***SpaCy***, por su amplio soporte de idiomas. Y para la parte de diccionario de sentidos, se decidió utilizar ***WordNet***, al ser una base de datos léxica bastante amplia, y sencilla de implementar.

4. Literatura sobre PLN

En este capítulo se mostrará la revisión de la literatura sobre PLN, para lo que se utilizaron artículos científicos de proyectos donde se hayan utilizado técnicas de PLN haciendo uso de datos extraídos de *Twitter*, así como también de técnicas de desambiguación de sentidos para los casos donde sean necesarias. La revisión de los textos científicos tuvo como finalidad observar buenas prácticas que pudieran ser aplicadas al desarrollo del módulo.

4.1 Revisión de la literatura

En [17] se plantean diferentes técnicas para clasificar datos extraídos de *Twitter* relacionados con eventos que serán realizados, para lo cual se utilizan diversas técnicas de PLN como la segmentación de entidades nombradas, y la extracción de los eventos mencionados en el texto. Para el proceso de clasificación de los eventos se utiliza un enfoque basado en modelos de variables latentes que infiere un conjunto apropiado de tipos de eventos. Con lo que se concluyó que, debido a que este enfoque puede aprovechar grandes cantidades de datos sin etiquetar, supera a una línea de base supervisada en un 14%.

En [18] se utilizan técnicas enfocadas al uso de la semántica para la clasificación de *tweets* en función de su tema. En este trabajo se presentan diferentes técnicas para el preprocesamiento de los *tweets*, con el fin de que el texto este limpio a la hora de hacer la clasificación. Con el enfoque semántico, se busca hacer uso de los diferentes significados de las palabras, así como los de sus sinónimos en ánimos de buscar una relación para que a la hora de realizar la clasificación, esta funcione de una manera más precisa. Para la clasificación de los tweets se utiliza una herramienta complementaria de *WordNet*, la cual es *eXtended WordNet Domains* (*XWND*). Se concluyó que este enfoque mejora la clasificación de tweets, comparado con un enfoque netamente léxico.

En [18] y [19] se presenta el método de *Specification Marks* (*SM*) para el proceso de desambiguación de sentidos de palabras, la hipótesis básica del método *SM* es que cuanto mayor es la similitud entre dos palabras, mayor cantidad de información es compartida por dos de sus conceptos, donde un concepto corresponde a un *SM* esculpido en forma de árbol. El algoritmo utilizado para la desambiguación de palabras recibe un conjunto de palabras $W = \{w_1, w_2, \dots, w_n\}$ en un contexto, luego para cada w_i se obtiene un conjunto de posibles sentidos $Si = \{Si_1, Si_2, \dots, Si_n\}$, y cada sentido tiene un conjunto de conceptos asociado.

En [19] y [20] se describen diferentes heurísticas que pueden ser aplicadas en caso de que el algoritmo de desambiguación de sentidos principal falle. Estas heurísticas son algoritmos que se utilizan de tal forma que se toman las palabras que necesitan ser desambiguadas, y se les aplican ciertos procedimientos que pueden llevar a que la ambigüedad desaparezca. Las heurísticas pueden ser usadas por separado o se pueden usar una seguida de la otra, pasándole a los algoritmos las palabras que no pudieron ser desambiguadas. Algunas de estas heurísticas son: heurística de definición, heurística de *SM* común, y heurística de dominio.

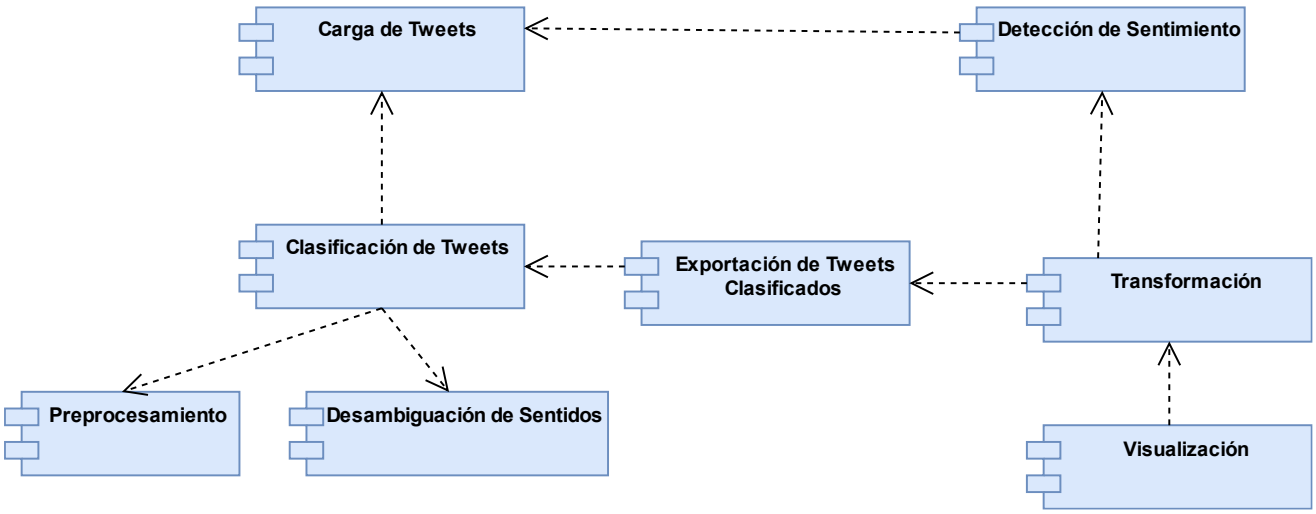
4.2 Conclusiones del capitulo

Toda la literatura revisada ayudó a aclarar el panorama sobre cómo abordar el desarrollo del módulo software. Se determinó la importancia de hacer un preprocesamiento de los *tweets* antes de realizar cualquier otro proceso, con el fin de sean más efectivos. Se observó que el método de *SM* para desambiguar sentidos es un método potente y de fácil implementación, y que es una buena opción para ser usado como método de desambiguación principal. Además, se observaron buenas prácticas en el proceso de PLN, tales como hacer uso de heurísticas en el caso de que el algoritmo de desambiguación de sentidos principal no pueda quitar la ambigüedad de alguna palabra. Y para la parte de clasificación, se destaca el uso de la herramienta *XWND* para usar los dominios conceptuales y obtener una clasificación óptima.

