

**UNIVERSIDAD POLITÉCNICA SALESIANA  
SEDE QUITO**

**CARRERA:  
INGENIERÍA DE SISTEMAS**

**Trabajo de titulación previo a la obtención del título de:  
Ingeniera de Sistemas**

**TEMA:  
DESARROLLO DE UN PROTOTIPO QUE EMPLEE LA PLATAFORMA  
LENGUAJE NATURAL DE GOOGLE CLOUD**

**AUTORA:  
AMANDA ELIZABETH CABASCANGO GARCIA**

**TUTOR:  
GUSTAVO ERNESTO NAVAS RUILOVA**

**Quito, febrero del 2020**

## CESIÓN DE DERECHOS DE AUTOR

Yo, Amanda Elizabeth Cabascango Garcia con documento de identificación N°.1724159957, manifiesto mi voluntad y cedo a la Universidad Politécnica Salesiana la titularidad sobre los derechos patrimoniales en virtud de que soy autora del trabajo de titulación con el tema: **DESARROLLO DE UN PROTOTIPO QUE EMPLEE LA PLATAFORMA LENGUAJE NATURAL DE GOOGLE CLOUD**, mismo que ha sido desarrollado para optar por el título de INGENIERA DE SISTEMAS en la Universidad Politécnica Salesiana, quedando la Universidad facultada para ejercer plenamente los derechos cedidos anteriormente.

En aplicación a lo determinado en la Ley de Propiedad Intelectual, en mi condición de autora me reservo los derechos morales de la obra antes citada. En concordancia, suscribo este documento en el momento que hago entrega del trabajo final en digital a la Biblioteca de la Universidad Politécnica Salesiana.



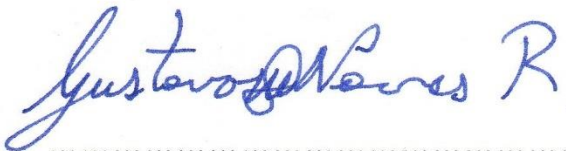
.....  
AMANDA ELIZABETH  
CABASCANGO GARCIA  
CI: 172415995-7

Quito, febrero del 2020

## DECLARATORIA DE COAUTORÍA DEL DOCENTE TUTOR

Yo declaro que bajo mi dirección y asesoría fue desarrollado el Artículo Académico, DESARROLLO DE UN PROTOTIPO QUE EMPLEE LA PLATAFORMA LENGUAJE NATURAL DE GOOGLE CLOUD realizado por Amanda Elizabeth Cabascango Garcia, obteniendo un producto que cumple con todos los requisitos estipulados por la Universidad Politécnica Salesiana, para ser considerados como trabajo final de titulación.

Quito, febrero 2020



.....  
GUSTAVO ERNESTO NAVAS RUILOVA

CI: 1705675625

## **Dedicatoria**

El presente artículo está dedicado para mi madre Eufemia, sé que tu partida fue muy temprana. Y desde el cielo te convertiste en un ángel que me ha cuidado y guiado en cada paso de la universidad. A mi padre Luis Eduardo, que dedicaste con mi mami muchas horas de trabajo duro, para apoyarme en mis estudios. A mi abuelita Carmen que es mi segunda madre, quien con su amor, paciencia y apoyo incondicional me han permitido llegar a cumplir esta meta. A mi hermano Misael por su apoyo y cariño, durante este proceso. Y a pesar de todas las adversidades que hemos pasado como familia, hemos salido adelante. Finalmente quiero dedicar este trabajo a mis amigos quienes se han convertido en personas importantes, demostrándome su lealtad y por el apoyo en cada momento, quienes sin esperar nada a cambio compartieron su conocimiento y fueron claves para cumplir esta etapa de mi vida.

Gracias a cada uno de ustedes, por ser una clave importante en esta etapa.

