



UNIVERSIDAD NACIONAL DE INGENIERÍA

FACULTAD DE CIENCIAS

ESCUELA PROFESIONAL DE CIENCIA DE LA COMPUTACIÓN

*Web Scraping y Lenguaje de Procesamiento
Natural en la polaridad de opinión de usuarios en
medios digitales y redes sociales*

SEMINARIO DE TESIS I

Autor: Kevin Daniel Molina Bejar

Asesor: Cesar Jesus Lara Avila

Diciembre, 2018

Resumen

En los últimos años las redes sociales juegan un rol muy importante en nuestra sociedad debido al tiempo que se les dedica interactuando entre la abundante información en tiempo real sobre diversos temas que allí se encuentran . En ese contexto buscamos como los medios de comunicación, a través de sus plataformas digitales, se relacionan con la polaridad en las opiniones de los usuarios en medios digitales como la de el Grupo El Comercio y redes sociales como Twitter en la ciudad de Lima-Perú.

Índice general

Resumen	III
1. Introducción	1
2. Estado del Arte	5
2.1. Twitter y la teoría de la Agenda-Setting: mensajes de la opinión pública digital [1]	5
2.2. A comparative study on different types of approaches to text categorization [2]	7
2.3. Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN)	8
2.3.1. Taller de análisis semántico del SEPLN [3]	8
2.4. Twitter Sentiment Analysis System [4]	9
2.5. Predictive Analysis on Twitter: Techniques and Applications [5] .	9
2.6. Focused Crawling Through Reinforcement Learning [6]	10
3. Marco Teórico y Metodología Aplicada	12
3.1. Web scraping	12
3.1.1. Librerías usadas en web scraping	13
3.2. Support vector machine [7]	14
3.2.1. Linear support vector classification	15
3.3. API Twitter [8]	15
3.3.1. Tweepy [9]	17
3.4. Análisis de Sentimiento	17
3.4.1. Natural Language Toolkit	18
3.4.2. Data recopilada	19

3.4.3. Data procesada	20
3.4.4. Modelo desarrollado	21
3.4.5. Data de entrenamiento	22
3.5. Base de datos	24
3.6. Gráficos	24
3.6.1. Matplotlib [10]	24
3.6.2. Heatmap.py	25
3.7. Amazon web services [11]	26
3.7.1. Elastic Compute Cloud	26
4. Evaluación y Resultados del Seminario	30
4.1. Recolección de noticias de medios digitales	30
4.2. Palabras claves o con mayor frecuencia usada por los medios digitales	32
4.3. Recolección de tweets usando la API de twitter	34
4.4. Modelo usado	36
4.4.1. Pre-procesamiento de datos	37
Tokenizar	37
Lematizar	37
CountVectorizer	37
Cambiar a minúsculas	37
Remover stopwords	37
4.4.2. Búsqueda de hiperparametros	38
4.5. Predicción de la polaridad de un tweet	38
4.6. Otras aplicaciones de Web Scraping	41
5. Conclusiones y Trabajos Futuros	47
5.1. Conclusiones	47
5.2. Trabajo Futuro	48

Índice de figuras

1.1. Audiencia web. Fuente: IAB-Comscore(2015)	2
2.1. Número de noticias de El País y El Mundo que recogen los trendings topics de Twitter España de marzo de 2013. Fuente: [1]	7
3.1. Dataset Iris. Fuente: https://scikit-learn.org/	15
3.2. Tipos de cuenta en la Api de Twitter. Fuente: https://developer.twitter.com/en/pricing	16
3.3. Canales RSS. Fuente: https://elcomercio.pe/rss/	20
3.4. Curva ROC. Fuente: Elaboración propia	21
3.5. Ejemplos simples de matplotlib. Fuente: https://matplotlib.org/	25
3.6. Ejemplo mapa de calor. Fuente: http://www.sethoscope.net/heatmap/	26
4.1. Links de las noticias publicadas. Fuente: Elaboración propia	31
4.2. Texto de noticias publicadas. Fuente: Elaboración propia	32
4.3. Palabras y su frecuencia para el dia 4/11/18. Fuente: Elaboración propia	32
4.4. Noticias del dia 4/10/18 con relacion a la palabra congreso. Fuente: Elaboración propia	33
4.5. Diseño primera parte del proyecto. Fuente: Elaboración propia	34
4.6. Atributos capturados. Fuente: Elaboración propia	35
4.7. Diseño segunda parte del proyecto. Fuente: Elaboración propia	36
4.8. Ejemplo de data limpia para dos tweets. Fuente: Elaboración propia	38
4.9. Mapa de calor de los tweets recolectados en Lima usando sus coordenadas. Fuente: Elaboración propia	39

4.10. Número de tweets por día. Fuente: Elaboración propia	40
4.11. Número de tweets por día. Fuente: Elaboración propia	40
4.12. Mapa de calor de los tweets recolectados en Lima con su polaridad. Fuente: Elaboración propia	41
4.13. Busqueda de alumnos y egresados. Fuente: http://www.orce.uni.edu.pe/buscaalu.php?op=buscaalu	42
4.14. Web scraping hecho en python. Fuente: Elaboración propia	43
4.15. Consulta miembro de mesa. Fuente: https://consultamiembrodemesa.onpe.gob.pe/	44
4.16. Web scraping a la página de ONPE. Fuente: Elaboración propia	44
4.17. Menor número dni vigente. Fuente: https://consultamiembrodemesa.onpe.gob.pe/	45

Índice de Acrónimos

API	Application Programming Interface
AUC	Area Under the Curve
AWS	Amazon Web Services
EC2	Elastic Compute Cloud
JSON	Javascript Object Notation
NLP	Natural Language Processing
NLTK	Natural Language Tool Kit
OSM	Open Street Map
RDBMS	Relational Data Base Management System
ROC	Receiver Operating Characteristic
RSS	Really Simple Syndication
SEPLN	Sociedad Española para el Procesamiento de Lenguaje Natural
SQL	Structured Query Language
SVM	Support Vector Machine
TASS	Taller de Análisis Semántico del Sepln
XML	eXtensible Markup Language

Capítulo 1

Introducción

Hoy en día la web en general almacena muchos datos privados como datos públicos los cuales se deben de tener un cuidado dedicado ya que estos datos podrían ser extraídos de forma automatizada por un script de manera eficiente y consumiendo pocos recursos.

Nuestros conocimientos adquiridos nos permite pasar estos datos de la web a formato legible para darles uso a nuestra conveniencia para distintos estudios o proponer distintas soluciones a diversos problemas.

A esto se le llama web scraping, programas que explotan la estructura de grafo de la web para moverse de una página a otra, muy interesante por su versatilidad para este proyecto y que decidimos utilizar en la práctica para extraer las noticias del medio digital más consumido en nuestra sociedad como lo es “Grupo El Comercio”. Elegimos este medio digital por ser el que posee la mayor concertación de medios desde el año 2015 (ver figura 1.1) y buscaremos alguna relación con las polaridades de la opinión pública de forma binaria (positivo o negativo).

Puede que este caso de estudio, sobre la influencia de los medios de comunicación en la opinión pública, no pertenezca a nuestra área pero podemos probar como nuestra especialidad puede adaptarse, aportando herramientas computacionales, a cualquier área útil a la sociedad.

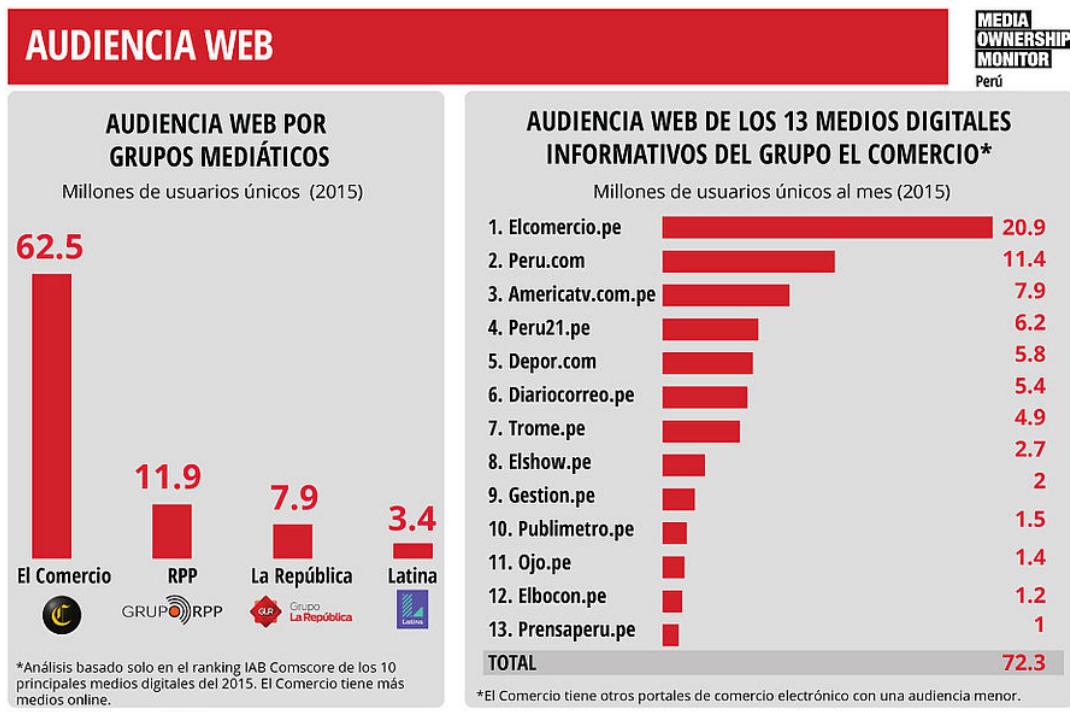


FIGURA 1.1: Audiencia web. Fuente: IAB-Comscore(2015)

Para brindar al lector una idea global del contenido de este trabajo, se hace una breve descripción del propósito de cada capítulo presente en este seminario de tesis.

■ Introducción:

En este capítulo introductorio se comenta la idea general del seminario indicando las técnicas sobre extracción de datos denominado web scraping y análisis de sentimiento.

■ Trabajos Relacionados y Estado del Arte:

En este capítulo analizamos trabajos similares que han sido considerados como referencia o precedentes, incluso trabajos que a pesar de ser obsoletos por el año de publicación han definido conceptos muy útiles. Luego se realizará algunos comentarios que contrastan estos trabajos valorados y la propuesta dada en este seminario. Además se mencionarán algunos otros artículos que utilizan técnicas distintas a lo que se plantea.

■ Metodología aplicada:

En este capítulo describimos en detalle los métodos y los recursos utilizados realizamos para este seminario.

■ Evaluación del proyecto:

Este capítulo contiene información y descripción de las tecnologías involucradas en el funcionamiento, desempeño y los pasos realizados en el seminario.

■ Conclusiones y Trabajos a Futuro:

En este capítulo se exponen las conclusiones obtenidas de este seminario. Adicionalmente, se proponen trabajos a futuro para cubrir áreas no consideradas o estudiar nuevas alternativas no contempladas como es el que caso del aprendizaje por Refuerzo.

Capítulo 2

Estado del Arte

Durante la etapa de investigación nos encontramos con muchos papers relacionados al tema, exponemos los siguientes que nos parecieron más interesantes y relevantes para nuestro caso de estudio.

2.1. Twitter y la teoría de la Agenda-Setting: mensajes de la opinión pública digital [1]

La teoría de la Agenda-Setting fue formalmente desarrollada en 1972 por el sociólogo estadounidense Maxwell McCombs junto a Donald Shaw a partir de los trabajos previos del periodista Walter Lippmann (1922). Ésta teoría se basa en que los medios de comunicación, al seleccionar los temas que incluyen u omiten en su agenda, ejercen gran influencia sobre el público y sobre su capacidad para opinar y debatir los asuntos públicos, ya que determinan los temas de interés informativo, su importancia y el espacio que se dedica a cada uno de estos temas.