5. Diseño del módulo software

En este capítulo se mostrará la parte de diseño del módulo software, en el cual se mostrarán las partes fundamentales que conforman el módulo. Luego se explicarán las entradas, procesos, y salidas de cada componente.

Figura 3. Diagrama de componentes del módulo software.



Fuente: Creación propia.

Tal como se observa en el diagrama de la ***Figura 3***, el diseño del módulo está compuesto por ocho componentes principales, el componente de carga de *tweets*, el de preprocesamiento, el de desambiguación de sentidos, el de clasificación de los *tweets*, el de detección de sentimientos, el de exportación de *tweets* clasificados, el de transformación, y el de visualización. Las líneas del diagrama expresan la dependencia entre los componentes, de tal forma que, por ejemplo, para que exista el componente de clasificación deben existir previamente los componentes de preprocesamiento y desambiguación de sentidos.

5.1 Componente de Carga de Tweets

Este componente recibe como entrada un archivo de valores separados por comas (*Comma Separated Values*, *CSV*), el cual contiene los *tweets* y otra información asociada a ellos, como puede ser el identificador de cada *tweet*, y la ubicación desde donde se realizó la publicación del *tweet*. Todos estos datos se cargan en una estructura de datos tipo tabla, de tal forma que sea más sencillo trabajar con esta información desde memoria, y poder extraer de la tabla solo los datos necesarios.

5.2 Componente de Preprocesamiento

El componente de preprocesamiento es donde ingresa el *tweet* tal como fue cargado en el componente anterior, y se le realizan distintitos procesos para limpiarlos de cosas innecesarias, tales como caracteres especiales, tabulaciones, emoticones. Para esto, se usan funciones que realizarán estas tareas. Una función encargada de tomar el *tweet* y quitarle todos los emoticonos que pueda contener. Luego se utiliza una función que modifique los *hashtags* de la siguiente forma, si el *tweet* tiene un *hashtag* ***#HolaColombia***, luego de realizar el proceso el *tweet* quedaría, ***Hola Colombia***. También se utiliza una función que elimina las menciones a otros usuarios, de tal forma que se elimine del *tweet* todo lo que haya después de un símbolo de arroba (@), pero antes del siguiente espacio en blanco. Adicionalmente, se utiliza una función que elimina cualquier URL que haya en el tweet, así como cualquier carácter especial, tal como signos de interrogación, símbolos matemáticos, etc. Por último, este componente de preprocesamiento se encarga de identificar las Entidades Nombradas (*NER*) en el *tweet*, de tal forma que si en el texto se nombra a ***Coca-Cola***, la función la identifiqué. De tal manera que la salida de este componente va a ser un contexto con todas las palabras del *tweet*, y sin todos los elementos que fueron descritos anteriormente.

5.3 Componente de Desambiguación de Sentidos

La entrada de este componente es la salida del anterior, y el primer proceso que se realiza es la búsqueda de los diferentes sentidos que puede tener cada palabra del *tweet*. Se crea

una lista con cada uno de los sentidos por cada palabra, para luego recorrer la lista y por cada palabra recorrer el árbol se sentidos siguiendo las indicaciones del método de *Specification Marks* (*SM*). Una vez realizado el proceso para todas las palabras del *tweet*, existe menos ambigüedad en el *tweet*; sin embargo, si continúan palabras que no se pudieron desambiguar por este método principal, entonces se procede a aplicar una heurística de desambiguación por definición, la cual se encarga de contabilizar el número de palabras del contexto que existen dentro de la definición de cada palabra del *tweet*, con el fin de quedarse con la palabra que posee esa definición y romper así la ambigüedad. La salida de este componente es una lista con todas las palabras desambiguadas.

5.4 Componente de Clasificación de Tweets

Este componente recibe de entrada una lista de palabras desambiguadas, luego busca cada palabra dentro de los *eXtended WordNet Domains* (*XWND*), donde cada palabra tiene un puntaje asignado en cada dominio. Se realiza una sumatoria de ese puntaje para luego buscar la mayor sumatoria de todas, y de esa forma el dominio con el que se haya conseguido la mayor sumatoria de puntaje será es escogido para ser asignado al contexto del *tweet*, y por tanto a cada *tweet*. Una vez realizado todo el proceso, el componente retorna un objeto con el *tweet* y el dominio en el que fue clasificado.

5.5 Componente de Exportación de Tweets Clasificados

Este componente toma los resultados del componente anterior y los guarda con determinada estructura en un archivo *CSV* que se guarda en disco de manera persistente, de tal forma que se puedan acceder a los datos con los tweets clasificados en un dominio de forma sencilla desde cualquier software que pueda leer archivos con ese formato.

5.6 Componente de Detección de Sentimiento

Este componente tiene como entrada los *tweets* que serán analizados realizando diferentes procesos que determinan una polaridad para cada *tweet*. Siendo los posibles sentimientos: Positivo, neutro, y negativo, y las posibles emociones de alegría, miedo,

tristeza, enfado, asco, y otro. Lo anterior reflejaría si el contenido del *tweet* está en alguna de esas polaridades y emociones, dando información importante para analizar el *tweet*, y para complementar el proceso de clasificación en dominios. La salida de este componente son los *tweets* junto con su respectiva polaridad asociada en un archivo *CSV*.