AMANDA ELIZABETH

# DESARROLLO DE UN PROTOTIPO QUE EMPLEE LA PLATAFORMA LENGUAJE NATURAL DE GOOGLE CLOUD

Amanda E. Cabascango<sup>1</sup>, Gustavo E. Navas<sup>2</sup>

## Resumen

El lenguaje natural en la actualidad se ha convertido una herramienta importante, donde nacen diversas aplicaciones las cuales cuyo objetivo es automatizar de cierta manera un proceso, convirtiéndose una herramienta esencial para el desarrollo de este prototipo. Debido a que Google Cloud una plataforma robusta la cual permite el desarrollo de dos prototipos por medio de lenguaje natural como herramienta para el desarrollo de análisis de opiniones y clasificación de textos ya sean por medio de librerías clientes, así como un modelo personalizado, el cual se puede uno crear acorde a la necesidad que se requiere clasificar.

Un análisis de opiniones dentro de la red social Twitter toma un rol importante para medir el nivel de impacto que está teniendo cierto hashtag y además la clasificación de textos para “Grounded Theory in Software Development” lo cual permitimos una clasificación automática basada en el modelo que sea necesario. Finalmente, como resultados se quiere ver que tan eficaz es el proceso de análisis para los dos casos anteriormente descritos.

**Palabras Clave:** Google Cloud, Lenguaje Natural, Twitter.

## Abstract

Natural language today has become an important tool, since several applications are born there, whose objective is to automate a process in a certain way, becoming an essential tool for the development of this prototype. Because Google Cloud is a robust platform which allows the development of two prototypes through natural language as a tool for the development of opinion analysis and text classification either through client libraries, as well as a personalized model, which one can create according to the need to classify.

An analysis of opinions within the social network Twitter takes an important role to measure the level of impact that a certain hashtag is having and also the classification of texts for “Grounded Theory in Software Development” which allows an automatic classification based on the model that be necessary. Finally, as results we want to see how effective the analysis process is for the two cases described above.

**Keywords:** Google Cloud, Natural Language, Twitter.

---

<sup>1</sup> Estudiante de Ingeniería de Sistemas – Universidad Politécnica Salesiana – Sede Quito. acabascango@est.ups.edu.ec

<sup>2</sup> Máster en Software Libre en la Universidad Abierta de Cataluña. gnavas@ups.edu.ec

## **1. Introducción**

La evolución de la información en los últimos tiempos por medio de la tecnología del internet, la misma que ha estimulado la proliferación de información; estas se pueden tener en varios campos desde bibliotecas virtuales hasta redes sociales, convirtiéndose en fuentes masivas. Como existe una gran cantidad de información y diversidad, por lo cual dentro de las bibliotecas virtuales se puede encontrar gran cantidad de datos científicos; mientras que en las redes sociales pueden tener información con multimedia generada por parte de los usuarios en tiempo real [1].

En este artículo se aborda la falta de automatización sin perder la esencia de razonamiento humano, por medio de un procesamiento de lenguaje natural (PLN). La cual permite establecer una comunicación entre humano y máquina [2] . Se considera el uso de Google Cloud Plataform (GCP) ya que unifica diferentes herramientas de desarrollo de varios campos. De acuerdo con el enfoque para el diseño de los prototipos propuestos. Por medio de la herramienta natural language se procesa los diferentes textos para el prototipo llamado *caso A* el cual maneja una cadena de caracteres con una longitud máxima de 280 llamado tweets; para el *caso B* maneja títulos de referencias bibliográficas. Se ofrece dos maneras de proceso: i) Por medio de librería cliente y ii) Un modelo personalizado. Las cuales son supervisadas de acorde a la necesidad del desarrollador.

Actualmente, las redes sociales se han tomado un rol importante, se considera para el estudio la red social twitter ya que se obtiene información por medio de los tweets en tiempo real, a nivel mundial [4]. Para el segundo caso se requiere un modelo de clasificación para las referencias bibliográficas, este proceso se lo realiza de manera manual; debido a que utiliza el método Systematic Mapping Studies (SMS) por parte del investigador.

El análisis de sentimientos dentro de los tweets permitirá ver principalmente el comportamiento de natural language de GCP versus a otra librería. Identificando el tipo de impacto que tiene los tweets incluyendo el hashtag #WWIII. Para el desarrollo del caso B se procedió por modelos de aprendizajes automatizados, desde los conjuntos de datos sin procesar hasta el modelo de entrenamiento llamado AutoML [4]. Además, al ofrecer una interfaz intuitiva permitiendo la carga de datos de preparación y así poder evaluar el modelo personalizado [5]; el desarrollo ambos prototipos por los diferentes procesos de GCP, permitiendo la aplicación de lenguaje natural con el fin de reducir la barrera de comunicación entre el hombre-máquina.