Se busca comprobar si la teoría de la Agenda-Setting mantiene su vigencia dentro del creciente entorno digital, concretamente dentro del contexto de las redes sociales como lo es Twitter en España.

Para esto se relaciona los trending topics de twitter en España con lo publicado en medios digitales como El País y El Mundo, los dos periódicos de información general más leídos en dicho país.

Utilizan una web que brinda información de los trending topics de twitter y así contabilizan los mismos para luego relacionar con lo publicado en los medios digitales.

De este estudio se desprenden dos aspectos importantes:

- La primera es que la opinión pública reflejada en la red social Twitter ha demostrado componer su agenda temática en torno a dos líneas distintas. Ha quedado probado que los usuarios están interesados tanto en la realidad más actual, aquella que suelen recoger los medios de comunicación, como en otro tipo de cuestiones más ligadas a la propia naturaleza de Twitter y al entretenimiento.

Dentro de los asuntos ligados a la información de actualidad, podemos destacar que los internautas han otorgado a las cuestiones relativas al mundo del deporte mayor importancia que al resto. Los asuntos deportivos han ocupado casi el 30% de la agenda pública, lo que evidencia un interés de los usuarios por todo tipo de deportes, entre los cuales ha predominado el fútbol.

- La segunda es que existe una clara correspondencia entre la agenda del público y la establecida por los medios de comunicación tradicionales. Dejando al margen la parte de la agenda pública referida a los temas originarios de Twitter, hemos observado que en lo que respecta a las temáticas más informativas, los asuntos comentados por los usuarios de Twitter en la red social han sido en su mayoría abordados también por los medios.

De esta forma, algunas cuestiones como los casos de corrupción en España, la muerte de Hugo Chávez, el cónclave para elegir nuevo Papa o el rescate de Chipre han protagonizado un interés mediático que los usuarios han recogido en Twitter en una proporción muy similar estos ejemplos resaltan otra observación importante y es que el nivel de influencia de los medios sobre la agenda pública depende y varía en función de los temas.

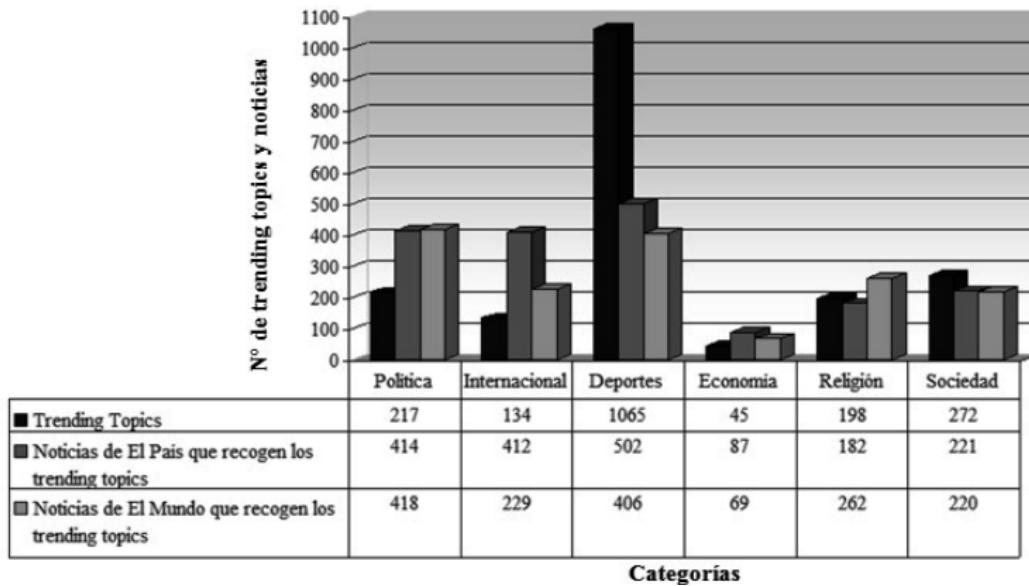


FIGURA 2.1: Número de noticias de El País y El Mundo que recogen los trending topics de Twitter España de marzo de 2013.

Fuente: [1]

2.2. A comparative study on different types of approaches to text categorization [2]

Nos centramos en este estudio para definir el modelo que usaremos en el seminario ya que aquí prueban distintos modelos para la clasificación de texto.

Este estudio busca encontrar una comparativa entre distintos modelos supervisados, no supervisados y semi supervisados para la clasificación de texto como son K-vecinos cercanos, Naive bayes, SVM, Árboles de Decisión, Backpropagation y algoritmo de Rocchio.

Actualmente, la investigación de categorización de texto está investigando las propiedades de escalabilidad de los sistemas de clasificación de texto, es decir, comprender si los sistemas que han demostrado ser los mejores en términos de efectividad sólo se enfrentan al desafío de tratar con un gran número de categorías. Se propusieron varios algoritmos o combinaciones de algoritmos como enfoques híbridos para la clasificación automática de documentos.

Entre estos algoritmos: SVM, Naive Bayes, k-vecinos cercanos y su sistema híbrido con la combinación de otros algoritmos y técnicas de selección de características se muestran más apropiados en la literatura existente.

Después de una revisión de los diferentes tipos de enfoques y comparar los métodos existentes basados con diversos parámetros, se concluye que el clasificador SVM ha sido reconocido como uno de los métodos de clasificación de texto más efectivos en las comparaciones de algoritmos de aprendizaje automático supervisado.

2.3. Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN)

El SEPLN es una asociación científica sin ánimo de lucro con el objetivo de promover todo tipo de actividades relacionadas con el estudio del procesamiento de lenguaje natural.

Para cumplir este objetivo cada año realiza un workshop incentivando el desarrollo del NLP en español y también atraen el interés de la comunidad de investigación con la clasificación de posturas, el manejo de la negación, la identificación de rumores, la identificación de noticias falsas, la extracción de información abierta, la extracción de argumentos, la clasificación de relaciones semánticas y cada vez sigue creciendo la lista de aplicaciones.

2.3.1. Taller de análisis semántico del SEPLN [3]

El TASS se ha llevado a cabo desde 2012, en el marco de la Conferencia Internacional de la Sociedad Española de Procesamiento del Lenguaje Natural (SEPLN). TASS fue la primera tarea compartida en el análisis de sentimientos en Twitter en español. El español es el segundo idioma más utilizado en Facebook y Twitter, que exige el desarrollo y la disponibilidad de métodos y recursos específicos del idioma para el análisis de sentimientos. El objetivo inicial de

TASS fue el avance de la investigación sobre el análisis de sentimientos en español con un interés especial en el idioma utilizado en Twitter.

Aunque el análisis de sentimientos sigue siendo un problema abierto, el Comité de Organización desea fomentar la investigación sobre otras tareas relacionadas con el procesamiento de la semántica de textos escritos en español.

Dicho Comité de Organización apela a la comunidad de investigación para proponer y organizar tareas de evaluación relacionadas con otras tareas semánticas en el idioma español. Las nuevas tareas brindan la oportunidad de crear recursos lingüísticos, evaluar su utilidad y promover la consolidación de una comunidad de investigadores interesados en los temas abordados. Por lo tanto, se alienta a la comunidad de procesamiento semántico a proponer y enviar una tarea de evaluación cada año.

2.4. Twitter Sentiment Analysis System [4]

En este documento se clasifica los sentimientos en Twitter con la ayuda de los algoritmos de Machine Learning y Lenguaje de Procesamiento Natural (NLP), utilizando conjuntos de datos de Kaggle que se rastrearon desde Internet y se etiquetaron como positivos o negativos. Los datos proporcionados vienen con emoticones, nombres de usuario y hashtags que deben procesarse (para que sean legibles) y se conviertan en una forma estándar de procesar.

En el documento se extraen características útiles de texto, como los unigramas y los bigramas, que es una forma de representación del "tweet". Utilizan varios algoritmos de Machine Learning basados en NLP para realizar análisis de sentimientos utilizando las funciones extraídas.

2.5. Predictive Analysis on Twitter: Techniques and Applications [5]

En este documento se estudian las técnicas, enfoques y aplicaciones de vanguardia del análisis predictivo de datos de Twitter. Específicamente, se

presenta un análisis detallado que involucra aspectos como el sentimiento, la emoción y el uso del conocimiento del dominio en el análisis general de los datos de Twitter para tomar decisiones y tomar acciones, y relatar algunas historias de éxito.

El análisis predictivo de los datos de las redes sociales, que se desarrolla en este trabajo tiene una considerable comunidad de investigación y del mundo empresarial debido a la información esencial y procesable que puede proporcionar. A lo largo de los años, se han llevado a cabo una extensa experimentación y análisis para obtener información sobre el uso de los datos de Twitter en varios dominios, como salud, salud pública, política, ciencias sociales y demografía.

2.6. Focused Crawling Through Reinforcement Learning [6]

En el documento, se propone un enfoque basado en el aprendizaje por refuerzo. El algoritmo que los autores plantean evalúan los hipervínculos más rentables a seguir a largo plazo y selecciona el enlace más prometedor basado en esta estimación.

Para modelar correctamente el entorno de crawling como un proceso de decisión de Markov (El framework matemático para definir una solución en el escenario de aprendizaje por refuerzo), se proponen nuevas representaciones de estados y acciones teniendo en cuenta tanto la información del contenido como la estructura del enlace. El tamaño del espacio de acción de estado se reduce por un proceso de generalización.

Los autores basándose en esa generalización, usan una aproximación de función lineal para actualizar funciones de valor. Se investiga la compensación entre los métodos síncronos y asíncronos y se compara el rendimiento de una tarea de crawling con y sin aprendizaje. Como resultado principal se muestra que los crawler basados en aprendizaje por refuerzo muestran un mejor rendimiento para diversos temas específicos.

Capítulo 3

Marco Teórico y Metodología Aplicada

3.1. Web scraping

La extracción de información se utiliza para motores de búsqueda, bibliotecas de noticias, manuales, textos específicos del dominio o diccionarios. Una forma de extracción de información es la extracción de texto, una tarea de recuperación de información dirigida a descubrir información nueva, previamente desconocida, extrayéndose automáticamente de diferentes recursos de texto. En la extracción de información, la extracción de texto se utiliza para eliminar información relevante de los archivos de texto basándose en algoritmos lingüísticos y estadísticos.

La búsqueda en la web y la extracción de información se realizan normalmente mediante web scrawler. Un web scrawler es un programa o script automatizado que navega por la web de manera metódica y automatizada. Una variante de los web scrawler son los raspadores web(web scrapers), que tienen como objetivo buscar ciertos tipos de información, como los precios de productos particulares de varias tiendas en línea, extraerlos y agregarlos a nuevas páginas web.

Los web scrapers se adoptan básicamente para transformar datos no estructurados y guardarlos en bases de datos estructuradas. Nosotros nos centramos en los web scrapers que extraen información textual de las páginas

web [12].

Para nuestro estudio usaremos esta técnica para simular la navegación de un usuario en la web e ir estructurando datos relevantes de manera automática y según lo requiera el programador.

Vale la pena señalar que el web scraping puede ir en contra de los términos de uso de algunos sitios web y algunas maneras de evitarlo pueden ser:

- Usar captchas.
- Bloquear el número de peticiones por dirección IP.
- Configurar el fichero robot.txt.
- Hacer uso de contenido dinámico en las web (javascript).
- Generar etiquetas html random para que la data no siga el mismo patrón.

3.1.1. Librerías usadas en web scraping

Se usaron las siguientes librerías en python para este proceso:

- **BeautifulSoup:** Es una biblioteca de Python que se usa para extraer datos de archivos HTML y XML. Funciona con su analizador favorito para proporcionar formas idiomáticas de navegar, buscar y modificar el árbol de análisis. Generalmente ahorra a los programadores horas o días de trabajo [13].
- **Requests:** Requests quita las complicaciones de trabajar HTTP/1.1 en Python - haciendo que la integración con servicios web sea transparente. No hay necesidad de agregar queries a tus URLs manualmente, o convertir tu información a formularios para hacer una petición POST. La reutilización de keep-alive y conexión HTTP se hace automáticamente, todo gracias a urllib3, el cual está integrado en Requests [14].

