UNIVERSIDAD DE CARABOBO
FACULTAD EXPERIMENTAL DE CIENCIAS Y TECNOLOGÍA
DIRECCIÓN DE POSTGRADO
ESPECIALIZACIÓN EN DESARROLLO DE SOFTWARE

SISTEMA DE RECOLECCIÓN Y PROCESAMIENTO DE DATOS EN *TWITTER* PARA EL ANÁLISIS DE SENTIMIENTO

AUTOR: Ing. Augusto González
TUTOR: Dra. Desirée Delgado

Trabajo Especial de Grado presentado
para optar al título de Especialista en Desarrollo de Software

Valencia, Marzo de 2022

UNIVERSIDAD DE CARABOBO
Facultad Experimental de Ciencias y Tecnología
Dirección de Postgrado FACYT

ACTA VEREDICTO DEL TRABAJO ESPECIAL DE GRADO
PROGRAMA: ESPECIALIZACIÓN EN DESARROLLO DE SOFTWARE

Quienes suscribimos, profesores **Desirée Delgado** C.I.V-7.352.958, **Mirella Herrera** C.I.V-8.044.677 y **Francisca Grimón** C.I.V-5.521.244, integrantes del Jurado designado Consejo de Postgrado de la Facultad Experimental de Ciencias y Tecnología (FACYT) de la Universidad de Carabobo, en su reunión ordinaria virtual No 03/2022 de fecha 28/03/2022, para considerar y evaluar el Trabajo Especial de Grado titulado **"SISTEMA DE RECOLECCIÓN Y PROCESAMIENTO DE DATOS EN TWITTER PARA EL ANÁLISIS DE SENTIMIENTO",** el cual fue presentado por el Ing. **Augusto Gonzalez, C.I. V-19.862.665,** en el IV Congreso Nacional de Investigación e Innovación en Ciencias Económicas y Sociales," Hacia la Nueva Visión del Planeta, para la obtención de su grado académico de Especialista en Desarrollo de Software, todo ello conforme a lo estipulado en el artículo 136 del Reglamento de los Estudios de Postgrado de la Universidad de Carabobo (REPUC). Dejamos constancia de lo siguiente:

1. El trabajo presenta un contexto centrado en conceptos y herramientas de la cibermetría y la webmetría, enfocado en la oferta de consultas dirigidas a usuarios investigadores o interesados en el estudio de fenómenos sociales, a partir de las opiniones expresadas por los usuarios de la red social Twitter, específicamente en el análisis de sentimiento.
2. Se aplicaron conceptos propios del cuerpo de conocimiento de la ingeniería de software como meta ulterior del programa de Especialización. Asimismo, desde el punto de vista metodológico, se cumplieron ciclos de la metodología Investigación-Acción, en conjunción con el uso de SCRUM en el desarrollo de software.
3. El trabajo realizado representa una posibilidad para el planteamiento de futuros proyectos en el área.

Emitimos por unanimidad el veredicto de **APROBADO** al trabajo sometido a nuestra consideración, todo conforme a lo dispuesto en las Normas para la Elaboración, Presentación y Evaluación del Trabajo de Especialización de la Facultad Experimental de Ciencias y Tecnologíade la Universidad de Carabobo.

En fe de todo lo cual levantamos y firmamos la presente acta de veredicto, el veintiocho del mes de marzo de dos mil veintidós.

Prof. DESIRÉE DELGADO
C.I. Nro. V-7.352.958
Tutor - Coordinador del Jurado
Dpto. de Computación FaCyT-UC

Prof. MIRELLA HERRERAC.I.
Nro. V-8.044.677
Miembro del Jurado
Dpto. de Computación FaCyT

Prof. FRANCISCA GRIMON
C.I. Nro. V-5.521.244
Miembro del Jurado
de Computación FaCyT-UC

Sistema de recolección y procesamiento de datos en *Twitter* para el análisis de sentimiento.

Augusto González[1,2], Desirée Delgado[3]