## **2. Trabajos relacionados**

Para el proceso de investigación, se partió indagando información sobre GCP y análisis de sentimientos. Se consideró como antecedentes los presentados a continuación:

En [6] habla sobre el procesamiento del lenguaje natural mediante el modelado de los procesos cognoscitivos. Entran en juego la comprensión del lenguaje para diseñar sistemas que realicen tareas lingüísticas complejas; como: traducción, resúmenes de textos y recuperación de información. El resultado a la investigación realizada se concluye que el lenguaje natural puede ser utilizado para analizar situaciones complejas, ya que en los lenguajes de programación manipulan información que son de consideraciones semánticos y pragmáticos.

En [7], cuyo objetivo principal es identificar las opiniones positivas o negativas generadas por medio de los usuarios. La problemática es debido a que los usuarios escriben en sus tweets sin ningún control hacia las diferentes opiniones que presenta cada uno de ellos a un tema referente. Análisis de sentimientos y la minería de opiniones han sido de gran utilidad dentro de la investigación. Dando como

resultado, la identificación de los tweets ya sean estos positivos o negativos.

Una publicación de un blog [8], resulta ser interesante dentro de la investigación. Implementa el uso de la GCP y la librería Keras que crea redes neuronales. Realiza su estudio con la finalidad de ver el impacto del final de temporada de la serie Game of Thrones. Por medio de la red social twitter, en donde, los diferentes usuarios compartían sus ideas antes, durante y después del capítulo, teniendo algunos resultados por medio la pila de tecnología de la plataforma de Google Cloud.

### 3. Metodología

La metodología utilizada para el desarrollo de la investigación dispondrá de seis fases genéricas dispuestas en la investigación de [9].

I. *Selección de fuentes de información* apropiadas.

II. *Preprocesamiento* agrupa tareas de transformación y filtrado de datos.

III. *Extracción* es la información relevante de los datos obtenidos.

IV. *Clasificación*, información es acuerdo a un modelo de dominio.

V. *Curación*, depura posibles errores y garantiza la calidad de información.

VI. *Acceso* es la definición de los métodos de acceso a la información.

Para los casos propuestos se tiene como objetivo común implementar el lenguaje natural y el procesamiento; permitiendo la extracción de información de una manera relevante por medio de algoritmos de NL. AutoML implementa el análisis semántico ya que así se verifica si la sentencia es coherente y tiene sentido de forma a la estructura que se desarrolla.

#### 3.1. Flujo de procesos caso A

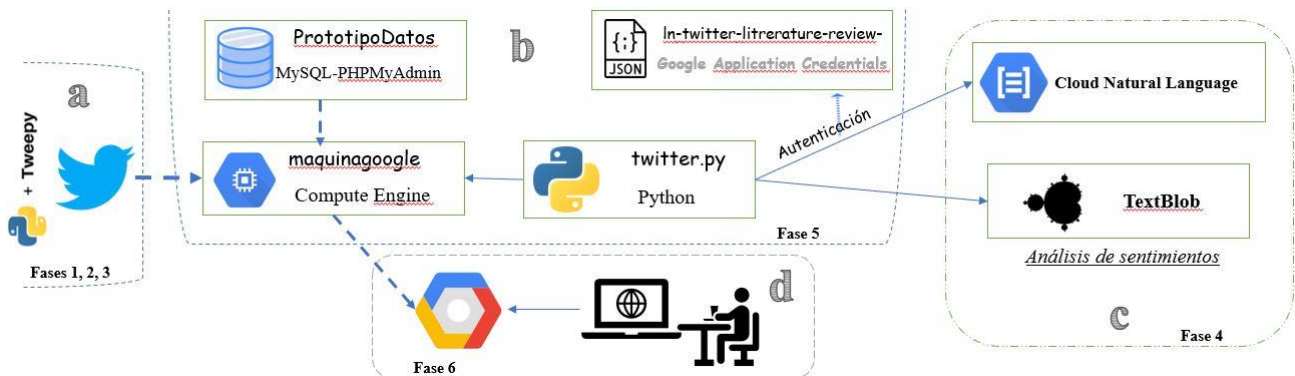


Figura 1. Flujo de procesos caso A. (a) Fases selección de fuentes de información, preprocesamiento, extracción (b) Fase curación (c) Fase clasificación (d) Fase acceso

### 3.1.1. Selección de fuentes de información

Definido el escenario de trabajo como se observa en la Figura 1A. En la red social twitter con ayuda de token de acceso, se procede con la búsqueda del hashtag. Se selecciona esta red social porque miles de usuarios twitteen alrededor de 65 millones según [10] en tiempo real. Nace la pregunta ¿Cómo se puede saber si es positivo o negativo el tweet?, todo esto se detallará en las siguientes fases.