5.7 Componente de Transformación

Este componente se encarga de recibir los resultados del componente de exportación de *tweets* clasificados, y del componente de detección de sentimiento. Luego de eso, se unifican ambos resultados para obtener un solo archivo con los respectivos datos asociados a cada *tweet*. Posteriormente el resultado se guarda en un archivo con formato *JSON* (*JavaScript Object Notation*), de tal forma que la estructura del archivo pueda ser más sencilla de reconocer por distintos lenguajes de programación y softwares.

5.8 Componente de Visualización

En este componente se ingresa el archivo con formato *JSON* resultante del componente anterior, y haciendo uso del lenguaje *JavaScript* se construye un mapa donde se muestran los *tweets* ubicados geográficamente desde donde fueron publicados, además en el mapa también se ubican distintos sitios turísticos de tal forma que se pueda ver la cercanía de estos con el lugar con la ubicación desde donde fue hecho el *tweet*. También, dentro de los marcadores que muestra el mapa se puede ver información relevante del *tweet*, tal como: Dominio, sitios cercanos, sentimiento, emoción, y si el sitio turístico cercano fue o no nombrado en el *tweet*.

6. Implementación de las rutinas necesarias para la clasificación de tweets

En este capítulo se describe el desarrollo de las rutinas necesarias para la clasificación de *tweets* en dominios conceptuales. Para lo cual se hace uso de *XWND* que proporciona los mencionados dominios conceptuales necesarios para el proceso de clasificación. Dentro de estos dominios, se encuentra el dominio del turismo, con lo cual se podrá determinar cuáles *tweets* pertenecen a ese dominio.

6.1 Rutina de Carga de Tweets

Esta rutina permite cargar todos los datos asociados a los *tweets* en una estructura de datos tipo tabla, donde cada columna corresponde a una característica del *tweet*, y cada fila es un *tweet* publicado. Dentro de estos datos están, el texto del *tweet*, el identificador del *tweet*, y la ubicación desde la que se publicó el *tweet*, entre otros.

La función recibe por parámetro el nombre del archivo con formato *CSV*, posteriormente lee los datos del archivo y los guarda en la estructura de datos `DatosDeTweets`. Luego se añaden dos nuevas columnas que servirán para guardar el dominio del *tweet*, y el puntaje que se obtuvo en ese dominio. Por último, la función retorna la estructura con los datos de los *tweets*.

```
1 función CargarTweets(NombreDelArchivo):
2
3     DatosDeTweets = LeerCSV(NombreDelArchivo)
4     DatosDeTweets.AñadirColumna('Dominio')
5     DatosDeTweets.AñadirColumna('Puntaje')
6     DatosDeTweets.AñadirColumna('Entidades')
7
8     retorna DatosDeTweets
```

6.2 Rutina de Preprocesamiento

En esta rutina se realiza toda la limpieza del texto del *tweet*. La función recibe como parámetros un *tweet* y un conjunto de palabras vacías (que son aquellas palabras que no tienen significado por sí solas, por ejemplo: por, para de, etc.). Dentro de la función se ejecuta la función `QuitarEmojis` la cual remueve todos los emoticones del texto, luego se ejecuta la función `ExtraerHashtags` que transforma los *hashtags* del texto tal como se describió en el capítulo anterior. La función `QuitarMenciones` remueve del texto todas las menciones a otros usuarios utilizando las reglas definidas en el capítulo anterior. `QuitarUrlsYCaracteresEspeciales` es una función que elimina del texto, todas las *URLs* que pueda contener, así como los caracteres especiales. La siguiente función es `LimpiarTexto` que se encarga de eliminar las palabras vacías y los espacios en blanco innecesarios en el texto, está función retorna un conjunto de palabras con sentido, llamadas contexto. La última función en ser ejecutada es `EncontrarEntidades` que se encarga de buscar y retornar las entidades nombradas en el *tweet*. Por último, se retorna el contexto del *tweet*, y las entidades que fueron encontradas en él.

```
1  función PreprocesamientoDeTweet(Tweet, PalabrasVacías):
2
3    Tweet = QuitarEmojis(Tweet)
4    Tweet = ExtraerHashtags(Tweet)
5    Tweet = QuitarMenciones(Tweet)
6    Tweet = QuitarUrlsYCaracteresEspeciales(Tweet)
7
8    Contexto = LimpiarTexto(Tweet, PalabrasVacías)
9    Entidades = EncontrarEntidades(Tweet)
11
12   retorna Contexto, Entidades
```

6.3 Rutina de Desambiguación de Sentidos

Esta rutina se encarga de desambiguar todos los sentidos de cada una de las palabras del contexto del *tweet*, para lo cual se hace uso del algoritmo de *Specification Marks* (*SM*), el cual es un algoritmo semántico, que se basa en la jerarquía de hiperónimo/hipónimo. Según la definición de la RAE un hiperónimo es aquella palabra cuyo significado está incluido en el de otras; por ejemplo, pájaro es hiperónimo de jilguero y de gorrión. Además,

un hipónimo es aquella palabra cuyo significado incluye el de otra, por ejemplo, gorrión es hipónimo de pájaro [21].