[1] correspondencia: aagg83556@gmail.com

[2] estudiante de Postgrado, FACYT-UC

[3] profesora del programa de postgrado FACYT-UC

Resumen

Las redes sociales son una de las herramientas utilizadas para comunicar y emitir opiniones. Una de las más usadas para emitir opinión es *Twitter*, en esta se encuentran millones de datos acerca de usuarios que comentan sobre cualquier tema. Esta data puede ser usada para generar tendencias, predicciones y tomar decisiones en función de lo que expresa, en lenguaje natural un sector de la población, acerca de un tema determinado. El propósito principal de esta investigación consistió en desarrollar una herramienta de software que permita recolectar datos desde *Twitter*, sobre opiniones expresadas (positiva, negativa, neutra o mixta) con respecto a un tema y en una región geográfica y ofrecerlos a investigadores para su procesamiento. La API extrae los datos y los analiza utilizando *Sentiment Analysis* agregando el sentimiento al tuit y vacía en un archivo Excel los datos clasificados. El usuario interactúa con la herramienta a partir de una interfaz sencilla en la que ingresa palabras clave y región geográfica para la búsqueda, y obtiene por pantalla el listado de opiniones, un cuadro resumen y la posibilidad de descargarlos en un archivo tipo Excel. La metodología utilizada fue la investigación acción participativa y para el desarrollo del software *SCRUM*. En los resultados se detectaron discrepancias en tuits que son irónicos o con nombres de usuarios iguales a la palabra clave, con lo cual se abre la posibilidad de nuevos trabajos que pudieran extraer un mayor número de tuits por petición a un almacenamiento para descartar aquellos duplicados y cuentas falsas.

Palabras clave: Análisis de sentimiento, *Twitter*, aprendizaje automático, procesamiento del lenguaje natural.

Introducción

En el año 2020 la manera en como las personas interactúan hizo un cambio radical debido a la pandemia mundial. Esto creo nuevos hábitos, nuevas maneras de trabajar y sobre todo aceleró el crecimiento del mundo digital. Por eso las redes sociales han tenido un crecimiento exponencial; el 55.1% de la población mundial usa las redes sociales activamente (Shum, 2021). Esto hace que exista una fuente de información directa sobre la opinión de dicha población, teniendo en consideración de que el 67.1% de la población mundial tiene acceso a un teléfono inteligente (Shum, 2021).

Se han realizado diversos estudios para explorar la posibilidad que nos ofrecen las redes sociales como *Twitter*, con el fin de analizar y usar técnicas del procesamiento del lenguaje natural para conocer la opinión de los usuarios. En este contexto, al *análisis de sentimiento* es un campo que se viene estudiando décadas atrás, por mencionar el trabajo de (Pang, Lee, & Vaithyanathan, Thumbs up? Sentiment Classification using Machine Learning Techniques, 2002) o de (Turney, 2002), en donde se muestran los primeros avances en el análisis de sentimiento. En los trabajos anteriores y en general el análisis se ha realizado para expresiones en el idioma inglés, sin embargo, en el idioma español se requiere reforzar las investigaciones, a fin de generar modelos de aprendizaje automático, considerando la complejidad que representa el idioma en cuanto a expresiones idiomáticas y modismos.

En este sentido, la investigación se propuso determinar el sentimiento general sobre una palabra clave en el idioma español utilizando los Trinos de la red *Twitter*, para una región geográfica determinada. Como consecuencia, la herramienta se enfoca en determinar la opinión caracterizada como positiva, negativa, neutra o mixta; sobre un producto, servicio, situación, partido político, fenómeno natural, entre otros; haciendo de ésta una herramienta que pueda ser utilizada para diversos fines en diferentes contextos.

Marco Teórico

Para contextualizar la investigación, a continuación se presentan un conjunto de trabajos en el área, así como los términos más importantes para su compresión.

Trabajos Relacionados

Entre los trabajos más destacados, se tienen:

- Trabajo Especial de Grado de Casaverde y Olarte (2020), titulado **"Análisis masivo de datos en *Twitter* para identificación de opinión"**, realizado en la Universidad Nacional de San Antonio Abad del Cusco, desarrolló un sistema para filtrar y hacer un análisis de sentimientos en los tuits usando aprendizaje automático. La implementación de máquinas de soporte vectorial para determinar opiniones constituyó un aporte fundamental para entender validaciones y herramientas.
- El trabajo de maestría de Sobrino Sande (2018), titulado **"Análisis de sentimientos en *Twitter*"**, realizado en la Universidad Oberta de Cataluña presenta un mapeo sistemático de la literatura hasta la fecha, en el área de análisis de sentimiento. Asimismo, desarrolla una solución propia de clasificador de sentimientos usando algoritmos de aprendizaje supervisado. Esta investigación permitió comprender la evolución del análisis de sentimiento en la red social *Twitter*.
- El artículo de Tapia, Aguinaga, y Luje (2018), titulado **"Detección de patrones de comportamiento a través de Redes Sociales como *Twitter*, utilizando técnicas de Minería de Datos como método para detectar el Acoso Cibernético"**, publicado en la *Conference On Software Process Improvement* (CIMPS), realiza un estudio de análisis de sentimiento en un tema particular como el ciber acoso en un medio como la red social *Twitter*. Este estudio contribuyó principalmente en la determinación del tipo de herramienta y arquitectura de software a utilizar.