3.2. Support vector machine [7]

Las máquinas de vectores de soporte (SVM) son un conjunto de métodos de aprendizaje supervisado que se utilizan para la clasificación, regresión y detección de valores atípicos.

Las ventajas de las máquinas de vectores de soporte son:

- Eficaz en espacios de alta dimensión.
- Sigue siendo efectivo en casos donde el número de dimensiones es mayor que el número de muestras.
- Utiliza un subconjunto de puntos de entrenamiento en la función de decisión (llamados vectores de soporte), por lo que también es eficiente en memoria.
- Versátil: se pueden especificar diferentes funciones de Kernel para la función de decisión. Se proporcionan núcleos comunes, pero también es posible especificar núcleos personalizados.

Las desventajas de las máquinas de vectores de soporte incluyen:

- Si el número de funciones es mucho mayor que el número de muestras, evite la adaptación excesiva en la elección de las funciones del núcleo y el término de regularización es crucial.
- Los SVM no proporcionan directamente estimaciones de probabilidad, estas se calculan utilizando una costosa validación cruzada.

SVM posee clases capaces de realizar una **Clasificación** de múltiples clases en un data set, podemos ver un ejemplo sobre Iris en la figura 3.1.

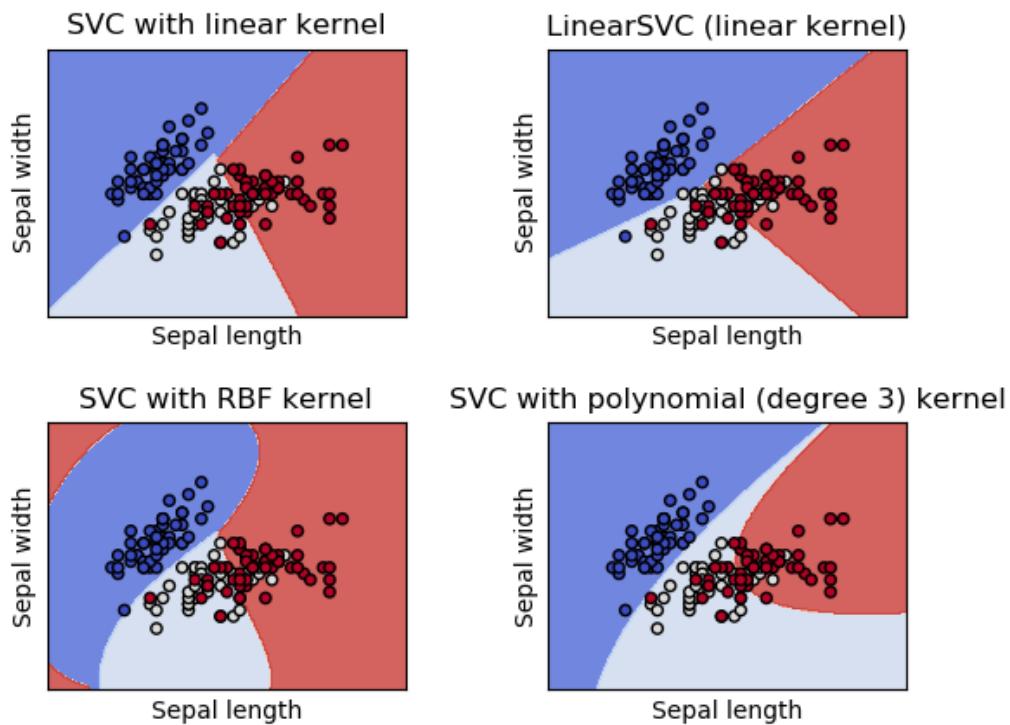


FIGURA 3.1: Dataset Iris. Fuente: <https://scikit-learn.org/>

3.2.1. Linear support vector classification

Para este trabajo necesitamos un clasificador de tipo binario es así como llegamos a LinearSVC.

LinearSVC es una implementación para casos de clasificación de SVM, cuenta con más flexibilidad en la elección de penalizaciones y funciones de pérdida, y escala mejor a un gran número de muestras.

3.3. API Twitter [8]

Twitter ofrece una API para que desarrolladores trabajen sobre sus datos, algunos casos de uso que se pueden dar son los siguientes:

- Buscar tweets: Te permite buscar tweets definiendo ciertos filtros como palabras clave, fecha, ubicación, hashtag, usuario, ciudad, lenguaje, hora, etc.

- Filtrar tweets en tiempo real: Te brinda los tweets generados en tiempo real.
- Actividad de cuenta: Te brinda cierta audiencia de Twitter de acuerdo a la línea de tu sitio web.
- Twitter para sitios web: Integra Twitter en tu página web.
- Anuncios: Te permite anunciar tu sitio web en Twitter.
- Mensajes directos: Te permite postear o enviar mensajes a usuarios de Twitter desde tu sitio web.

La API de Twitter ofrece 3 tipos de cuentas para desarrolladores como vemos en la figura 3.2, nosotros usaremos la cuenta gratuita pero las limitaciones son un problema ya que solo podemos acceder a tweets de 7 días atrás, para solucionar esto usamos la captura de tweets en tiempo real y vamos almacenando en una base de datos para el posterior uso de estos.

Tweets	Standard (free)	Premium	Enterprise
Publish and engage	✓		
Search Tweets: 7-days	✓		
Search Tweets: 30-days		✓	✓
Search Tweets: Full-archive		✓	✓
Filter Tweets	✓		✓
Sample Tweets	✓		✓
Batch Tweets			✓

FIGURA 3.2: Tipos de cuenta en la Api de Twitter. Fuente:
<https://developer.twitter.com/en/pricing>

Para acceder a la API de Twitter desde python, nuestro lenguaje de programación base, usaremos la librería tweepy.

3.3.1. Tweepy [9]

Tweepy es una librería para Python 2.6, 2.7, y 3.x para acceder al API de Twitter. Provee acceso a todos los métodos API RESTful, incluyendo la lectura y escritura de tweets. Tweepy soporta autenticación con OAuth esto nos facilita el acceso y uso a los servicios de la API de Twitter.

3.4. Análisis de Sentimiento

El análisis de sentimiento es un proceso de recopilación y análisis de datos basados en los sentimientos, las opiniones y los pensamientos de las personas. A veces al análisis de sentimiento se le conoce como opinion mining, ya que extrae la característica importante de las opiniones de la gente.

El análisis de sentimiento se realiza mediante el uso de varias técnicas de aprendizaje automático, modelos estadísticos y procesamiento de lenguaje natural (NLP).

El análisis de sentimiento se puede hacer a nivel de documento, frase u oración. En el nivel del documento, primero se toma un resumen del documento completo y luego se analiza si el sentimiento es positivo, negativo o neutral. En el nivel de frase, el análisis de frases es el análisis de cada oración, esto se toma en cuenta para verificar la polaridad. En el nivel de oración, cada oración se clasifica en una clase particular para proporcionar el sentimiento.

El análisis de sentimiento tiene varias aplicaciones:

- Se utiliza para generar opiniones para las personas de las redes sociales mediante el análisis de sus sentimientos o pensamientos que proporcionan en forma de texto
- Para obtener comentarios sobre cualquier producto en sus lanzamientos.
- Para obtener valoraciones de películas en sus estrenos a partir de los comentarios generados por la crítica.
- Para la creación de chatbots en atención al cliente.

- Encontrar polaridad positiva en noticias para dirigir anuncios a usuarios.
- En el marketing los comentarios positivos le da un valor agregado al producto, aplicando análisis de sentimiento podemos generar puntuaciones en base a los comentarios.

Twitter es una plataforma de microblogging en la que cualquier persona puede leer o escribir mensajes cortos que se llaman tweets. La cantidad de datos acumulados en twitter es muy grande. Estos datos no están estructurados y están escritos en lenguaje natural.

El análisis de sentimiento de Twitter es el proceso de acceder a los tweets para un tema en particular y predice el sentimiento de estos tweets como positivo, negativo o neutral con la ayuda de diferentes algoritmos de aprendizaje automático.

3.4.1. Natural Language Toolkit

NLTK es una biblioteca en Python, que proporciona una base para la creación de programas y la clasificación de datos. NLTK es una colección de recursos para Python que se puede usar para el procesamiento, la clasificación, el etiquetado y la tokenización de textos. Esta librería desempeña un papel clave en la transformación de los datos de texto en los tweets en un formato que se puede utilizar para extraer el sentimiento de ellos.

NLTK proporciona varias funciones que se utilizan en el preprocesamiento de datos para que los datos disponibles en Twitter se adapten a las funciones de minería y extracción de características. NLTK es compatible con varios algoritmos de machine learning que se utilizan para entrenar algún clasificador y para calcular su precisión.

Usaremos Python como lenguaje de programación base y NLTK es una biblioteca que desempeña un papel muy importante en la conversión de texto de lenguaje natural a un sentimiento positivo o negativo.

NLTK también proporciona diferentes conjuntos de datos que se utilizan para el entrenamiento de clasificadores. Estos conjuntos de datos están

estructurados y almacenados en la biblioteca de NLTK, a la que se puede acceder fácilmente con la ayuda de Python.

Para encontrar la polaridad de un tweet existen ya varios modelos y librerías pero con soporte para palabras en inglés, por ejemplo:

- Text blob [15]: Biblioteca para el procesamiento de datos textuales. Proporciona una API simple para sumergirse en tareas comunes de procesamiento de lenguaje natural (NLP), como etiquetado de parte del discurso, extracción de frases nominales, análisis de sentimientos, clasificación y traducción.
- Sentiwordnet [16]: Es un recurso léxico para opinion mining. SentiWordNet asigna tres puntuaciones de sentimiento: positividad, negatividad, objetividad.
- indicio.io: Esta es una web muy completa para el análisis de sentimiento pero es de paga.

Por lo anterior nos encontramos con 2 opciones: traducir cada tweet recolectado y usar dichos modelos para el inglés o entrenar un modelo usando un data estructurada en español (corpus).

3.4.2. Data recopilada

Mediante el web scraping accedemos a los medios digitales usando sus RSS(ver figura 3.3), este es un formato XML que se usa para distribuir su contenido de manera estructurada en la web. De esta forma capturamos todo el texto de las noticias publicadas en su web en una base de datos para luego hacer uso de las mismas.

También con el web scraping y la API de twitter podemos capturar todos los tweets originados en Lima y almacenarlas en una base de datos.

The screenshot shows the URL <https://elcomercio.pe/rss/> at the top left. The header features the El Comercio logo with a yellow background and a black circular icon. Below the header, the word "CANALES RSS" is centered. A horizontal line separates this from the main content. The text reads: "Ya no necesitas buscar las noticias: ahora ellas te encuentran. Anímate a usar la tecnología RSS (**Real Simple Syndication**), un sistema que te alertará automáticamente de las novedades de elcomercio.pe y de otros sitios de su interés." Another horizontal line follows. Below this, a section titled "PORTADA" is shown with five RSS feed icons: Política, Perú, Lima, Tecnología y ciencias, Mundo, and Economía.

FIGURA 3.3: Canales RSS. Fuente: <https://elcomercio.pe/rss/>

3.4.3. Data procesada

La data extraída de Twitter aún no puede ser procesada porque a pesar de ser solo texto, esta data aún no es data limpia ya que contiene:

- Caracteres no alfabéticos, esto lo solucionamos eliminándolos a través de expresiones regulares.
- Stop-word, se denomina así a las palabras sin significado como artículos, pronombres, preposiciones, palabras propias del sitio web, etc. Esto lo solucionamos importando stopwords.words('spanish') de nltk.corpus, una función de la librería NLTK.
- Links propios de Twitter o emoticones, esto lo solucionamos usando expresiones regulares

3.4.4. Modelo desarrollado

Respecto a los diferentes modelos de clasificación que estaban disponibles se decidió por usar SVM, en especial su variante LinearSVC. Para este propósito, se hace uso de la biblioteca de python scikit-learn [7], que proporciona un conjunto muy completo de herramientas para el aprendizaje automático. El objetivo es producir un clasificador binario que sea capaz de juzgar si el tweet extraído tenía mas probabilidades de ser positivo(igual a 1) o negativo(igual a 0).