Para el desarrollo se define el hashtag #WWIII, para el estudio del comportamiento por medio del análisis de sentimientos por las dos librerías definidas. El nacimiento es debido a que puede desatarse una tercera guerra mundial que por sus siglas es World War III.

### 3.1.2. Preprocesamiento

Como se observa en la Figura 1A, se describe el cómo es la extracción de los tweets, por medio de la librería tweepy. Se realiza la conexión con Twitter, realizando búsquedas correspondientes. La selección de tweets será en idioma inglés definido dentro del algoritmo de programación. Para una mejor valoración de sentimientos para las dos librerías google-natural-language y textblob.

### 3.1.3. Extracción

Miles de usuarios twitteen y esta es mostrada en tiempo real. Para su correcta extracción es necesario solicitar un token de acceso como desarrollador dentro de twitter.

Debido al gran volumen de datos se requiere tener una idea clara del análisis al cual se quiere llegar. Se toma en consideración que textblob trabaja correctamente con el idioma previamente definido; debido a que su estructura NLP es muy básica. Mientras que la librería cliente de GCP es mejor estructurada por el manejo de

inteligencia artificial. Todos estos algoritmos están siendo manejadas con lenguaje de programación Python.

En la Figura 1A se considera el proceso desde su extracción de los tweets por medio la librería tweepy, que se encuentra en el script llamado twitter.py. Una vez obtenida la información se procede analizar los tweets para obtener los sentimientos por medio de las dos librerías clientes; luego de ese proceso todos los resultados son guardados en la base de datos.

### 3.1.4. Clasificación

En las fases anteriores el objetivo analizar la opinión por medio de un algoritmo de análisis de sentimientos, una vez extraído los tweets.

Tabla 1. Valores análisis de sentimientos

Valor	Descripción
1	Valor más alto Positivo
0	Valor Neutral
-1	Valor más alto Negativo

En la Tabla 1, se observa los valores máximos. Se considerará un rango neutral el cual podemos determinar durante la ejecución del algoritmo.

Considerando que la red social Twitter según [10] alrededor de 30 millones de tweets son publicados en tiempo real alrededor del mundo, por ende, se clasifica 689 tweets para los respectivos análisis comparativos.

Tabla 2. Clasificación por categoría de sentimientos dos librerías

Clasificación	Descripción
Positivo	Textblob y CNL <sup>3</sup> positivos
Negativo	Textblob y CNL negativos
Neutral	Textblob y CNL con valores dentro de un rango

<sup>3</sup> Cloud Natural Language CLN



Positivo Negativo	Textblob y CNL cualquiera de ellos sea positivo y el otro negativo
Neutral Positivo	Textblob y CNL cualquiera de ellos sea neutral y el otro positivo
Neutral Negativo	Textblob y CNL cualquiera de ellos sea neutral y el otro negativo

En la Tabla 2, se considera seis tipos de clasificación; dando como resultados dichas combinaciones. Debido a que se comparará por dos librerías y se analizarán las precisiones por cada una de ellas.

### 3.1.5. Curación

Esta fase incluye datos de manejo de exactitud donde se analizará con mayor

profundidad en la sección de resultados del artículo.

### 3.1.6. Acceso

En la Figura 1D, el usuario accede al prototipo por medio de conexión SSH y por navegador web para observar los resultados. Todo se encuentra dentro de GCP desde la máquina virtual llamada maquinagoogle, la cual tenemos nuestros scripts Python, en este caso llamado *twitter.py*.

Para ver los resultados accede por dirección IP externa que la misma GCP da, se debe considerar que si se suspende la máquina virtual esta dirección se cambiara al reiniciar.