Definiciones

- **Procesamiento del Lenguaje Natural:** Según Cortez Vásquez, Vega Huerta, Pariona Quispe, & Huayna (2009) y Hernández y Gómez (2013), conocido también por sus siglas en inglés *Natural Language Processing* (NLP), es un campo de las ciencias de la computación, de la inteligencia artificial y de la lingüística que estudia las interacciones entre las computadoras y el lenguaje humano. Se ocupa de la formulación e investigación de mecanismos eficaces computacionalmente para la comunicación entre personas y máquinas por medio del lenguaje natural, es decir, de las lenguas del mundo. Entre las aplicaciones del procesamiento del lenguaje natural mas populares e importantes tenemos: recuperación de información, traducción automática de textos, reconocimiento del habla, extracción de información y análisis de sentimientos.
- **Análisis de sentimiento:** Es un campo de investigación dentro del PNL que trata de extraer de manera automática y mediante técnicas computacionales, información subjetiva

expresada en el texto de un documento dado y acerca de un determinado tema. De esta forma, mediante el análisis de sentimientos podremos saber si un texto presenta connotaciones positivas o negativas. Una definición ampliamente extendida de este concepto es la ofrecida por los investigadores (Pang & Lee, Opinion Mining and Sentiment Analysis, 2008) y que define análisis de sentimientos como: *"Tratamiento computacional de opiniones, sentimientos y subjetividad en textos"*.

- **Aprendizaje automático:** Es el estudio de algoritmos computacionales que mejoran mediante la experiencia, es decir, que pueden aprender y mejorar sus conocimientos previamente definidos. El aprendizaje automático es un subconjunto de la inteligencia artificial, cuyos algoritmos construyen un modelo matemático basado en datos de prueba, conocidos como "datos de entrenamiento" con fines de predicción y toma de decisiones (Bishop, 2006).
- **API (***Application Programming Interfaces):* Es un conjunto de definiciones y protocolos utilizados para desarrollar e integrar el software de las aplicaciones, permitiendo la comunicación a través de un conjunto de reglas. Así pues, una API es una especificación formal que establece cómo un módulo de un software se comunica o interactúa con otro, para cumplir una o muchas funciones. Todo ello dependiendo de las aplicaciones que las utilicen y los permisos que les dé el propietario de la API a los desarrolladores de terceros (Fernández, 2019).

Metodología

La investigación-acción es una metodología que busca la interpretación de un aspecto social a través del uso de la investigación activa, constante y colaborativa de investigadores, con la intención de provocar un cambio positivo de orden social, en donde se involucra la toma de decisiones. Esta definición implica combinar la práctica con la teoría, así como cuatro fases o momentos diferenciados: observación, participación, planificación y la reflexión (Velásquez & Cádiz, 2011). Cada uno implica una mirada retrospectiva, y una intención prospectiva que forman una espiral auto reflexiva de conocimiento y acción. La Figura 1 a continuación, muestra los momentos de la investigación-acción, de acuerdo con Latorre (2003).

Figura 1. Ciclos de la investigación-acción. Fuente: *(Latorre, 2003)*.

En el contexto del trabajo realizado, cada una de las fases de la metodología conllevó a un conjunto de actividades y entregables, según se señalan en la Tabla 1 siguiente:

Tabla 1. Actividades del proyecto, según la metodología investigación-acción.

Fases metodología investigación-acción	
Planificación	Investigar referencias sobre modelos de extracción de datos, filtrado de información relevante, análisis de sentimiento, algoritmos de aprendizaje automático y plataformas en la nube.
	Determinar los requisitos funcionales y no funcionales
	Seleccionar metodología de desarrollo de software (*SCRUM*)
	Realizar el plan de trabajo
Acción	Implementar la herramienta siguiendo la metodología de sw
	Iterar sobre alternativas y mejoras en el análisis de sentimiento
	Realizar pruebas de funcionalidad y rendimiento
Observación	Presentar resultados y análisis de las pruebas
Reflexión	Realizar sugerencias para futuras versiones

Fuente: Autor.

Tal como se muestra, en la Fase de Planificación se integró *SCRUM***.** El nombre *SRUM*, no es un acrónimo, proviene del *Rugby* y significa melé, y al igual que en *Rugby*, todos los jugadores actúan como una unidad para hacer avanzar la pelota. En *SCRUM* todos los componentes de un equipo realizan actividades de forma iterativa e incremental para desarrollar el proyecto. *SCRUM* es una metodología de desarrollo ágil simple, que se adapta continuamente a la evolución del proyecto. En ésta se priorizan los objetivos/requisitos en función de las necesidades del cliente y lo esencial para iniciar el desarrollo. Sigue un proceso incremental basado en iteraciones y revisiones, orientándose más a las personas que a los procesos (Gorakavi, 2009; Hundermark, 2009).