Como este problema es un problema de clasificación binaria, una buena métrica es el Área bajo la Curva ROC (ver figura 3.4), que tiene en cuenta tanto los Falsos positivos (es decir, tweets negativos que fueron clasificados como positivos) como los Falsos Negativos (es decir, los tweets positivos que fueron clasificados como negativos).

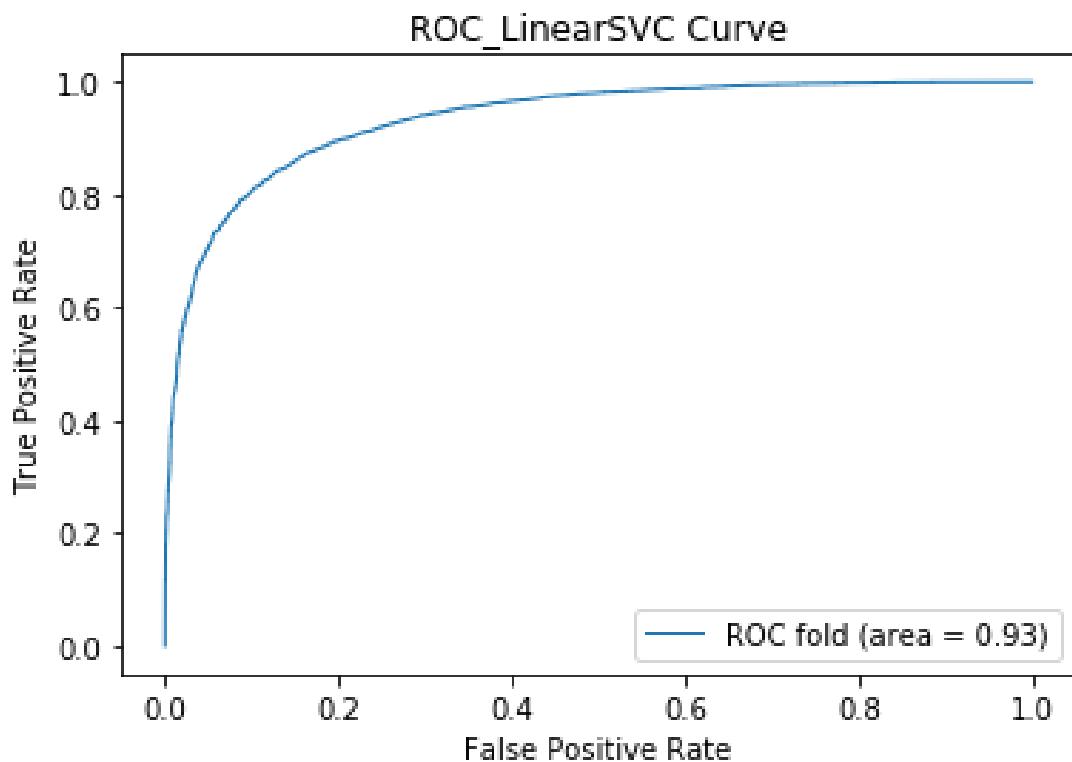


FIGURA 3.4: Curva ROC. Fuente: Elaboración propia

LinearSVC fue el que produjo mejores valores de AUC (area under the curve).

Una vez hemos seleccionado nuestro clasificador, hacemos una búsqueda en rejilla (GridSearchCV) para encontrar los mejores parámetros para nuestro modelo.

GridSearchCV itera sobre los modelos especificados con el rango de parámetros indicados y nos devuelve el modelo cuyos parámetros proporcionan los mejores resultados.

3.4.5. Data de entrenamiento

Una vez elegido el modelo, nuestro siguiente paso es obtener muestras ya clasificadas para el entrenamiento, para esto durante la investigación nos encontramos con el TASS el cual nos brinda una data de aproximadamente 55000 tweets etiquetados con sus respectivas polaridades.

La data estructurada que nos facilita el TASS (corpus) nos brinda características como “id del tweet”, “id del usuario”, “contenido del tweet”, “idioma del tweet” y “polaridad del tweet” que puede tomar 4 valores posibles:

- **N** (Negativo): los tweets pueden contener palabras ofensivas o sentimientos negativos.
- **P** (Positivo): los tweets muestran sentimientos de afecto o elogios.
- **NEU** (Neutral): los tweets no reflejan sentimientos positivos ni negativos.
- **NONE** (Ninguno): no se pudo determinar.

Recordemos que estas polaridades son generadas por el mejor modelo entrenado de cada competencia realizada un año antes, con esto se busca ir desarrollando este tipo de aprendizaje para el idioma español.

La estructura del corpus que nos brinda el TASS con el que trabajaremos es el siguiente:

```

1 <?xml version="1.0" encoding="UTF-8"?>
2 <tweets>
3   <tweet>
4     <tweetid>772432598027145216</tweetid>
```

```
5   <user>71546415</user>
6   <content>Sin ser fan de Juan Gabriel, siempre supe que era una fuerza
7   de la naturaleza. Hoy escuch\'e "Querida", y me dio una ternura enorme.
8   </content>
9   <date>Sun Sep 04 13:54:17 +0000 2016</date>
10  <lang>es</lang>
11  <sentiment>
12    <polarity><value>P</value></polarity>
13  </sentiment>
14  </tweet>
15  <tweet>
16    <tweetid>771715645843079169</tweetid>
17    <user>106919551</user>
18    <content>ayer preguntaban y d\'onde est\'an las solteras!!!! todo mi grupo
19    alza la mano y yo la \'unica que no y todas voltean a verme AJAJAJAJAJJA
20    </content>
21    <date>Fri Sep 02 14:25:22 +0000 2016</date>
22    <lang>es</lang>
23    <sentiment>
24      <polarity><value>NEU</value></polarity>
25    </sentiment>
26  </tweet>
27  <tweet>
28    <tweetid>774703192105971712</tweetid>
29    <user>280686132</user>
30    <content>Hoy me sent\'i como grace de "al fondo hay sitio" cuando la
31    atropellaron, solo que no fue de mentira y casi no la cuento que horrible
32    </content>
33    <date>Sat Sep 10 20:16:48 +0000 2016</date>
34    <lang>es</lang>
35    <sentiment>
36      <polarity><value>N</value></polarity>
37    </sentiment>
38  </tweet>
39  <tweet>
40    <tweetid>767891418849292288</tweetid>
41    <user>759585997</user>
```

```

42   <content>es el tercer d\ía de clases, ya tengo 4 tns de tarea, no he
43   instalado ning\'un software y acabo de despertar de una siesta de 7 horas
44   </content>
45   <date>Tue Aug 23 01:09:15 +0000 2016</date>
46   <lang>es</lang>
47   <sentiment>
48     <polarity><value>NONE</value></polarity>
49   </sentiment>
50   </tweet>
51   <tweet>
```

3.5. Base de datos

Para la administracion de la base de datos relacional se hara uso de MySQL (RDBMS). MySQL es la base de datos de código abierto más popular del mundo. Con su rendimiento, confiabilidad y facilidad de uso comprobados, MySQL se ha convertido en la principal opción de base de datos para aplicaciones basadas en la Web, utilizada por propiedades web de alto perfil como Facebook, Twitter, YouTube, y los cinco principales sitios web. Además, es una alternativa extremadamente popular como base de datos integrada basado en lenguaje de consulta estructurado (SQL).

3.6. Gráficos

Para los gráficos nos apoyaremos en la librería de python matplotlib la cual provee muchas herramientas para distintos tipos de gráficos y también usaremos Heatmap.py para los mapas de calor con los tweets ya clasificados.

3.6.1. Matplotlib [10]

Matplotlib es una biblioteca de trazado 2D de Python que produce gráficos de calidad para papers en una variedad de formatos y entornos interactivos en

todas las plataformas. Matplotlib se puede usar en scripts de Python e IPython shells, Jupyter notebook, en los servidores de aplicaciones web, etc. Vemos un ejemplo básico en la figura 3.5.

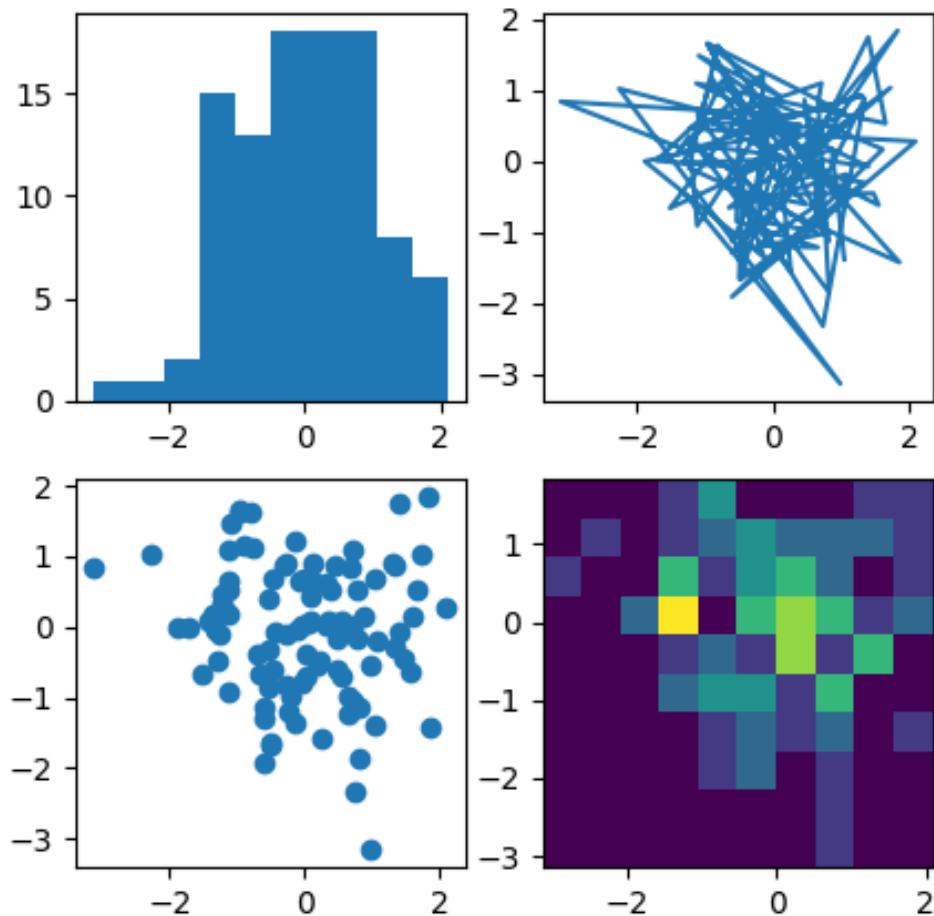


FIGURA 3.5: Ejemplos simples de matplotlib. Fuente:
<https://matplotlib.org/>

3.6.2. Heatmap.py

Usaremos este programa por su simplicidad de ser un único script el cual solo necesita las coordenadas y muestra la densidad de los datos como en la figura 3.6.

Este programa se apoya en Open Street Map [17] para las locaciones al momento de hacer el mapa de calor.