La función recibe como parámetro el contexto de un *tweet*, y se hace uso de la función `ConceptosDePalabras` para obtener todos los conceptos que puede tener una palabra dentro del contexto, posteriormente se recorre cada uno de los conceptos (los cuales tiene una palabra y un conjunto de sentidos asociado) para acceder a la palabra y los sentidos de cada uno.

```
1  función DesambiguarSentidos(Contexto):
2
3   Datos = ConceptosDePalabras(Contexto)
4   ListaDeSM = []
5   PalabrasAmbiguas = []
6
7   Para Conceptos en Datos:
8     Para cada Palabra, Sentidos en Conceptos:
9      Bandera = Verdadero
10
11    Mientras Bandera = Verdadero:
12      ListaTemporal = []
13      Para cada Sentido en Sentidos:
14        ListaTemporal.Añadir({Sentido.NombreSentido,  Sentido.Palabras,
  Sentido.Palabras.Longitud()})
16
17      Si ListaTemporal no es Vacía:
18       ListaDeLongitudes = ObtenerListaDeLongitudes(ListaTemporal)
19       Máximo = ObtenerMáximo(ListaDeLongitudes)
20       CantidadDeMáximos = ListaDeLongitudes.Contar(Máximo)
21
22        Si CantidadDeMáximos = 1:
23         Índice = ListaDeLongitudes.ObtenerÍndice(Máximo)
24         Si ListaTemporal[Índice].Longitud() > 0:
25           DatosDeHiperónimo = ListaTemporal[Índice]
26           Nombre = ListaTemporal[Índice].Sentido.NombreSentido
27           ListaDeSM.Añadir({Palabra:{Nombre:DatosDeHiperónimo}})
28           Bandera=Falso
29      Si no:
30        Si Sinónimos(Palabra) no es Vacía:
31          PalabrasAmbiguas.Añadir(Palabra)
32          Bandera=Falso
33   Si PalabrasAmbiguas:
34     ListaDeSM  =  HeuristicaDeDefinición(ListaDeSM,  PalabrasAmbiguas,
  contexto)
36   retorna ListaDeSM
```

Luego se realiza un ciclo que se ejecuta mientras la `Bandera` sea verdadera. Se recorren todos los sentidos de una palabra, y se guardan en una `ListaTemporal` el nombre del sentido, la lista de palabras asociadas a él, y la longitud de la lista de palabras. Se comprueba si `ListaTemporal` no es vacía, para luego escoger la lista con mayor número de palabras, luego se obtiene los datos del hiperónimo y del sentido, y se guardan en `ListaDeSM`. Si `ListaTemporal` está vacía, se buscan los sinónimos de la palabra para añadirla a la lista de `PalabrasAmbiguas`, y posteriormente ejecutar la función `HeuristicaDeDefinición` la cual recibe todas las palabras que no pudieron ser desambiguadas en el método principal. Por último, se retorna la `ListaDeSM`.

6.4 Rutina de Clasificación de Tweets

La rutina de clasificación de *tweets* recibe como parámetro los datos de los tweets que retorna la función de carga de *tweets*. Para cada *tweet*, se extrae el texto y se le aplica la función `PreprocesamientoDeTweet` que fue descrita anteriormente, cuyos contextos retornados son guardaos en la lista `Contextos` para su posterior acceso. Una vez todos los *tweets* han sido preprocesados, se procede a recorrer los contextos limpios para aplicarles la desambiguación mediante la función `DesambiguarSentidos`.

Una vez se han desambiguado los contextos, se procede a obtener el dominio de cada uno de los *tweets*, así como el puntaje que se obtuvo en ese dominio, haciendo uso de la función `ObtenerDominio`. Esta función busca las palabras de cada contexto dentro de los dominios de *XWND* y obtiene un puntaje para cada palabra en un determinado dominio, estos puntajes se suman para obtener el puntaje del contexto en ese dominio (este proceso es realizado para cada contexto en todos los dominios de *XWND*). Una vez obtenidos todos los puntajes para cada contexto en cada dominio, se comparan todos los puntajes obtenidos y se elige el puntaje mayor (lo que significa que el contexto del *tweet* pertenece a ese dominio). Una vez hecha la clasificación de los *tweets*, se guardan los valores obtenidos en las listas `Dominios` y `Puntajes`, para posteriormente guardar los valores en las columnas de Dominio y Puntaje de `DatosDeTweets`, y finalmente retornar esta estructura de datos.

```
1  función ClasificaciónDeTweets(DatosDeTweets):
2
3    Contextos = []
4    ListaDeEntidaes = []
5    Para cada Tweet en DatosDeTweets.Texto:
6      Contexto, Entidades = PreprocesamientoDeTweet(Tweet,
   PalabrasVacías)
8      Contextos.Añadir(Contexto)
9      ListaDeEntidades.Añadir(Entidades)
10
11   Dominios = []
12   Puntajes = []
13   Para cada Contexto en Contextos:
14     SM = DesambiguarSentidos(Contexto)
15
16     Si SM no es Vacía:
17       Clasificados = ObtenerDominio(SM)
18       Dominios.Añadir(Clasificados.Dominios)
19       Puntajes.Añadir(Clasificados.Puntajes)
20     Si no:
21       Dominios.Añadir('Indefinido')
22       Puntajes.Añadir(0)
23
24   DatosDeTweets.Dominio = Dominios
25   DatosDeTweets.Puntaje = Puntajes
26   DatosDeTweets.Entidades = ListaDeEntidades
27
28   retorna DatosDeTweets
```

6.5 Rutina de Exportación de Tweets

Esta rutina recibe los datos de los *tweets* retornados por la función de clasificación, y un nombre de archivo. La función de exportación ejecuta una función que crea un archivo *CSV* con el nombre pasado por parámetro, y guarda todos los datos de los *tweets* en ese archivo. De tal forma que los *tweets* clasificados quedan guardados de forma persistente para su posterior consulta, e integración con otros módulos.

```
1 función ExportarTweets(DatosDeTweets, NombreArchivo):
2   EscribirCSV(DatosDeTweets, NombreArchivo)
```