Figura 2. Página principal del caso A

Como se observa dentro de la Figura 2, se tiene un formulario el cual el usuario deberá llenar los campos, en donde la fecha de búsqueda coincidirá con alguna almacenada dentro de la base de datos. Para facilidad por medio de un combo se extraerá las búsquedas

automáticamente desde nuestra base de datos, dando mejor accesibilidad al proceso como se observa en la Figura 3.

Fecha de Búsqueda: 2020-01-03

Palabra clave: #WWIII

Rango: 0.0995

Graficar

- #Australia
- #Bolivia
- #Colombia
- #EEUU
- #InstagramDown
- #Iran
- #Iraq
- #Venezuela
- #Worldwar3
- #WWIII
- comercio
- Iran

Figura 3. Vista de datos página principal

Por último, se observa un campo rango, el cual este cumple un papel muy importante en este estudio dando resultados para la búsqueda. Este dato será variante debido al

criterio que dará el usuario para saber la neutralidad que desea tener para los tweets.

Una vez enviado los datos de la página principal, se tiene un reporte general de la librería textblob vs natural language de GCP, de los resultados obtenidos y por medio de la fecha que se desee analizar el comportamiento de ambas librerías visualizando como cada tweet toma su valor dentro de las tablas anteriormente visualizadas, así como se observa en la Figura 4.

Para finalizar se observa la página con los resultados de la comparación de las dos librerías de manera general, categorizada como se describió anteriormente.