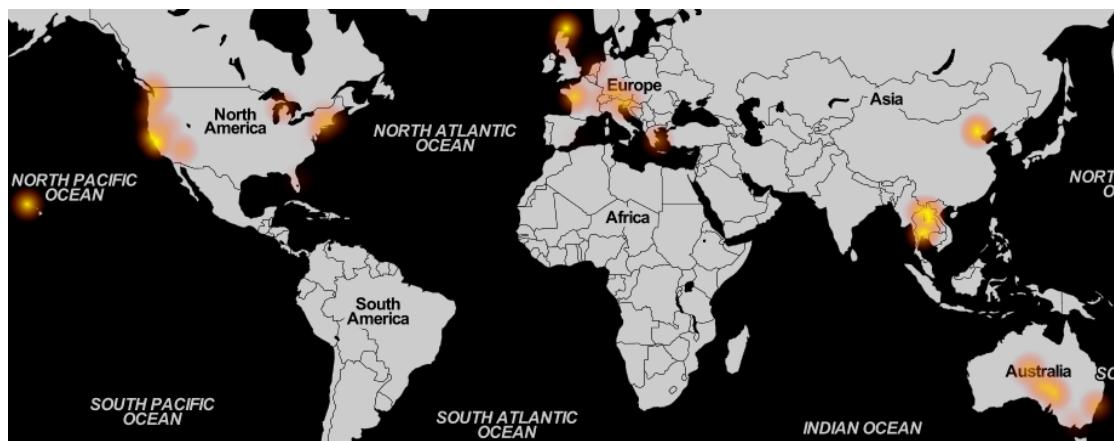


FIGURA 3.6: Ejemplo mapa de calor. Fuente: <http://www.sethoscope.net/heatmap/>

3.7. Amazon web services [11]

AWS es una colección de servicios de computación en la nube pública que en conjunto forman una plataforma de computación en la nube, ofrecidas a través de Internet por Amazon.com. Es usado en aplicaciones populares como Dropbox, Foursquare, HootSuite. Es una de las ofertas internacionales más importantes de la computación en la nube y compite directamente contra servicios como Microsoft Azure y Google Cloud Platform. Es considerado como un pionero en este campo.

Para nuestro estudio nos centraremos en el servicio gratuito que nos ofrece AWS por un año con ciertas limitaciones pero no afecta en nuestro trabajo, nos referimos a EC2 de AWS free tier.

3.7.1. Elastic Compute Cloud

EC2 es un servicio web que proporciona capacidad informática en la nube, segura y de tamaño modifiable. Está diseñado para simplificar el uso de la informática en la nube a escala web para los desarrolladores.

La sencilla interfaz de servicios web de Amazon EC2 permite obtener y configurar capacidad con una fricción mínima. Proporciona un control completo sobre los recursos informáticos y puede ejecutarse en el entorno informático acreditado de Amazon. Amazon EC2 reduce el tiempo necesario para obtener e iniciar nuevas instancias de servidor en cuestión de minutos, lo que permite escalar rápidamente la capacidad, ya sea aumentando o reduciendo, en función de sus necesidades.

Amazon EC2 cambia el modelo económico de la informática al permitir pagar solo por la capacidad que utiliza realmente. Amazon EC2 nos brinda a los desarrolladores las herramientas necesarias para crear aplicaciones resistentes a errores y para aislarlas de los casos de error comunes.

Detallaremos la metodología usada en resumen para resolver los distintos problemas presentados durante el desarrollo de este informe, acompañado de un marco teórico correspondiente

- **Definimos la estructura de los datos a recolectar en los medios digitales.**

Este punto tiene como objetivo definir la estructura de datos a recolectar mediante el uso del web scraping y generar una base de datos usando MySQL.

- **Reconocemos las palabras más usadas por dia.**

Se busca las palabras con mayor frecuencia usadas por los medios digitales a fin de buscar estas mismas palabras en los tweets creados por usuarios de Lima.

- **Usamos el twitter API como medición de la opinión pública.**

Una vez definida las palabras más frecuentes se pasa a la recolección de las opiniones de twitter usando la API de la red social para la recolección de tweets en tiempo real acotando a la zona de la ciudad de Lima y generando una base de datos con lo recolectado.

- **Definimos los sentimientos reflejados en cada tweet.**

Buscamos obtener la polaridad del tweet de un determinado tema en

forma binaria(positivo=1 y negativo=0) usando análisis de sentimiento con algún algoritmo clasificador de machine learning.

- **Mostramos los resultados obtenidos en graficos.**

Se busca mostrar los resultados de manera legible usando gráfico de barras y mapas de calor.

Además que se resolverá los siguientes ítems en el capítulo siguiente:

- Como la teoría de la Agenda-Setting se cumple, buscamos el medio digital de mayor alcance a los peruanos para realizar nuestro estudio, en ese sentido nos encontramos que los medios digitales con más visitas pertenecen a un mismo dueño: “grupo El Comercio” (ver figura 1.1).
- Implementamos un clasificador binario con el algoritmo SVM, en su variante LinearSVC, ya que nos dará mejores resultados.
- Entrenamos el clasificador binario con el corpus brindado por el TASS juntando la data de todas sus ediciones para tener más data de entrenamiento categorizada ya que una de las características del LinearSVC es que funciona mejor con mayor cantidad de data de entrenamiento.

Capítulo 4

Evaluación y Resultados del Seminario

Como se ha visto en el capítulo 3, definimos las herramientas usadas para este proyecto, ahora detallamos los pasos realizados.

Nos enfocamos en 2 estructuras principales para este proyecto donde en la primera se aplica el web scraping(ver figura 4.5) y la segunda hacemos uso del análisis de sentimiento (ver figura 4.7)

4.1. Recolección de noticias de medios digitales

En este primer paso nos apoyamos en el formato RSS de las web de noticias(elcomercio.pe, trome.pe, peru21.pe, peru.com) la cual tienen la siguiente forma:

```
1 <rss version="2.0">
2   <item>
3     <title>
4       <! [CDATA[
5         Piden que salgan de la Subcomision quienes estaban en chat La Botica
6       ]]>
7     </title>
8     <description>
9       <! [CDATA[
10        Cesar Vasquez, vocero de APP, considera que no es correcto que sigan pertenecie
11       ]]>
12   </description>
```

```

13   <link>
14     http://elcomercio.pe/politica/piden-salgan-subcomision-quienes-estaban-ch
15   </link>
16   <category domain="http://elcomercio.pe/politica">Politica</category>
17   <guid isPermaLink="true">
18     http://elcomercio.pe/politica/piden-salgan-subcomision-quienes-estaban-ch
19   </guid>
20   <pubDate>Sun, 04 Nov 2018 10:16:32 -0500</pubDate>
21 </item>
```

Podemos observar que toda publicación realizada en su web tendrá esta estructura con los mismos atributos, lo cual nos permite capturar los links de cada noticia usando web scraping como vemos en la figura 4.1.

link	date	category
http://elcomercio.pe/lima/seguridad/siete-ninos-pe...	Tue, 13 Nov 2018 22:19:14 -0500	lima
https://peru.com/redes-sociales/facebook/facebook-...	Fri, 02 Nov 2018 17:04:53 -0500	redes-sociales
https://peru.com/actualidad/politicas/reyes-espana...	Sun, 11 Nov 2018 22:11:02 -0500	actualidad
https://peru.com/entretenimiento/celebrities/insta...	Sat, 03 Nov 2018 12:21:41 -0500	entretenimiento
http://elcomercio.pe/eldominical/ideas-filosoficas...	Tue, 13 Nov 2018 22:03:36 -0500	eldominical
http://elcomercio.pe/lima/sucesos/crimen-barranca-...	Wed, 14 Nov 2018 22:38:48 -0500	lima
http://trome.pe/viral/facebook/facebook-viral-qued...	Mon, 12 Nov 2018 22:58:32 -0500	Actualidad
http://elcomercio.pe/tvmas/farandula/instagram-ric...	Tue, 06 Nov 2018 22:45:01 -0500	tvmas

FIGURA 4.1: Links de las noticias publicadas. Fuente: Elaboración propia

Con los links almacenados podemos hacer un request a cada uno para almacenar los textos de cada noticia a nuestra base de datos(ver figura 4.2) para usarlos más adelante.

fecha	título	subtítulo	contenido
14/11/2018 - 11:59h	Boca Juniors vs River Plate: Carlos Bianchi causa polémica con este mensaje sobre la final VIDEO FOTOS	El técnico más ganador de Boca Juniors, Carlos Bianchi, emitió un comentario sobre cómo tiene que afrontar el hincha la final de la Copa Libertadores ante River Plate.	Cuando los hinchas de Boca Juniors están comiéndose las uñas pensando en la final de la Copa Libertadores en cancha de River Plate, apareció Carlos Bianchi, una institución en el cuadro xeneize y desde su retiro envió un mensaje. Un socio de Boca Juniors quien mantiene diálogo con Carlos Bianchi le hizo la consulta al entrenador, campeón en tres ocasiones de la Copa Libertadores, después de la final de ida contra River Plate. Julio Cronopiano, como se llama este socio de Boca Juniors publicó la breve respuesta del 'Virrey' en su cuenta de Twitter. El estratega ante la consulta sobre ¿cómo espera la revancha del sábado 24 de noviembre ante River Plate, en el Monumental? le escribió "Il Faut y Croire !!!" que en español significa "Hay que creer". Carlos Bianchi respondió en francés y a su vez generó gran expectativa entre los hinchas de Boca Juniors quienes le preguntaron a Julio Cronopiano si este mensaje era real. El socio reconoció que el DT ya le había mandado antes un audio cuando el equipo estaba a punto de quedar eliminados. "Confirmons, es la palabra del más grande".
14/11/2018 - 12:15h	Perú vs Ecuador: Con Raúl Ruidíaz recuperado la selección entrena en el Estadio Nacional VIDEO FOTOS	Ricardo Gareca incluyó a Raúl Ruidíaz dentro de los jugadores hábiles para el amistoso de la selección peruana este jueves ante Ecuador.	La selección peruana realizó la mañana del miércoles su entrenamiento en el Estadio Nacional y una de las novedades en los trabajos que supervisó Ricardo Gareca ha sido la presencia de Raúl Ruidíaz, quien hasta ayer parecía estar descartado para el duelo de mañana ante la selección de Ecuador. Desde las 9 de la mañana Ricardo Gareca llegó junto al plantel nacional hasta el coloso de la calle José Díaz. Desde el ingreso del bus se pudo observar la presencia de Raúl Ruidíaz dentro del plantel y luego se le vió con ropa de trabajo y participando del tradicional 'camotito' para luego trabajar en el partido de práctica previo al Perú vs Ecuador. Raúl Ruidíaz solo habría requerido un día de descanso, para superar una fuerte contusión en la cabeza durante el duelo de la MLS con los Seattle Sounders. Hoy con chaleco GPS, al igual que el resto de sus compañeros, fue monitoreado para ver las condiciones en las que llega para este duelo y confirmar si entra en la lista definitiva de Ricardo Gareca. La selección peruana trabajó por espacio de una hora y media con trabajos tácticos de ataque y defensa en espacio reducido. Asimismo, Jefferson Farfán, Yoshimar Yotún y Miguel Trauco ensayaron tandas de tiros libres y otras jugadas de balón detenido.

FIGURA 4.2: Texto de noticias publicadas. Fuente: Elaboración propia

4.2. Palabras claves o con mayor frecuencia usada por los medios digitales

Usando la librería NLTK podemos hallar las palabras claves, de significado relevante, con mayor frecuencia usadas por los medios de comunicación digitales. Por ejemplo podemos ver los resultados para el dia 4 de octubre en la figura 4.3.