Más específicamente, *SCRUM* se basa en ciclos de trabajo llamados *Sprints*, los cuales son iteraciones de 1 a 4 semanas de forma consecutiva, con una duración fija y con fechas de culminación previamente establecidas. Cada *Sprint* además, está conformado por los requerimientos a desarrollar, a partir de una lista priorizada. Cabe señalar, que esta metodología incluye una serie de eventos y ceremonias que permiten mantener al equipo de trabajo integrado y enfocado en el proceso de desarrollo, en sus diversos momentos. La Figura 2 presenta un ciclo o *sprint* de *SCRUM*

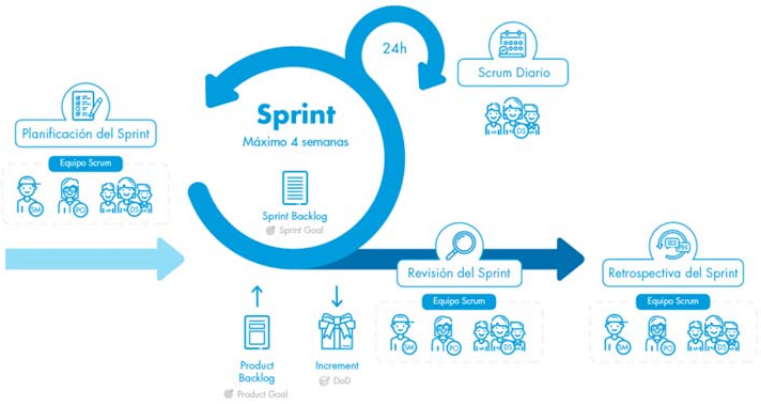

Figura 2. Ciclo de *SCRUM*. Fuente: (Rodríguez & Álvarez, 2020)**.**

Resultados

A continuación se presentan los resultados, organizados en *sprints* con una duración promedio de 3 semanas cada uno.

***Sprint* 0:**

Definición de Requisitos: En este ciclo se definieron los requisitos funcionales y no funcionales del proyecto (Tablas 2 y 3 respectivamente)

Tabla 2. Requisitos Funcionales

Código	Requerimiento	Prioridad
REQF01	Permitir buscar información en *Twitter* ingresando una palabra clave	Esencial
REQF02	Permitir filtrar la información de *Twitter* por ubicación geográfica con el nombre de una ciudad o región	Esencial
REQF03	Clasificar cada tuit de acuerdo al sentimiento que expresa en positivo, negativo, mixto o neutro	Esencial

REQF04	Desplegar los datos de cada tuit en un archivo en formato texto para su posterior análisis	Media
REQF05	Desplegar un resumen de los datos recolectados en términos de porcentaje por categoría de sentimiento	Media

Fuente: Autor

Tabla 3. Requisitos No Funcionales

Código	Requerimiento
REQNF01	La herramienta debe ser capaz de extraer la data y analizarla con un tiempo de respuesta menor a 5 minutos
REQNF02	La interfaz debe ser fácil de usar
REQNF03	El formato de salida debe ser un archivo tipo Excel para su posterior procesamiento

Fuente: Autor

Modelos de extracción de datos de la red social *Twitter*:

Durante la investigación de modelos de extracción de datos, filtrado y análisis de sentimiento en la red social *Twitter*, se determinó que la forma más sencilla de extraer datos es mediante el uso de *Twitter Api*, la cual se describe a continuación

***Twitter Api*:** Permite acceder a datos y escribirlos en Twitter, sin embargo, su acceso es restringido y limitado en el caso de requerir más de 500K tuits por mes (*Twitter*, Inc.). Por esta razón, se gestionó una solicitud especial de uso con fines académicos en el portal de *Twitter*, la cual fue respondida en 2 semanas aproximadamente.

Una vez obtenido el acceso requerido, se evaluó la documentación encontrándose referencias al uso de *Tweepy* como librería de terceros desarrollada en *Python*, que facilitaban la conexión a *Twitter Api*. Cabe señalar, que dicha librería cuenta con una extensa comunidad de desarrolladores y amplia documentación. Aunado a lo anterior, se seleccionó *Flask* como *framework* de desarrollo en *Python*, por sus características de flexibilidad y facilidad de uso (Dev.to, 2019).

Al final de este *sprint*, también se analizaron los modelos de aprendizaje automático, así como el procesamiento de una frase para determinar el análisis de sentimiento.

***Sprint* 1:**

En este *sprint* se realizó una revisión de las diferentes herramientas y librerías *open source* existentes en el mercado, para evaluar el análisis de sentimiento.