6.6 Rutina de Transformación

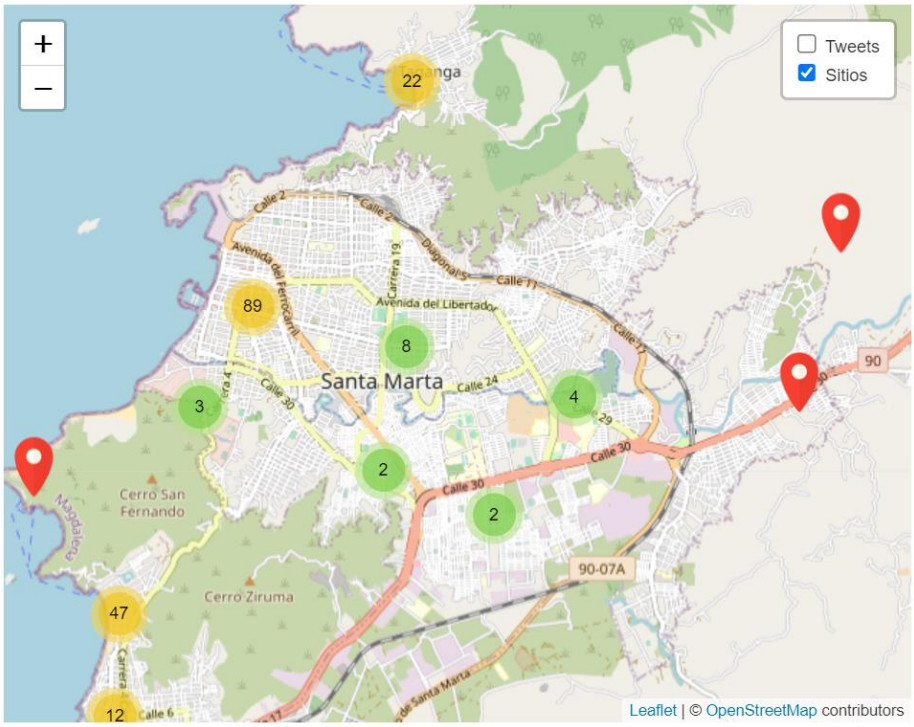
Esta rutina recibe dos archivos con formato *CSV*, uno con los datos de los *tweets* clasificados, y otro con los *tweets* asociados con un sentimiento y una emoción. Lee los datos de los archivos, se guarda el contenido de cada uno en una estructura de datos, para luego añadir el sentimiento y la emoción al conjunto de *tweets* clasificados. Por último, se guarda la estructura de datos combinada en un archivo con formato *JSON* para poder utilizar de forma más eficiente los datos de los *tweets* con el lenguaje de programación *JavaScript* en el componente de visualización.

```
1  función UnificarDominioYSentimiento(ClasificadosCSV, SentimientosCSV):
2    TweetsClasificados = LeerCSV(ClasificadosCSV)
3    TweetsSentimientos = LeerCSV(SentimientosCSV)
4
5    Para Índice = 0 hasta TweetsClasificados.Longitud():
6      TweetsClasificados[Índice].AñadirColumna('sentimiento',
   TweetsSentimientos[Índice].sentimiento)
8      TweetsClasificados[Índice].AñadirColumna('emoción',
   TweetsSentimientos[Índice].emoción)
10
11   EscribirJSON(TweetsClasificados)
```

7. Construcción de las visualizaciones

Para el componente de visualización se usó el lenguaje de programación *JavaScript* para construir un mapa, en el que se ubicaron marcadores que hacen referencia a la posición desde la que se publicaron los *tweets*. Esto fue posible gracias a que dentro de la información de los *tweets* se encontraba tanto la longitud como la latitud desde la que se publicó cada uno de los *tweets*. Además de ubicar los *tweets* en el mapa, también se ubicaron sitios relevantes (lugares turísticos, históricos, hoteles, restaurantes, etc.) con el fin de determinar si el *tweet* se realizó cerca de alguno de esos sitios.

Figura 4. Mapa con la capa de los sitios relevantes visible.

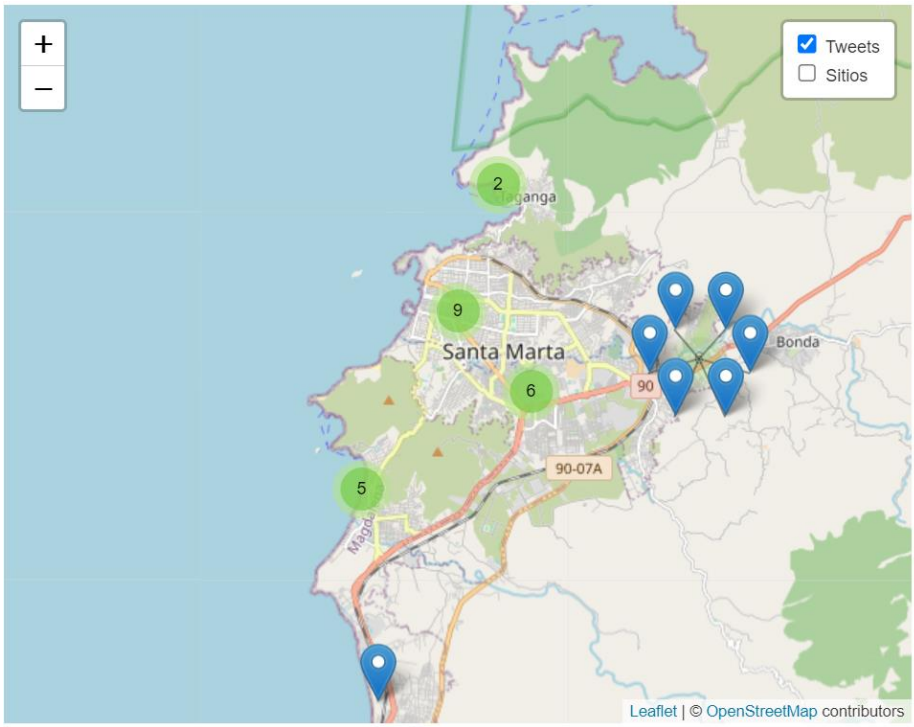


Fuente: Creación propia.