Out[20]:	palabra	frecuencia	fecha
	16 Elecciones	35	4
	17 Fujimori	31	4
	18 candidatos	23	4
	19 Alberto	17	4
	20 indulto	16	4
	21 Congreso	15	4

FIGURA 4.3: Palabras y su frecuencia para el dia 4/11/18. Fuente: Elaboración propia

Tambien podemos verificar las noticias de la figura 4.3 gracias a consultas SQL y obtener el siguiente resultado, por ejemplo para la palabra “congreso”, mostrado en la figura 4.4.

fecha	título
04.10.2018 / 08:40 pm	Congreso no acusará a ex miembros del CNM por organización criminal
04.10.2018 / 07:21 pm	Congreso aprueba acusar a Hinostroza por pertenecer a una organización criminal
04.10.2018 / 06:47 pm	Congreso debate acusaciones contra Hinostroza y ex consejeros [GALERÍA]
04.10.2018 / 03:25 pm	Pleno del Congreso debate denuncia constitucional contra César Hinostroza [EN VIVO]
04.10.2018 / 02:40 pm	Iván Noguera ante el pleno del Congreso: "Estoy orgulloso de mis audios"
04.10.2018 / 07:55 am	Congreso aprueba que las cuatro reformas constitucionales pasen a referéndum
04.10.2018 / 07:30 am	Fusiones y adquisiciones: Los detalles del PL que avanza al Pleno del Congreso
04.10.2018 / 20:57 PM	Congreso no acusó de organización criminal a 4 exmiembros del CNM
04.10.2018 / 19:24 PM	César Hinostroza fue destituido e inhabilitado por 10 años por el Congreso
04.10.2018 / 14:09 PM	Iván Noguera ante el pleno del Congreso: "Estoy orgulloso de mis audios"
04.10.2018 / 10:04 AM	César Hinostroza: Congreso aprobó incluir delito de crimen organizado en informe
04.10.2018 / 09:18 AM	Congreso aprueba la no reelección inmediata de congresistas
04.10.2018 / 09:12 AM	Congreso debate este jueves informe sobre César Hinostroza y ex consejeros del CNM
04/10/2018 - 21:08h	César Hinostroza destituido e inhabilitado por 10 años por el Congreso de la República VIDEO FOTOS
04/10/2018 - 17:08h	Iván Noguera armó un show durante su defensa en el Congreso: "Estoy orgulloso de mis audios" VIDEO

FIGURA 4.4: Noticias del dia 4/10/18 con relacion a la palabra congreso. Fuente: Elaboración propia

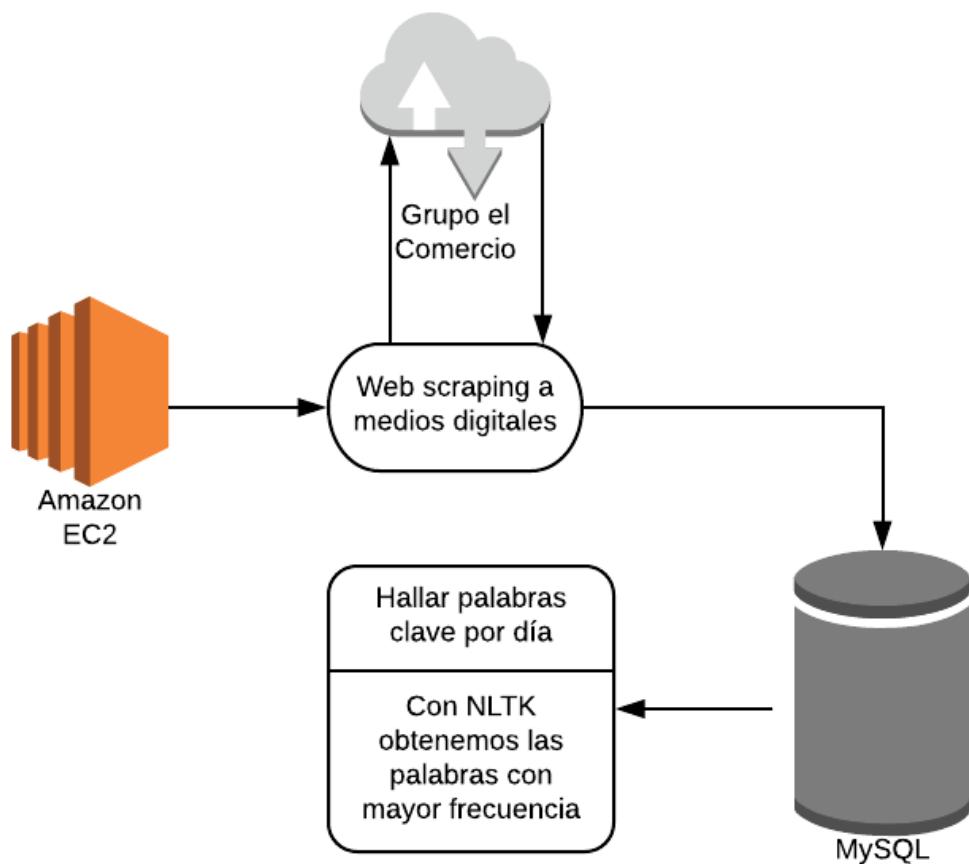


FIGURA 4.5: Diseño primera parte del proyecto. Fuente:
Elaboración propia

4.3. Recolección de tweets usando la API de twitter

Para este proceso nos encontramos con un problema, lo cual nos impedía obtener tweets de alguna fecha específica ya sea por el límite de consultas o por no tener una cuenta premium en la API de Twitter lo cual nos limitaba.

Solucionamos este inconveniente haciendo consultas en tiempo real y almacenando cada tweet creado en la zona de Lima con los siguientes atributos mostrados en la figura 4.6.

text	user	location	created
Gareca lo dejó más que claro!	@RenzoGQ10		Sun Nov 04 18:03:26 +0000 2018
Y siempre, siempre nos perdona! "La Misa del domi...	@Kamaridul		Sun Nov 04 18:03:20 +0000 2018
#TeaTime #FamilyPUCP en Las Bolena TeaRoom & R...	@marianaalinne	-12.1315868,-77.023733	Sun Nov 04 18:03:06 +0000 2018

FIGURA 4.6: Atributos capturados. Fuente: Elaboración propia

Haciendo uso de la librería NLTK podemos encontrar las palabras con mayor frecuencia usada en las noticias para encontrar los tweets que tengan relación con estas palabras mediante consultas SQL a nuestra base de datos.

En 20 días aproximadamente pudimos capturar alrededor de 12000 tweets con relación a las noticias publicadas en los medios digitales los cuales incluyen publicaciones no solo en español, por lo que tuvimos que hacer uso de las librerías textblob, langid y detect para separar los tweets escritos en inglés con lo que nos quedamos con alrededor de 10900 tweets clasificados como español para el procesamiento.

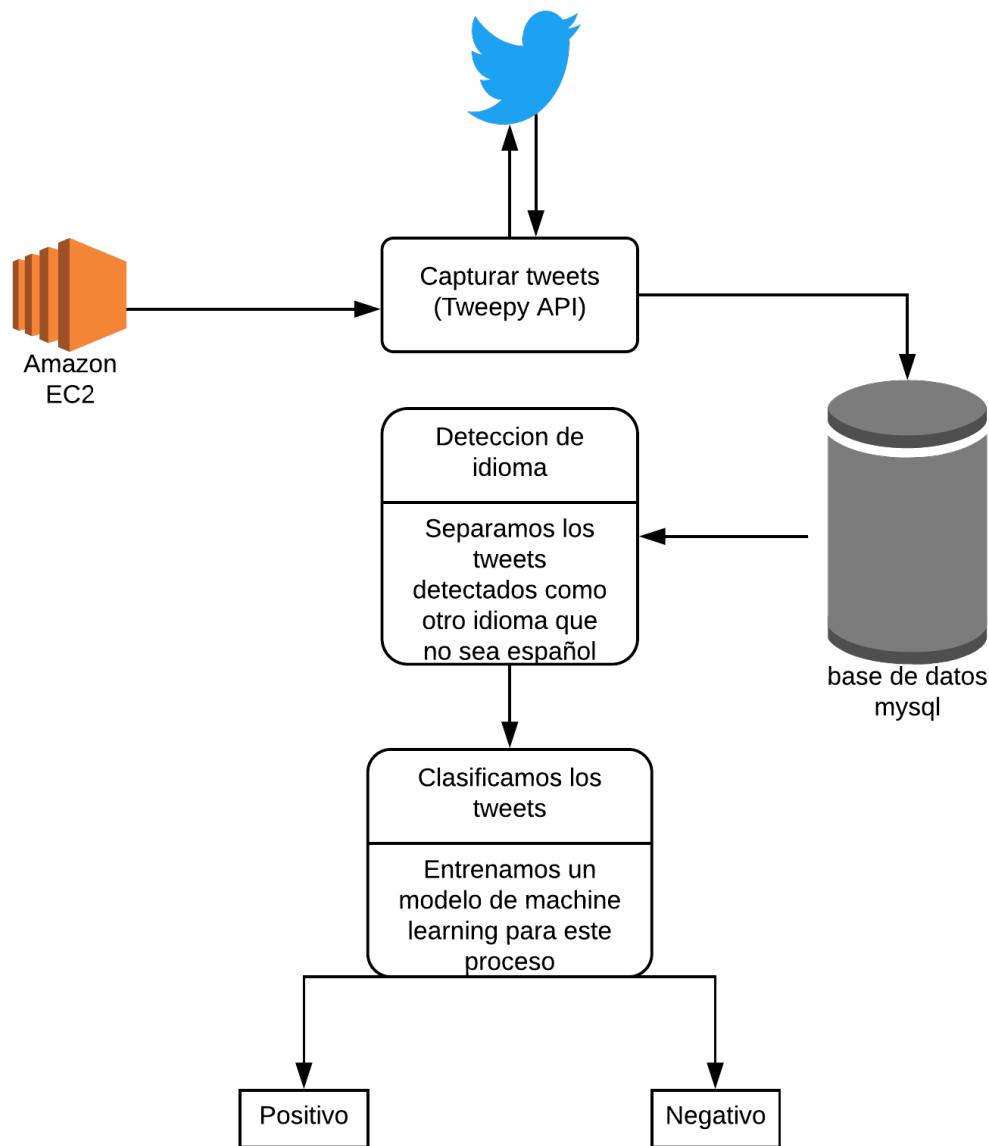


FIGURA 4.7: Diseño segunda parte del proyecto. Fuente:
Elaboración propia

4.4. Modelo usado

Como ya vimos en el capítulo 2 el modelo que usaremos será el algoritmo LinearSVC, un aspecto importante a considerar es elegir una métrica apropiada para evaluar este modelo. Lo que buscamos es un problema de clasificación binaria (predecir si un tweet es positivo =1 o negativo=0).

El siguiente paso sería limpiar la data para el entrenamiento del modelo, a esto llamaremos pre-procesamiento de datos:

4.4.1. Pre-procesamiento de datos

Para entrenar el modelo necesitamos data limpia que pueda ser procesada por el algoritmo LinearSVC para esto necesitamos tokenizar, lematizar y otros pasos previos que se mencionan a continuación:

Tokenizar

Usamos nltk.word_tokenize de la librería NLTK para este proceso, este tokenizador también remueve los signos de puntuación y separa la oración en palabras.

Lematizar

Usamos nltk.stem de la librería NLTK para este proceso, luego de la tokenización necesitamos obtener la raíz de cada palabra.

CountVectorizer

Nos apoyamos de esta herramienta de la librería sklearn para convertir el texto del tweet en una matriz en la que cada palabra es una columna cuyo valor es el número de veces que dicha palabra aparece en cada tweet.

Cambiar a minúsculas

Usamos CountVectorizer de la librería sklearn.feature_extraction.text y pasamos el parámetro lowercase como True esto convierte todos las letras mayúsculas a minúsculas.

Remover stopwords

Usamos stopwords de la librería nltk.corpus lo cual nos provee soporte para las palabras vacías o sin valor del idioma español.

```

7    Más de mañana en Gaceta. UPyD contará casi seguro con grupo gracias al Foro Asturias. Eso se dice en el Congre
50
7    Las remuneraciones económicas son lo que todos esperan, pero un "me encanta lo que has hecho", lo vale todo. E
l cliente es primero.
Name: content, dtype: object
['asturi', 'casi', 'cliente', 'congres', 'cont', 'dic', 'econom', 'encant', 'esper', 'for', 'gacet', 'graci', 'gru
p', 'hech', 'mas', 'mañan', 'per', 'primer', 'remuner', 'segur', 'tod', 'upyd', 'val']
[[1 1 0 1 1 1 0 0 0 1 1 1 0 1 0 0 0 1 0 1 0]
 [0 0 1 0 0 0 1 1 1 0 0 0 0 1 0 0 1 1 0 2 0 1]]

```

FIGURA 4.8: Ejemplo de data limpia para dos tweets. Fuente:
Elaboración propia

Con este pre-procesamiento obtenemos una matriz necesaria para que el algoritmo inicie el entrenamiento, como podemos ver en la figura 4.8 vemos como queda la matriz para dos valores luego de todos los procesos anteriormente mencionados.