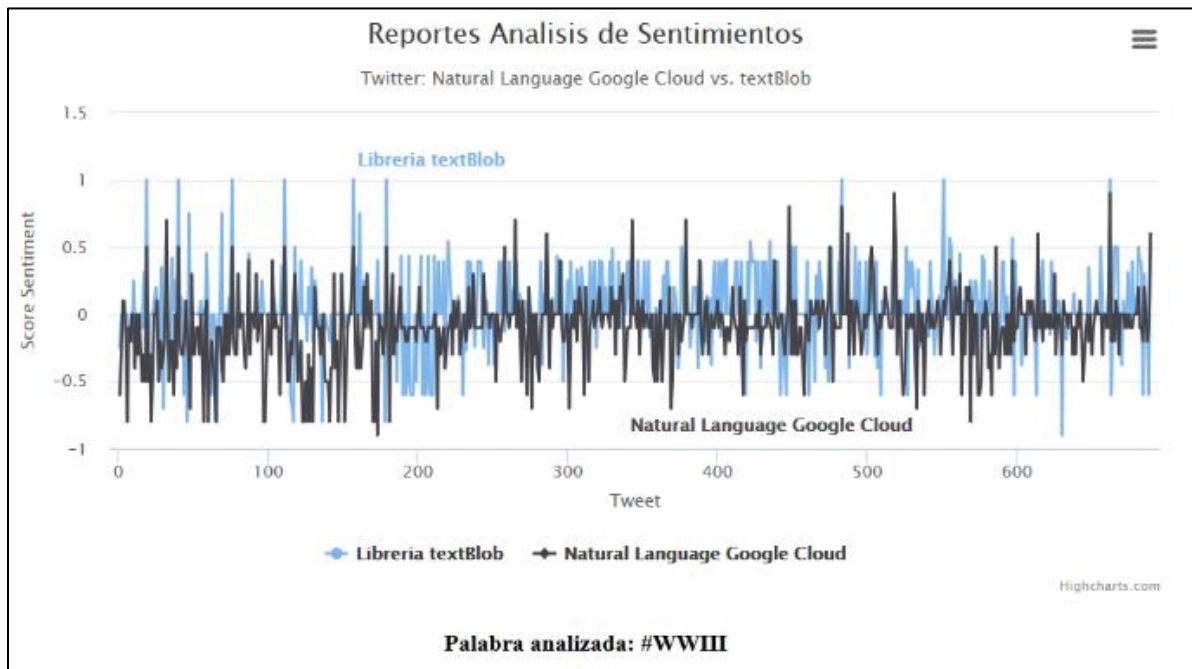


Figura 4. Página reporte general

### 3.2. Flujo de Procesos caso B

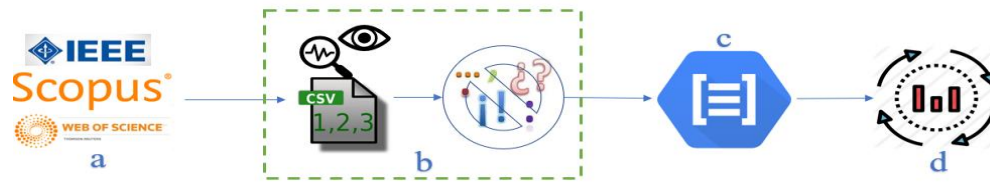


Figura 5. Flujo de procesos caso B

#### 3.2.1. Selección de fuentes de información

El investigador selecciona bibliotecas virtuales, las cuales el investigador realiza una búsqueda de referencias bibliográficas de “Grounded Theory in Software Development”.

#### 3.2.2. Preprocesamiento

Se observa en la Figura 5b la forma de pre-proceso, donde se ingresa primero un archivo .csv la cual este debe tener un texto y etiqueta. Este archivo debe contener un formato utf-8. Se realiza de forma manual la eliminación de signos de puntuación tales como: “,” y “;”. Además de símbolos que puedan alterar el significado.

El investigador por medio de SMS la cual describe al proceso en su etapa de búsqueda de artículos científicos alrededor de revisiones sistemáticas donde se observa diferentes tipos de objetivos, amplitud, problemas de validez, ya que así de forma complementaria se rige bajo el siguiente proceso [12]:

- a. Definición de preguntas de investigación
  - a. Revisión del alcance.
- b. Realizar búsqueda
  - a. Todos los artículos.
- c. Revisando los artículos
  - a. Criterios de inclusión
  - b. Criterios de exclusión
  - c. Artículos relevantes

#### 3.2.3. Extracción

Una vez extraída las referencias bibliográficas, el investigador obtiene una información

completa del tema. En ocasiones los motores de búsqueda extraen citas de referencias que incluían no en su totalidad. En ocasiones pueden repetirse información, por ende, el investigador realiza un proceso manual clasificando desde temas cercanos hasta los más lejanos a la búsqueda principal.

Debido que se requiere extraer de la búsqueda los artículos científicos más relevantes del tema “Grounded Theory in Software Development”. Pueden existir un gran volumen de datos y realizarlo de manera manual con llevaría demasiado tiempo. Por parte del investigador contempla la siguiente tabla de etiquetación:

Tabla 3. Tabla de etiquetación por parte del investigador Gustavo Navas

Etiqueta	Rango
AAA – AD	Mas cercano
AE – AH	Punto Medio
AS – AZ	Contiene palabras Grounded Theory
Mx	Archivos relacionados a la medicina
Nx	Relacionado a industrias

Como se observa en la Tabla 3, la información que se exporta de las fuentes principales sus motores de búsquedas toma en ocasiones algunas palabras de la búsqueda realizada, considerándoles como palabras claves, ya que por esta razón se pueden encontrar referencias bibliográficas de diversos campos.

Se considera fundamental que esta información previa a su análisis, el archivo .csv este delimitado por coma; dentro del contenido

se debe considerar que no exista el signo de puntuación coma “,”. En caso de no eliminar puede afectar a la interpretación de la información al momento de ser cargada al AutoML.

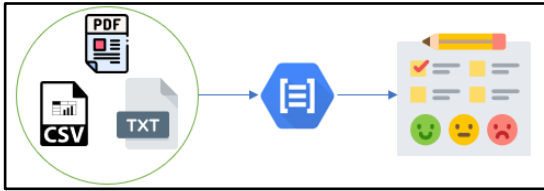


Figura 6. Funcionamiento AutoML Natural Language

En la Figura 6 se observa el funcionamiento del AutoML Natural Language desde la carga de documentos que contiene una etiqueta de texto basado en palabras claves y frases de dominio específico, entrenar el modelo donde ahí se clasifica, extrae y detecta sentimientos, entidades y textos [13].

### 3.2.4. Clasificación

Para este caso se considera, a partir del documento que fue facilitado por parte del investigador una nueva clasificación y etiquetación a criterio del desarrollador.