- Librería "*sentiment spanish":* Cuenta con licencia MIT (Bello, 2020) y trabaja a partir de un modelo pre entrenado en español. En el análisis se realizaron un conjunto de pruebas a fin de determinar la velocidad y eficiencia, resultando un tiempo de respuesta promedio entre 5 y 6 segundos por frase. Específicamente, en las pruebas se utilizó la frase de 4 palabras: *"La playa es lo mejor",* ejecutada en una computadora con procesador i9. El equipo presentó un recalentamiento importante, por lo que al seguir investigando se halló que por lo general este tipo de librerías es ejecutada en equipos especializados o en su defecto equipos en la nube de algún proveedor como *Amazon, Microsoft, IBM o Google*, con sus consecuentes costos por servicios de cómputo.
- API de análisis de sentimiento: Se realizó un análisis de las distintas alternativas en la nube, considerando: costos por servicio, rendimiento y usabilidad (Tabla 4). Entre las opciones: *Amazon Comprehend, Google Cloud Natural Language, IBM Watson, Microsoft Azure Cognitive Services*. Cabe señalar que algunas proveen una *free tier o* capa gratuita de peticiones por mes, haciendo factible este estudio enmarcado en el ámbito académico (Deducive, 2018).

Tabla 4. Soluciones API en la nube para el análisis de sentimiento (AS)

Variables	Archivos	Caract/ seg	Amazon	Google	IBM	Microsoft
Opinión AS	10.000	550	$6	$10	$30	$75
AS *Twitter*	1.000.000	75	$100	$1000	$1000	$2500
Tema doc AS	5.000	10.000	$51	$100	$30	$75

Fuente: (Deducive, 2018)

Los factores claves considerados en el análisis fueron el costo y la usabilidad. Por razones de costos se descartaron *Microsoft e IBM*, a pesar de que *IBM* (con la *IA Watson*), es preciso y ofrece un precio competitivo en cuanto al análisis de temas o tópicos, sin embargo, para opiniones de clientes con pocos archivos (menos de 10.000 tal como el caso que nos ocupa) resulta más costoso.

Las nubes de *Amazon y Google* son las más usadas y a pesar de que Amazon es la de menor costo; requería una curva de aprendizaje más prolongada por lo que se decidió utilizar *Google Cloud Natural Language.* Adicionalmente, ésta última cuenta con una extensa documentación y resultó fácil de usar.

Para finalizar este *sprint*, se realizaron pruebas de rendimiento en el equipo i9, la API de *Google* y con la misma frase del experimento anterior (*"La playa es lo mejor")*, resultando un tiempo de respuesta promedio menor a 0.6 segundos, lo cual representó una reducción drástica del tiempo en casi 10 veces y sin ningún recalentamiento del procesador.

***Sprint* 2:**

En este *sprint* se plantea la arquitectura como se muestra en la Figura 3

Figura 3. Diagrama de arquitectura. Fuente: Autor.

Vista del Servidor: Se desarrolló en *Flask*, el cual permite consultar a la API de *Twitter* y extraer los últimos 100 tuits (limitación de tuits por petición) relacionado con una palabra clave. Como entregable, *Twitter* devuelve la data para una ubicación geográfica y sin ningún orden de relevancia. Como limitación clave está el número de peticiones por intervalo de tiempo, la cual es 450 peticiones en 15 minutos.

Una vez obtenida la data, se ejecuta la API de *Google Cloud*, la cual procesa cada tuit para analizar el tipo de sentimiento (positivo, negativo y neutro o mixto). Esta API está disponible para 15 idiomas y tiene reconocimiento automático del mismo, haciendo posible la escalabilidad del proyecto para múltiples idiomas.

Para restringir la búsqueda de la frase a una ubicación geográfica específica, se utilizó la API de *Google* de geolocalización, la cual dada una palabra clave, ejecuta *Google Street.*

***Sprint* 3:**

Vista del cliente: El cliente consta de una aplicación web en *React* con *Typescript*, la cual es una de las librerías más usadas en la actualidad (Stack Overflow, 2021). La interfaz inicial se muestra en la figura 4 a continuación:

Figura 4. Interfaz de búsqueda. Fuente: Autor.

Es una interfaz sencilla que presenta un campo para introducir la palabra clave en idioma español y luego la ubicación geográfica en términos de ciudad o población. El resultado corresponderá a los últimos 100 tuits relacionados con esa palabra clave en la región indicada. La respuesta se muestra en la figura 5, a continuación:

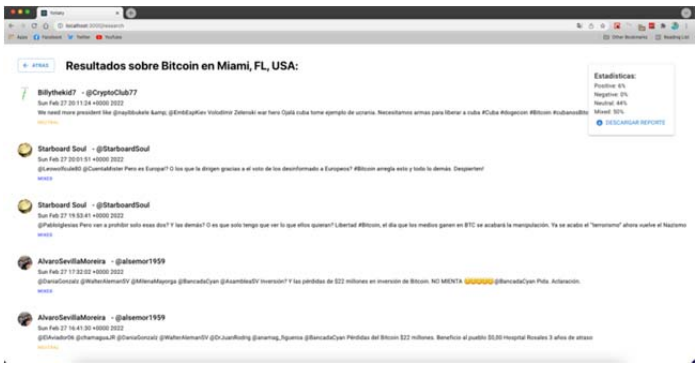

Figura 5. Interfaz de resultados. Fuente: Autor.