Para la visualización del mapa se utilizaron capas, donde una de las capas del mapa muestra todos los sitios relevantes, y otra capa muestra los *tweets*. Ambas capas se

pueden visualizar de forma independiente o juntas, de tal forma que se puedan ver los sitios que hay cerca de la ubicación de cada *tweet*. Los marcadores que muestran el punto exacto de los sitios son de color rojo (ver ***Figura 4***); mientras que los marcadores que muestran la ubicación de los *tweets* son de color azul (ver ***Figura 5***), con el fin de que sea más fácil distinguirlos con un vistazo rápido. En la ***Figura 6*** se observan ambas capas visibles en el mapa y se evidencia como es fácilmente distinguible un sitio de un *tweet*, además de poder ver los sitios que se encuentran cercanos a los puntos desde donde se realizaron los *tweets*.

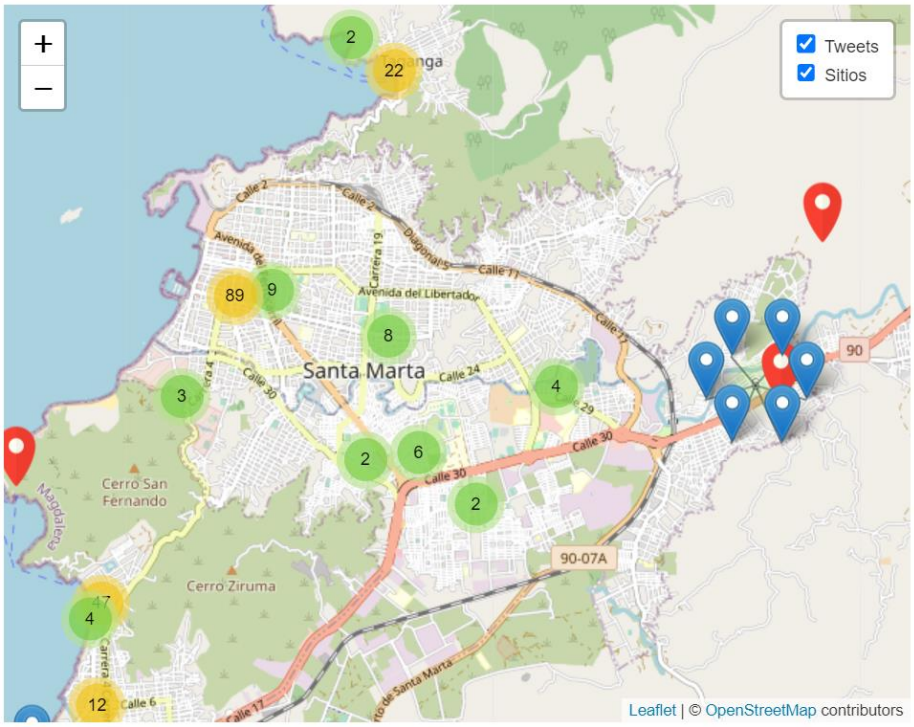
Figura 5. Mapa con la capa de tweets visible.



Fuente: Creación propia.

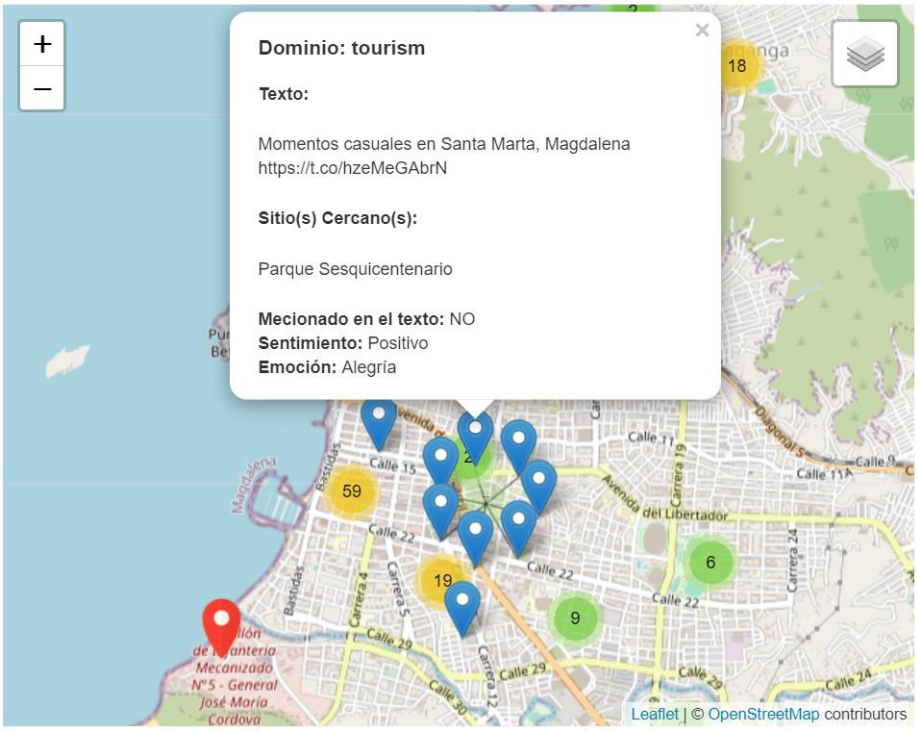
Además, en la ***Figura 7*** se observa que cada marcador azul contiene la información relacionada al tweet, tal como el dominio al que pertenece, el texto, los sitios cercanos, si el sitio es o no mencionado en el texto, además del sentimiento y la emoción que se identificaron.