Tabla 4. Nueva clasificación y etiquetación

Antigua Etiqueta	Nueva Etiqueta
AAA-AB	N0
AC-AD	N1
AE-AH	N2
AS-AX	N3
AZ- Mx-Nx	N4

Como se observa en la Tabla 4, se tiene una nueva clasificación considerada por parte del desarrollador. El cual permite que se tenga una clasificación de acorde a niveles; N0 representa temas más cercanos, dado a que N4 se considera el nivel más alejado. Debido a que estos se encuentran temas que se desarrollan en diferentes campos como es la medicina, industrias, entre otros temas no equivalentes a la búsqueda.

### 3.2.5. Curación

Esta fase incluye datos de manejo de exactitud donde se analizará con mayor profundidad en la sección de resultados del artículo.

### 3.2.6. Acceso

Se ejecuta el script de este caso, por SSH debido a que esta almacenado de igual manera que en el caso anterior. Una vez adentro de la ventana ejecutamos la siguiente sentencia, tal como se observa en la Figura 7. Donde se coloca el nombre del script, el path de ubicación del archivo a evaluar y por último el path del proyecto del modelo este dato es dado por la misma GCP. Se describe dentro del comando la creación de un nuevo archivo con extensión .csv, por el cual se coloca los resultados automáticos de la clasificación

Los resultados obtenidos los tenemos como impresión dentro de la misma la misma plataforma, así como se muestra en la Figura 8.

```
amandacabascango@maquinagoogole:~$ python prediccion4.py '/home/amandacabascango/textopruebaA.txt' projects/478616320640/locations/us-central1/models/TCN4125028977834196992
```

Figura 7. Comando de ejecución

```
We never thought of a vasectomy: a qualitative study of men and women's counsel
ling around sterilization.

Predicted class name: N4
Predicted class score: 0.999980449677
Predicted class name: N3
Predicted class score: 1.51013036884e-05
Predicted class name: N0
Predicted class score: 2.94444998872e-06
Predicted class name: N1
Predicted class score: 9.42899475831e-07
Predicted class name: N2
Predicted class score: 4.3682362616e-07
```

Figura 8. Muestra del ejemplo de clasificación

## 4. Resultados

### 4.1. Recursos de Hardware

Para el desarrollo del prototipo para los casos A y B. Se tiene una sola máquina virtual, de las siguientes características Figura 12.

Imagen	Tamaño (GB)
ubuntu-1604-xenial-v20191024	10

Figura 9. Características VM

### 4.2. Caso A

Tabla 5. Información almacenada de tweets

	Tweet	vlgoogle	vltextblob
#WWIII	me acting mentally ill so i don't get drafted #WWIII <a href="https://t.co/FiyV9COFbP">https://t.co/FiyV9COFbP</a>	-0.6	-0.25
#WWIII	RT @Abdel7akime: Me: 3rd day at 2020 please 2020: be a good...	-0.3	0
#WWIII	RT @canjeer0: Me and the boys calculating our K/D ratios after #WWIII <a href="https://t.co/6PqfOljzrZ">https://t.co/6PqfOljzrZ</a>	0.1	0
#WWIII	I hope I make the all pro team after #WWIII is over	0.1	0
#WWIII	RT @7roximity: Xbox players when they see homosexuals can't be drafted: #WWIII <a href="https://t.co/jfFi2YOvKm">https://t.co/jfFi2YOvKm</a>	-0.2	0
#WWIII	RT @daniah_abd: Me purposely breaking my leg before the recruiters bust my door and take me #WWIII <a href="https://t.co/AbWAHmfcs8">https://t.co/AbWAHmfcs8</a>	-0.8	0
#WWIII	RT @metro_b00lin: When you hiding in the dead bodies and fart #WWIII <a href="https://t.co/KWX31om6o0">https://t.co/KWX31om6o0</a>	-0.1	-0.2

El enfoque de este análisis es ver el comportamiento de las dos librerías. En las etapas de preprocesamiento dentro de esta metodología, se utiliza dos librerías: textblob y cloud natural language. Su estructura interna de procesamiento por lo cual se obtiene en

ciertas ocasiones una contrariedad dentro de un mismo tweet; cómo se puede observar en la Tabla 5. El día que se realiza la búsqueda fue el tres de enero de 2020, se observa que se obtuvo 689 tweets que contiene #WWIII.