La pantalla muestra una lista de todos los tuits, con las siguientes propiedades: nombre, usuario, fecha, foto de perfil, tuit y una etiqueta clasificándolo como positivo, negativo, mixto o neutro. En la parte superior derecha aparece un recuadro con los resultados generales expresados en términos de porcentaje, de los últimos 100 tuits, clasificándolos como positivo, negativo, mixto o neutro. Y una opción dentro del mismo recuadro que permite descargar un archivo Excel con las columnas: “Date”, “Tweet”, “Url”, “Fullname”, “Username” y “Sentiment”, tal como lo muestra el formato de la figura 6 a continuación:

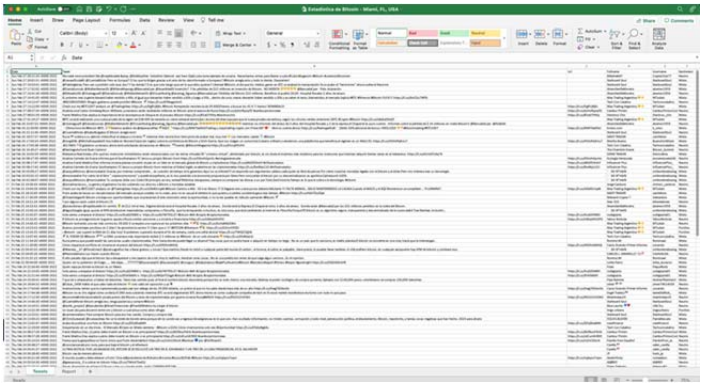

Figura 6. Excel resultante para Bitcoin en Miami, FL, USA. Fuente: Autor.

***Sprint* 4:**

Este *sprint* se enfocó en la interpretación de los resultados, así como las mejoras en cuanto a detalles de interfaz.

Interpretación de resultados en *Google Sentiment Analysis:* Para el análisis se utilizó la documentación de *Google Cloud* (Google, n.d.), conformado por valores numéricos *score* y *magnitude*. A continuación, un ejemplo de una corrida del *analyzeSentiment*:

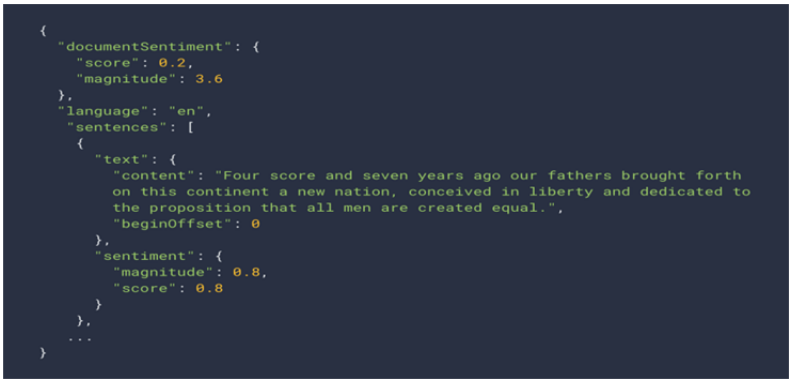

Figura 7. Formato de respuesta API de *Google Sentiment Analysis.*

Fuente: (Google, n.d.).

Estos valores de campo se describen a continuación:

- *documentSentiment* contiene la opinión general del documento, que consta de los siguientes campos:
  - *score* de las opiniones oscila entre -1.0 (negativo) y 1.0 (positivo), y corresponde a la tendencia emocional general del texto.

    - *magnitude* indica la intensidad general de la emoción (tanto positiva como negativa) en un determinado texto, entre 0.0 y +inf. A diferencia de *score*, *magnitude* no está normalizada; cada expresión de emoción en el texto (tanto positiva como negativa) contribuye a *magnitude* de este (por ende, las magnitudes podrían ser mayores en bloques de texto más extensos).
- *language* contiene el idioma del documento, ya sea que se haya pasado en la solicitud inicial o se haya detectado automáticamente si estaba ausente.
- *sentences* contiene una lista de las oraciones extraídas del documento original
    - *sentiment* contiene los valores de nivel de opiniones de la oración adjuntos a cada oración, conformado por los valores *score* y *magnitude* descritos.