Figura 6. Mapa con las capas de sitios y tweets visible.



Fuente: Creación propia.

Figura 7. Mapa con la información que contiene cada marcador.



Fuente: Creación propia.

8. Conclusiones

El sector del turismo es fundamental para la ciudad de Santa Marta, así como para el departamento Magdalena, contribuyendo en gran medida a la economía de la región, y teniendo un gran potencial de crecimiento para los próximos años. Por lo cual, cualquier desarrollo que contribuya al potenciamiento de este sector (en la toma de decisiones, en mejorar el acceso a la información, en mostrar información relevante, etc.) tiene un alto valor para los actores del sector tales como turistas, comerciantes, inversores, funcionarios públicos, etc. El desarrollo que se realizó y que fue descrito en este trabajo, es un aporte al mejoramiento del sector turismo mediante el procesamiento, clasificación, y presentación de datos obtenidos de la red social *Twitter*, con el fin de que sea de utilidad para la comunidad en general.

Hacer la revisión de la literatura sobre técnicas y conceptos de PLN fue fundamental para comprender la mejor forma de abordar las demás fases del proyecto. Gracias a la revisión literaria se pudo realizar de una forma más eficiente el inventario de las herramientas de utilidad para realizar procesos de PLN, de las cuales se eligieron las más idóneas para ser utilizadas en la fase de implementación. Una vez elegidas las mejores herramientas para realizar PLN, se procedió a realizar un diseño con cada uno de los componentes del módulo, y a describir su funcionamiento para tener una idea mucho más clara del comportamiento que debería tener el módulo y los procesos de este. A partir de ese diseño, la fase de implementación fue realizada de forma guiada, codificando las funciones necesarias para la construcción de cada uno de los componentes, en los cuales se aplicaron técnicas para cargar los datos de los *tweets*, también técnicas para el preprocesamiento, detección de entidades nombradas, y clasificación de los *tweets* en dominios conceptuales.

Teniendo como resultado a los tweets clasificados y unificados con los resultados del módulo de análisis de sentimiento, se construyeron visualizaciones en las cuales se

ubicaron los tweets en el lugar donde fueron realizados, para determinar si estaban cerca de algún sitio relevante (hotel, restaurante, playa, etc.). Gracias a todo lo anteriormente descrito se pudo determinar que muchos de los *tweets* fueron realizados cerca de algún lugar relevante de la ciudad de Santa Marta, y en algunos casos estos sitios eran nombrados en el contenido del *tweet*.

En cuanto al impacto que podría tener el desarrollo de una plataforma para el análisis de datos para un turismo 2.0, se pueden destacar muchos usos para este tipo de plataformas, como por ejemplo que los prestadores de servicios turísticos la utilicen para dirigir de mejor manera sus catálogos de productos y servicios con el fin de llegar al mayor número de usuarios de manera efectiva. La plataforma también podría ser usada por los propios turistas para determinar cuáles destinos son los que más les convienen. Otros entes que podrían usar la plataforma serían los gobiernos distritales y departamentales, ya que podrían utilizar la información proporcionada para tomar decisiones o medidas políticas respecto de, por ejemplo, ciertos destinos o lugares turísticos que tengan problemas de seguridad y que sean manifestados por los turistas; o para sancionar lugares como restaurantes u hoteles que incumplan con medidas de sanidad. Los entes nacionales también podrían hacer uso de estos datos, para determinar en qué vías hay más afluencia de viajeros con el fin de tomar decisiones referentes a la movilidad y al mejoramiento de las carreteras. Esta plataforma también podría ser usada por inversores que deseen saber cómo está el panorama turístico en determinado subsector, y puedan tomar decisiones informadas de inversión.

En cuanto al planteamiento de mejoras para trabajos futuros, se puede plantear la utilización de más algoritmos heurísticos en el proceso de desambiguación de sentidos después de aplicar el algoritmo *SM*, de tal forma que el contexto del *tweet* contenga los significados más precisos posibles, y cuando se realice el proceso de clasificación se obtengan mejoras en el resultado.

También se puede plantear el enriquecimiento del contexto del *tweet* mediante la búsqueda de más características semánticas relevantes utilizando algún recurso especializado, de tal forma que incluso el proceso de desambiguación de sentidos funcione de una mejor forma pues los sentidos serían más precisos, y por tanto las definiciones de cada palabra serían más adecuadas, favoreciendo a los resultados del proceso de clasificación.

Bibliografía

[1] Ministerio del comercio industria y turismo, “Estadísticas Nacionales - Flujo de Turistas - Turismo Receptor.” https://www.citur.gov.co/estadisticas/df_viajeros_ciudad_destino/num_viajeros/2?t=1#gsc.tab=0

[2] Ministerio de comercio industria y turismo, “Estadísticas Departamentales.” https://www.citur.gov.co/estadisticas/departamental/index/47#gsc.tab=0

[3] Sistema e información turística del Magdalena y Santa Marta, “Motivo principal de un viaje turístico,” 2019. https://www.siturmagdalena.com/indicadores/receptor

[4] K. M. Castellar Melo and D. C. Polo Rosales, “INCIDENCIA DEL TURISMO EN EL DESARROLLO SOCIOECONOMICO DE SANTA MARTA, DURANTE EL PERIODO 2010 A 2016.,” Santa Marta, 2017. [Online]. Available: https://repository.ucc.edu.co/bitstream/20.500.12494/13712/1/2017_incidencia_turismo_desarrollo.pdf