4.4.2. Búsqueda de hiperparametros

Luego del pre-procesamiento necesitamos los mejores valores para iniciar el algoritmo, para esto hacemos uso de GridSearchCV, esto nos devuelve un conjunto de valores llamados hiperparametros los cuales nos generan la mejor AUC, para nuestro caso obtenemos un valor de 0.8521.

Con esto ya podemos entrenar el modelo y pasar a predecir la polaridad de los tweets capturados.

4.5. Predicción de la polaridad de un tweet

Durante el mes que se hizo la recolección de tweets tambien se pudo obtener el origen de algunos de estos, con estas coordenadas se realizó un mapa de calor como se muestra en la figura 4.9

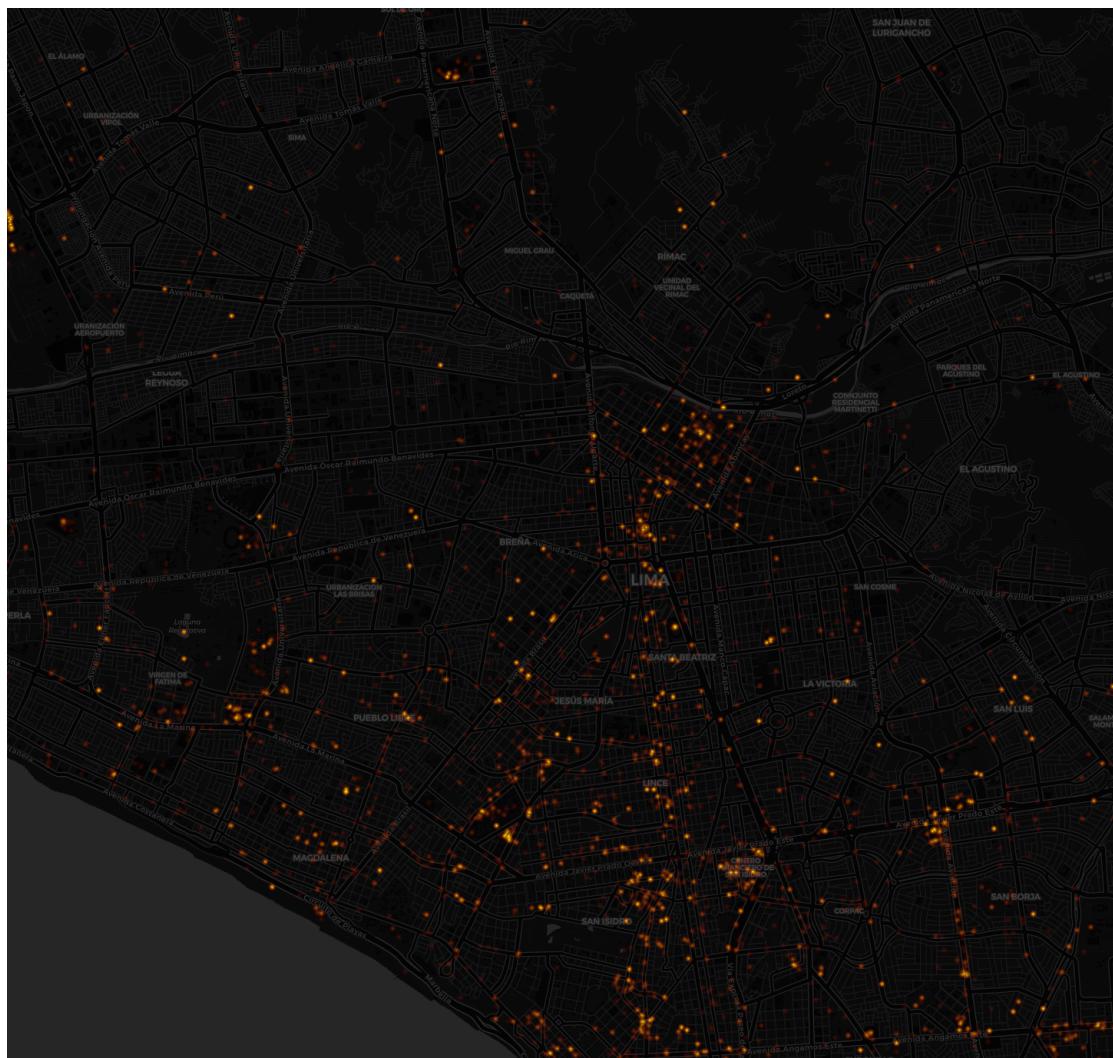


FIGURA 4.9: Mapa de calor de los tweets recolectados en Lima usando sus coordenadas. Fuente: Elaboración propia

Ya con el modelo entrenado y con los tweets solo en español pasamos a predecir la polaridad de cada uno obteniendo los siguientes resultados como se muestra en la figura 4.10 donde podemos ver el número de tweets totales por día.

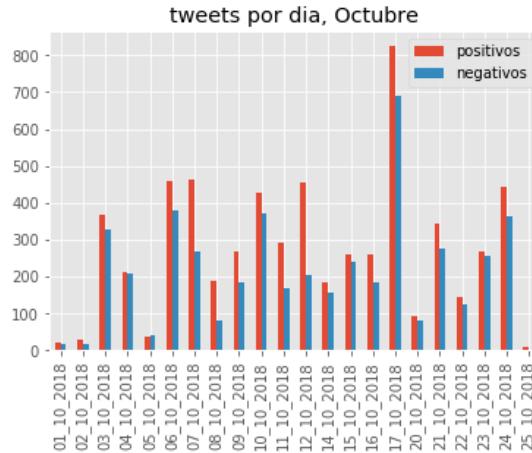


FIGURA 4.10: Número de tweets por día. Fuente: Elaboración propia

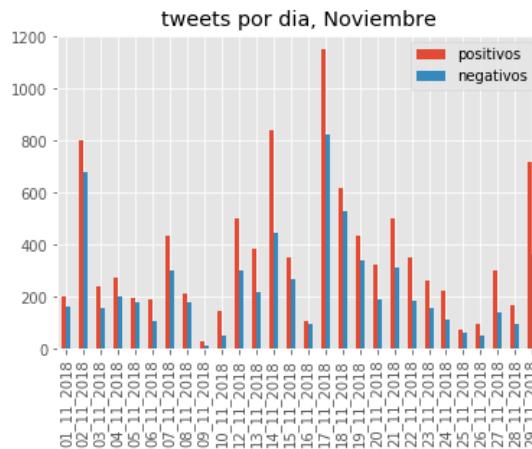


FIGURA 4.11: Número de tweets por día. Fuente: Elaboración propia

Con los tweets clasificados también podemos hacer un mapa de calor separando por colores(rojo=negativo y verde=positivo) como vemos en la figura 4.12.

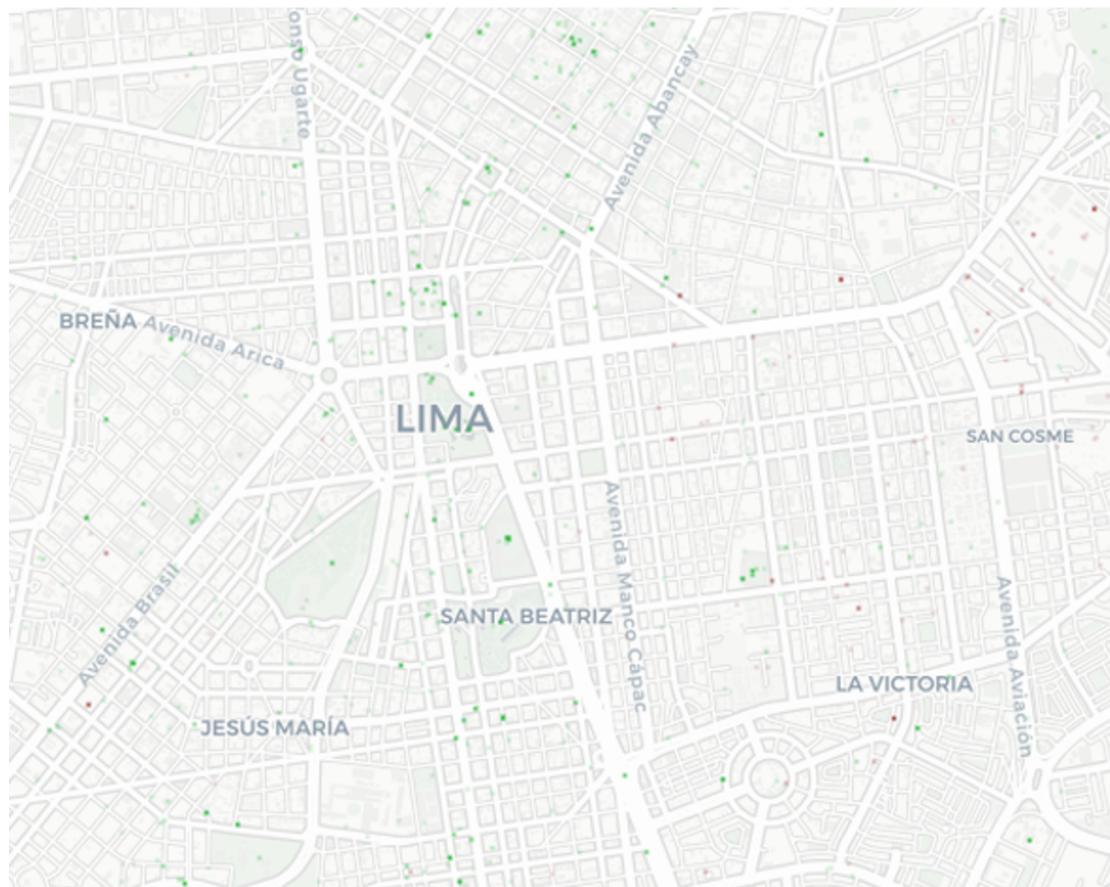


FIGURA 4.12: Mapa de calor de los tweets recolectados en Lima con su polaridad. Fuente: Elaboración propia

4.6. Otras aplicaciones de Web Scraping

Durante el proceso de aprendizaje del web scraping se realizó algunas pruebas, podemos mostrar algunos trabajos propios donde no se tiene cuidado con el uso de esta técnica:

- **ORCE:** Esta web muestra la información de todos los alumnos y egresados de nuestra casa de estudio(ver figura 4.13)



Búsqueda de Alumnos y Egresados

Código UNI: CODIGO UNI
 Apellido Paterno: APE. PATERNO
 Apellido Materno: APE. MATERNO
 Nombre: NOMBRE

Nota: Para efectuar la búsqueda, sólo es necesario colocar uno o más datos solicitados, haga click en los campos para mostrarle un ejemplo.