Un valor de respuesta del ejemplo anterior de 0.2 de puntuación, indica que un documento es ligeramente positivo en emoción, mientras que un valor de magnitud de 3.6 indica un documento relativamente emotivo, dado su tamaño pequeño (alrededor de un párrafo). Por lo tanto, la primera oración de la dirección de Gettysburg contiene un score altamente positivo de 0.8 (Google, n.d.).

Teniendo esto en consideración, se puede observar que existen textos que se identifican como neutros o mixtos, lo cual no proporciona indicios claros de un sentimiento positivo o negativo. Más aún, en el lenguaje hay modismos, acrónimos y emojis; que vuelven más compleja la tarea de medir un sentimiento (Cogito, 2021), incluso utilizando algoritmos sofisticados entrenados por *Google*, se observa que se requiere seguir avanzando en las interpretaciones del lenguaje (Darr, 2019).

Otro aspecto a destacar es la cantidad de ruido que presenta *Twitter*, tales como, *bots* de noticias y cuentas que tienen cierta popularidad como portales de noticias o personas influyentes, lo cual hace más compleja la tarea de extraer datos y esto teniendo en consideración que *Twitter* tiene un sistema de autenticación de cuentas y políticas *anti-bot*. La determinación de este tipo de distorsiones es una tarea importante para los investigadores en este campo ya que posibilitaría el incremento en el nivel de confianza de los datos.

Conclusiones

Este trabajo de investigación se adentró en un tema de actualidad que involucra la red social *Twitter* y el análisis de sentimiento como forma de expresión categorizada como positivo,

negativo, mixto o neutro, a partir del análisis resultante de herramientas automatizadas, basadas en el aprendizaje de máquina. La herramienta desarrollada ofrece un uso potencial para la realización de estudios sociales, económicos y políticos; y en general de cualquier tema en tendencia.

Un complemento importante es la posibilidad de filtrar los datos obtenidos por ubicación geográfica, lo que permite realizar estudios segmentados por lugar o región, sin embargo, esto pudiera ser también una desventaja para aquellos lugares donde el uso de *Twitter* no esté tan extendido.

Otro hallazgo importante en los resultados, es el impacto que tiene la interpretación de los datos como base para medir con cierto nivel de confianza, el análisis de sentimiento. Cabe destacar que esta es una tarea compleja incluso para la mente humana, a pesar de tener el entendimiento del lenguaje, algunas opiniones pueden resultar ambiguas y generar tanto sentimientos positivos como negativos. Como resultado de las pruebas preliminares en el proyecto, el algoritmo utilizado por la librería de *IBM Watson* arrojó mayor precisión al identificar sentimientos sobre diversos temas, ya que no solamente indica cual es el sentimiento general, sino también el sentimiento sobre un tópico. Una de las limitaciones de la API de *Google Cloud* es que la información se presenta como algo global y no acerca de la palabra clave. Otro factor importante para considerar es la cantidad de ruido en la data de *Twitter*, por lo que integrar la posibilidad de realizar un filtrado sería una mejora a considerar para una próxima iteración.

Considerando un posible escalamiento para un uso más robusto, se recomienda integrar *Amazon Comprehend* y comparar los resultados de análisis, ya que *Amazon* también ofrece una capa gratuita y los costos son menores para volúmenes más grandes de datos.

Para finalizar, se concluye que los objetivos del proyecto fueron logrados y que se coloca a disposición de investigadores una herramienta que les permitirá realizar estudios de tendencia, predicción y toma de decisiones, a partir del análisis de sentimiento en campos tan diversos como estudios sociales, políticos, económicos y de mercado, entre muchos otros.

Referencias Bibliográficas

Bello, H. (14 de Abril de 2020). Recuperado el Febrero de 2022, de https://github.com/sentiment-analysis-spanish/sentiment-spanish

Bishop, C. (2006). *Pattern Recognition and Machine Learning.* Singapore: Springer.

Casaverde, A., & Olarte, A. (2020). Análisis Masivo de datos en *Twitter* para identificación de opinión.

Cogito. (3 de Enero de 2021). Obtenido de https://www.cogitotech.com/blog/sentiment-analysis-types-how-it-works-why-difficult

Cortez Vásquez, A., Vega Huerta, H., Pariona Quispe, J., & Huayna, A. M. (30 de Diciembre de 2009). Procesamiento de lenguaje natural . *Revista De investigación De Sistemas E Informática*, 45-54.