[5] Alcaldía de Santa Marta, “Componente Estratégico Plan de Desarrollo ‘Santa Marta Corazón del Cambio’ 2020-2023,” *2020*. https://www.santamarta.gov.co/portal/archivos/documentos/transparencia/2020/PDD/COMPONENTE_ESTRAT%C3%89GICO_PDD_2020-2023.pdf

[6] C. Batista Gondim, A. L. de Castro Seabra, and L. Mendes-Filho, “EMPODERAMIENTO PSICOLÓGICO DE LOS ANFITRIONES DE AIRBNB EN BRASIL A TRAVÉS DE LAS COMUNIDADES EN LAS REDES SOCIALES,” *Estudios y perspectivas en turismo*, vol. 29, no. 2, pp. 349–368, Apr. 2020, [Online]. Available: http://www.scielo.org.ar/scielo.php?script=sci_arttext&pid=S1851-17322020000200349&lang=es

[7] A. Falcão Durão, A. Jacinto dos Santos, M. Raquel Avelino, and C. Borba da Mota Silveira, “COMIENDO VIRTUALMENTE CON LOS OJOS Un estudio sobre el uso de Instagram por parte de los prestadores de servicios turísticos de gastronomía de Recife (Brasil),” *Estudios y perspectivas en turismo*, vol. 26, no. 4, pp. 964–977,

Sep. 2017, [Online]. Available: http://www.scielo.org.ar/scielo.php?script=sci_arttext&pid=S1851-17322017000400011&lang=es

[8] M. Á. Sánchez Jiménez, “Análisis de la estrategia en las redes sociales oficiales desarrollada por el Consejo de Promoción Turística de México,” *Cimexus*, vol. 13, no. 1, pp. 13–50, 2018.

[9] R. Bellman, *An introduction to artificial intelligence : can computers think?* San Fransisco: Boyd & Fraser Publishing Company, 1978.

[10] S. Russell and P. Norvig, “¿Qué es la IA?,” in *INTELIGENCIA ARTIFICIAL UN ENFOQUE MODERNO*, 2nd ed., pp. 2–6.

[11] A. Gelbukh, “Procesamiento de lenguaje natural y sus aplicaciones,” *Komputer Sapiens*, vol. 1, pp. 6–11, 2010, [Online]. Available: https://d1wqtxts1xzle7.cloudfront.net/30768432/Procesamiento_de_lenguaje_natural_y_sus_aplicaciones-with-cover-page-v2.pdf?Expires=1635909525&Signature=D-m-TErFa-buYARh0O~BN46ehqHl~bH5dGUruvPIJs5O4TyAst8KuKMhWYG1ThadlalxX2HLFGh3vlkA1GtSpqpsSlWRyito5ZIKWR9vbtkLgezvZ6rIbmjLC~eDGDmKWAWXLzSQZy6PLgBpSWMpD4FSnOilnsxpRPfm3NwbZG6YcwEYCfcblw6WPiVwq2k6QkHHOadj0T4OR8SogWHXOJQslMH59LF~8W6scZf2nAgaj3gfv466AHsuOrOYeoooqhZVrx-SBPx0v9zkAgjFR5VOm~SCo9fGSxh06ZEwQBkuRVtg3yKPyW1-B54cKJYDj4Qky8MI9-5~gNp9WfDOIA__&Key-Pair-Id=APKAJLOHF5GGSLRBV4ZA

[12] C. Russo, H. Ramón, N. Alonso, B. Cicerchia, L. Esnaola, and J. P. Tessore, “Tratamiento Masivo de Datos Utilizando Técnicas de Machine Learning”, [Online]. Available: https://repositorio.unnoba.edu.ar/xmlui/bitstream/handle/23601/107/1_resource.pdf?sequence=1&isAllowed=y

[13] L. García Brime, “TURISMO 2.0: UNA REVOLUCIÓN EN LA FORMA DE VIAJAR.,” León, 2014. [Online]. Available:

https://buleria.unileon.es/bitstream/handle/10612/4136/45688635D_GADE_septiembre2014.pdf

[14] M. Borja Quevedo, “Análisis de las Herramientas de Procesamiento de Lenguaje Natural para estructurar textos médicos,” Donostia - San Sebastián, 2020. [Online]. Available: https://dadun.unav.edu/bitstream/10171/60003/1/Quevedo%20Marcos_Borja_904447_MII.pdf

[15] Princeton University, “About WordNet,” *WordNet*, 2010. https://wordnet.princeton.edu/

[16] University of Amsterdam, “Welcome to EuroWordNet,” *EuroWordNet*, 2001. https://archive.illc.uva.nl//EuroWordNet/

[17] A. Ritter, Mausam, S. Clark, and O. Etzioni, “Open Domain Event Extraction from Twitter,” Association for Computing Machinery, 2012, pp. 1104–1112. doi: 10.1145/2339530.2339704.

[18] B. L. Ibtihel, H. Lobna, and B. J. Maher, “A Semantic Approach for Tweet Categorization,” *Procedia Comput Sci*, vol. 126, pp. 335–344, 2018, doi: 10.1016/j.procs.2018.07.267.

[19] A. Montoyo, A. Suárez, G. Rigau, and M. Palomar, “Combining Knowledge-and Corpus-based Word-Sense-Disambiguation Methods,” 2005. [Online]. Available: http://www.cogsci.princeton.edu/

[20] G. Rigau, J. Atserias, and E. Agirre, “Combining unsupervised lexical knowledge methods for word sense disambiguation,” 1997, pp. 48–55. doi: 10.3115/976909.979624.

[21] *Diccionario de la lengua española*, 23rd ed. REAL ACADEMIA ESPAÑOLA, 2021. [Online]. Available: https://dle.rae.es