FIGURA 4.13: Busqueda de alumnos y egresados. Fuente:
<http://www.orce.uni.edu.pe/buscaalu.php?op=buscaalu>

Pude probar que con un simple script descargué toda la data que muestra esta web como vemos en la figura 4.14.

```
kevin@kevin-Inspiron-14-3467: ~/Escritorio/python
20072053H, C1, ALEGRE-MILLA-ENRIQUE GREGORIO
20072049K, M3, ALVINO-RODRIGUEZ-FRANCIS
20072021I, M6, ALVARADO-ALVA-DANIEL ANGEL
20072018H, C1, ACOSTA-LEZAMA-JOSE HUMBERTO
20072012J, L2, ANDRADE-TEEVIN-RUBEN LEONARDO
20072008B, M6, ALBITES-SANABRIA-JOSE LUIS
20072007F, C1, ACOSTA-GRAÑEZ-HUGO LEONIDAS
20072004G, L2, AGAMA-MOLINA-JOSE DANIEL
20072004G, L2, AGAMA-MOLINA-JOSE DANIEL
20071402I, A1, ABSI-MEJIA-MARLO ISRAEL
20071387J, A1, AYALA-ARANGO-HERNAN MIGUEL MAC
20071355K, P3, ASCENCIOS-ALBAN-ERNESTO DAVID
20071350I, E3, APEÑA-CHILI-CHRISTIAN IVAN
20071346A, Q1, ALVARADO-BALTAZAR-ROBERT TEODORO
20071344I, N3, AYALA-PIZARRO-DAGOBERTO MANUEL
20071343B, E3, ALVA-RODRIGUEZ-EDNAR OSCAR
20071342F, Q1, AREVALO-FLORES-JORGE ARTURO
20071338I, E3, ALVARADO-VALLEJOS-ARNALDO EDUARDO
20071319D, E1, AQUINO-VALLEJOS-CARLOS JUNIOR
20071318H, P3, ALAMA-OTAEGUI-ALONSO WILFREDO
20071312J, Q1, ARESTEGUI-ROMAN-NOE
20071284F, E1, ARUATA-MAMANI-JORGE LUIS
20071281G, C1, ANCCASI-CANDIOTTI-FREDDY
20071280K, Q1, ANCALLE-DE LA CRUZ-ROSA MERCEDES
20071274K, G1, ALCANTARA-CHUCO-OCTAVIO FERMIN
20071262B, A1, ALVAREZ-ANGULO-MARIA VICTORIA
20071242A, S1, AURIS-CORTEZ-YESICA FIORELA
20071235E, G3, ALVAREZ-HUAYHUA-MIGUEL GRIMALDO
20071234I, Q2, ALBERTI-MAYS-MARYSABEL ELAINE
20071234I, Q2, ALBERTI-MAYS-MARYSABEL ELAINE
20071223G, C1, ALDAVE-GUILLEN-LUIS ALEJANDRO
20071217G, Q1, ALMONACID-CHURA-CARLOS ALBERTO
20071208H, M4, ACOSTA-CIRIACO-FRANK ENRIQUE
20071182I, M4, ALDANA-CORDOVA-JOSEPH JULIO
20071173J, M4, ALVAREZ-ESCOBAR-RAFAEL
20071165G, Q1, AMASIFEN-BRAVO-VIRGINIA NATALY
20071150J, G1, ARIZA-SANCHEZ-ERICK GODOFREDO
20071143C, I1, ALVITES-CARPIO-DANIEL ALEJANDRO
20071098H, L1, ACASiete-VEGA-KEVIN ODYN SKY
20071092J, S1, AYUQUE-MENDOZA-EDZON RHOMARIO
20071090G, I1, ARAGON-VALLADOLID-JAVIER ENRIQUE
20071088B, I1, ALVAREZ-HERRERA-RICARDO DIONICIO
20071064F, M6, ACEVEDO-PASCUAL-GLIDVER FRANCO
20071058F, C1, ALVAREZ-JARA-KIRO IRWING
20071052H, M4, ALMIDON-FLORES-JOSEPH MIJAIL
20071049G, P2, AQUINO-HANCO-NESTOR
20071046H, M4, ALBINO-MARTINEZ-CESAR CRISTIAM
20071039A, I2, ATALLUZ-HUARI-OMAR ENMANUEL
20071034J, I2, ATOCHE-BRAVO-JUNIOR ALIN
20070435K, A1, AVILA-COLCHAO-JULIO CESAR
```

FIGURA 4.14: Web scraping hecho en python. Fuente: Elaboración propia

- ONPE: Para el proceso de elección de alcaldes se habilitó una web para consultar miembros de mesa(ver figura 4.15)



FIGURA 4.15: Consulta miembro de mesa. Fuente: <https://consultamiembrodemesa.onpe.gob.pe/>

En esta web tampoco se tuvo los cuidados para evitar el web scraping ya que esta data resulta ser valiosa para ciertas empresas, como podemos ver en la figura 4.16 se realiza la extracción sin ningún problema.

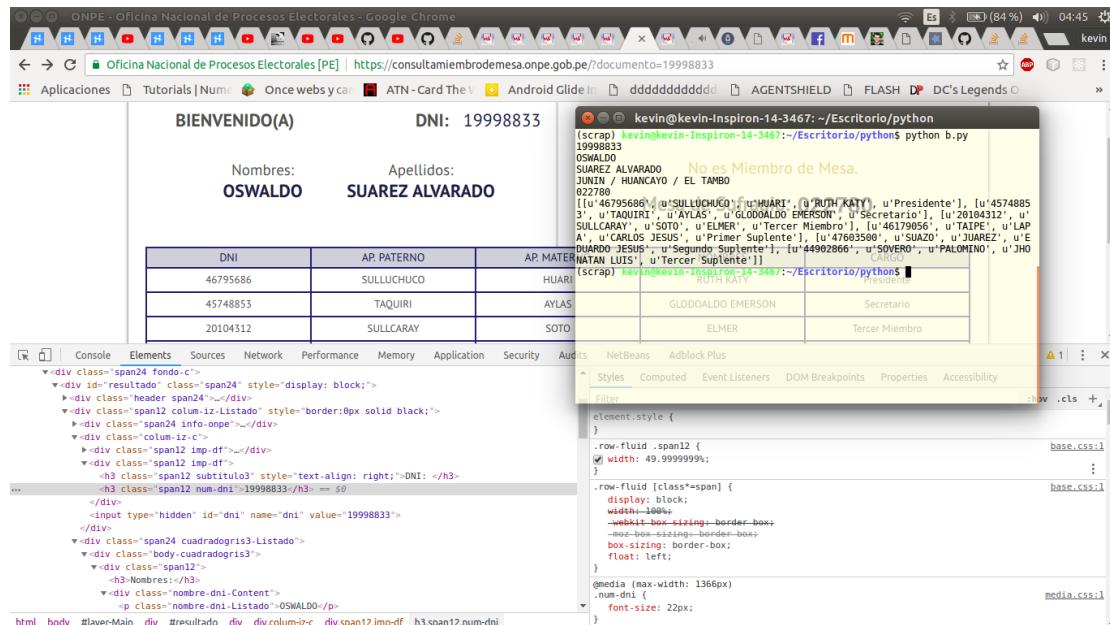


FIGURA 4.16: Web scraping a la página de ONPE. Fuente: Elaboración propia

Como resultado interesante obtuvimos que el dni vigente de menor cifra fue el 00000317 (ver figura 4.17) o que existe el dni 00001000.

BIENVENIDO(A)	DNI: 00000317	UCAYALI / CORONEL PORTILLO / CALLERIA
Nombres: MIGUEL	Apellidos: RODRIGUEZ DEL AGUILA	No es Miembro de Mesa. Mesa de Sufragio: 077363

FIGURA 4.17: Menor número dni vigente. Fuente:
<https://consultamiembrodemesa.onpe.gob.pe/>

Capítulo 5

Conclusiones y Trabajos Futuros

Describiremos las conclusiones y el trabajo a futuro del proyecto realizado.

5.1. Conclusiones

A partir de este trabajo pasamos a mencionar las conclusiones a las que llegamos:

- El análisis de sentimiento se utiliza para identificar la opinión, la actitud y los estados emocionales de las personas, esto relacionado a la coyuntura del país podemos llamarla opinión pública. Las opiniones de las personas pueden ser positivas, negativas o neutrales.
- Para realizar el análisis de sentimiento de los tweets, el sistema propuesto primero extrae las publicaciones de twitter por usuario. El sistema también puede calcular la frecuencia de cada palabra en un tweet. El uso del enfoque de un aprendizaje supervisado, como lo es SVM, ayuda a obtener los resultados.
- El usar el corpus del TASS como muestra de entrenamiento no ayuda mucho en la precisión ya que esta data no está clasificada por un humano, y aun así si fuese hecha por un humano aquí entraría a tallar el punto de vista de cada quien, por este motivo es que el análisis de sentimiento en español es una ciencia que aún está en desarrollo.

- Twitter es una gran fuente de datos, lo que lo hace más atractivo para realizar análisis de los sentimientos. Realizamos un análisis de alrededor mes y medio de recolección de tweets con las palabras claves obtenidas del web scraping realizado pero nos afecta no ser un usuario premium en la API de Twitter por las limitaciones que esto conlleva.
- Encontramos que en los medios digitales las palabras con mayor frecuencia son las relacionadas al fútbol ya sea nacional o internacional es lo que más vende en nuestra sociedad, por eso tratamos de quitar estas palabras y encontrar resultados distintos.
- Para llegar a concluir el objetivo planteado en esta investigación necesitamos aún más tiempo de recolección para analizar un patrón en los sentimientos reflejados de la opinión pública.
- Otro factor que no consideramos en el análisis de sentimiento es el hecho de las fallas ortográficas y el uso de muchas jergas que comúnmente se dan en nuestra sociedad lo cual afecta la predicción.

5.2. Trabajo Futuro

En este punto se mencionan las mejoras que podemos realizar a esta trabajo.

- Pasar de un clasificador binario a uno multiclasificación para valores como neutral, muy positivo, positivo, negativo y muy negativo.
- Podemos incluir otra rama del análisis de sentimiento como es la categorización por temas lo cual ayudará a relacionar mejor los temas que promueven los medios digitales con las opiniones que vemos en Twitter.
- Hacer uso de focused scrawler para no solo centrarnos en el medio más influyente si no hacer una búsqueda en toda la web.
- Crear un grupo interdisciplinario para el aporte de más indicadores que desconocemos por no ser afín a nuestra carrera y poder mejorar el estudio realizado.

- Aplicar el análisis de sentimiento a los títulos de las noticias y así se podría generar publicidad de acuerdo a la polaridad del tema en cuestión.

Bibliografía

- [1] R. Rubio García, "Twitter y la teoría de la agenda-setting: mensajes de la opinión pública digital," *Estudios sobre el mensaje periodístico*, vol. 20, no. 1, pp. 249–264, 2014.
- [2] Y. P. Pratiksha and S. H. Gawande, "A comparative study on different types of approaches to text categorization," *International Journal of Machine Learning and Computing*, vol. 2, 2012.
- [3] TASS-2018: Workshop on Semantic Analysis at SEPLN. <http://www.sepln.org/workshops/tass/2018/>.
- [4] S. Joshi and D. Deshpande, "Twitter sentiment analysis system," *International Journal of Computer Applications*, vol. 180, 2018.
- [5] M. G. Ugur Kursuncu, U. Lokala, A. S. Krishnaprasad Thirunarayan, and B. Arpinar, "Predictive analysis on twitter: Techniques and applications," 2018.
- [6] P.-H. W. Miyoung Han and P. Senellart, "Focused crawling through reinforcement learning," 2018.
- [7] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
- [8] Twitter API. <https://developer.twitter.com/en/docs>.
- [9] Tweepy: An easy-to-use Python library for accessing the Twitter API. <https://tweepy.readthedocs.io/en/v3.5.0/>.

- [10] J. D. Hunter, "Matplotlib: A 2d graphics environment," *Computing In Science & Engineering*, vol. 9, no. 3, pp. 90–95, 2007.
- [11] AWS, "Overview of Amazon Web Services (White Paper)." <https://d1.awsstatic.com/whitepapers/aws-overview.pdf>, April 2017.
- [12] A. Mehlührer, "Web scraping: A tool evaluation," Master's thesis, Wien University, 2009.
- [13] *Beautifulsoup4*. <https://pypi.org/project/beautifulsoup4/>.
- [14] *Requests: HTTP for Humans*. <http://docs.python-requests.org/en/master/>.
- [15] *Textblob*. <https://textblob.readthedocs.io/en/dev/>.
- [16] S. Baccianella, A. Esuli, and F. Sebastiani, "Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining," Master's thesis, Istituto di Scienza e Tecnologie dell'Informazione, 2013.
- [17] *Open Street Map*. <https://www.openstreetmap.org/>.