Darr, S. (25 de Septiembre de 2019). Recuperado el Febrero de 2022, de https://www.impression.co.uk/blog/testing-sentiment-with-googles-natural-language-api/

Deducive. (29 de Agosto de 2018). Recuperado el Febrero de 2022, de https://www.deducive.com/blog/2018/8/29/how-much-does-sentiment-analysis-in-the-cloud-actually-cost

Dev.to. (18 de Noviembre de 2019). Recuperado el Febrero de 2022, de https://dev.to/detimo/python-flask-pros-and-cons-1mlo

Fernández, Y. (23 de Agosto de 2019). *Xataka.* Obtenido de https://www.xataka.com/basics/api-que-sirve

Google. (s.f.). Recuperado el Enero de 2022, de Interpretar los valores de análisis de opiniones: https://cloud.google.com/natural-language/docs/basics#interpreting_sentiment_analysis_values

Gorakavi, P. (2009). Build your Project using Agile Methodology.

Hernández, M., & Gómez, J. (31 de Julio de 2013). Aplicaciones de Procesamiento de Lenguaje Natural. *Revista Politécnica, 32*, 96-96.

Hundermark, P. (Noviembre de 2009). Un mejor Scrum -Un conjunto no oficial de consejos e ideas sobre cómo implementar Scrum. Ciudad de Cabo.

Latorre, A. (2003). *La investigación-acción: Conocer y cambiar la práctica educativa.* Barcelona: Editorial Graó.

Pang, B., & Lee, L. (January de 2008). Opinion Mining and Sentiment Analysis. *Foundations and Trends in Information Retrieval, 2*(1-2), 1-135.

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? Sentiment Classification using Machine Learning Techniques. *Conference on Empirical Methods in Natural Language Processing* (págs. 79-86). Association for Computacional Linguistics.

Rodríguez, M., & Álvarez, A. (2020). Scrum: el pasado y el futuro. *Revista Agilidad Empresarial*.

Shum, Y. M. (2 de Mayo de 2021). Obtenido de https://yiminshum.com/internet-social-media-2021/

Sobrino Sande, J. C. (Junio de 2018). Análisis de Sentimientos en *Twitter*.

Stack Overflow. (Mayo de 2021). Obtenido de https://insights.stackoverflow.com/survey/2021#section-most-popular-technologies-web-frameworks

Tapia, F., Aguinaga, C., & Luje, R. (2018). Detection of Behavior Patterns through Social Networks like *Twitter*, using Data Mining techniques as a method to detect Cyberbullying. *7th International Conference On Software Process Improvement (CIMPS)* (págs. 111-118). Guadalajara: IEEE.

Turney, P. D. (2002). Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, (págs. 417-424). Philadelphia.

Twitter, Inc. (s.f.). Recuperado el Febrero de 2022, de https://developer.twitter.com/en/docs/twitter-api/getting-started/getting-access-to-the-twitter-api

Velásquez, C., & Cádiz, A. (Febrero de 2011). Recuperado el Febrero de 2022, de https://www.slideshare.net/ajavier81/metodologa-investigacin-accion-ia/10

Valencia, Venezuela, IVCNICEyS-2021

Ciudadano(a) Investigador(a)
Desirée Delgado y Augusto González
Presente

Nos complace informar a usted que el trabajo titulado **«Sistema de recolección y procesamiento de datos en Twitter para el análisis de sentimiento»,** fue **APROBADO** luego de realizado el arbitraje doble ciego juicio de pares, para ser **presentado oralmente** en el **IV Congreso Nacional de Investigación e Innovación en Ciencias Económicas y Sociales "Hacia La Nueva Visión del Planeta",** organizado por la Facultad de Ciencias Económicas y Sociales de la Universidad de Carabobo. Los detalles sobre la presentación oral de su trabajo a través de ZOOM (día, hora) le serán comunicados oportunamente por la Comisión Académica en fecha previa al Congreso.

El Congreso se estará realizando del 29 al 31 de marzo de este año 2022 y su lema **"Hacia La Nueva Visión del Planeta**" da continuidad a la voluntad manifiesta de nuestra gestión decanal y directiva por transitar los senderos de construcción de un mundo mejor, donde haya posibilidad de ver alcanzados los **17 Objetivos de Desarrollo Sostenible ODS de la Agenda 2030 de la Organización de Naciones Unidas.**

Le expresamos el agrado de la Universidad de Carabobo a través de su Facultad de Ciencias Económicas y Sociales y el Comité Organizador del Congreso, por contar con su participación y le recordamos que este trabajo, previa revisión y aprobación del Comité de Publicaciones y realizados de su parte los ajustes necesarios recomendados por los evaluadores, podrá formar parte de los textos digitales editados por el Comité Editorial del congreso.

En Valencia a los 19 días del mes de marzo de 2022.

Atentamente,

Dr. Williams Aranguren
Coordinador Comisión de Arbitraje