**Tabla 6.** Datos caso A

Verdaderos Positivo	Verdaderos Negativos	Falsos Positivos	Falsos Negativos
126	376	80	246

Como se observa en la Tabla 6, se considera a falsos negativos y positivos a la diferencia de tweets entre ambas librerías, dado el rango de 0.099; fue utilizado para la neutralidad de este, se realiza el análisis de exactitud para las dos librerías de manera independiente por medio de la siguiente fórmula:

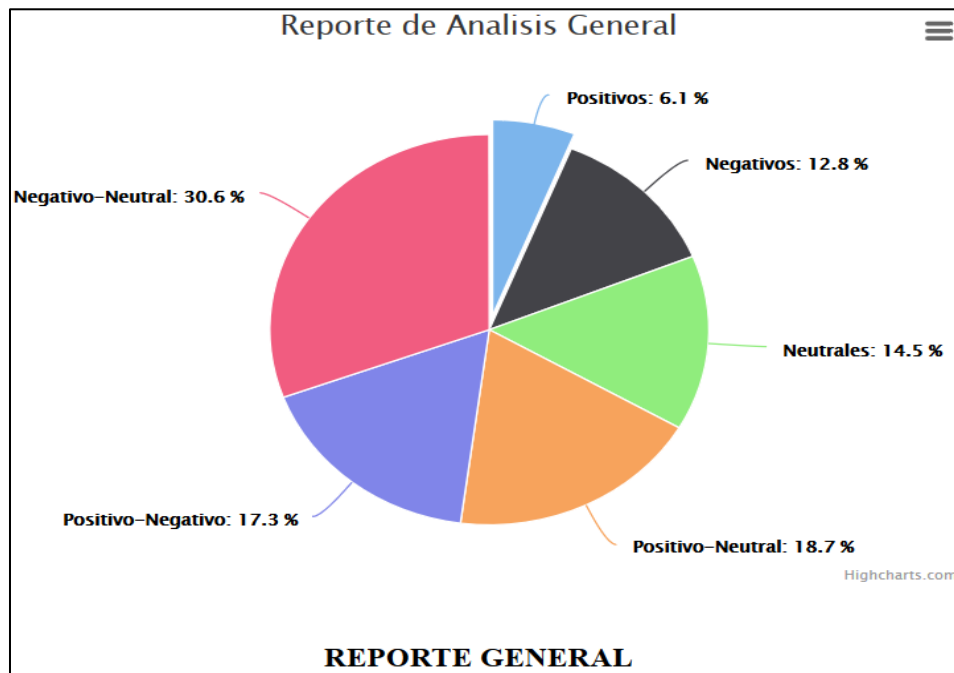
$$Exactitud = \frac{VP + VN}{VP + FN + FP + VN} \quad [1]$$

De la Ecuación 1 se considera las siguientes definiciones:

- VP: Verdadero Positivo
- VN: Verdadero Negativo
- FP: Falso Positivo
- FN: Falso Negativo

El cual aplicando los respectivos datos en la ecuación para cada librería se obtiene como resultado el *61% de exactitud* para Natural Language de GCP y con el *51%* a Textblob, por tal motivo se define a la librería Natural Language con mejor exactitud.

Se observa la predicción por las dos librerías tenemos resultados en un mismo tweet en donde existe una coincidencia de que son negativos. También existe diferente valoración dentro de las dos librerías y esto es debido a que textblob está manejando una polaridad y una subjetividad; en donde para nuestro sistema nos interesa la polaridad y para CLN tenemos propiamente el sentimiento. Durante el procesamiento de los tweets se observa que en ocasiones una librería obtiene una valoración negativa y la otra una positiva. Para tener un mejor análisis de cada tweet deberíamos considerar que las dos librerías coincidan por lo menos en la designación de negatividad, positivismo y neutralidad.



**Figura 10.** Reporte comparativo

Al analizar toda la información de los tweets y categorizando de acorde a la clasificación que se mencionó en la sección de clasificación; recalamos que ningún sistema es 100% exacto y como se vio en la ecuación 1 entre las dos librerías se maneja un porcentaje diferenciado de exactitud debido a que cada una maneja una estructura de procesamiento diferente. Por lo cual se considera un resultado aceptable que más del 50% de los tweets analizados corresponda a una categorización principal, es decir, a las tres clásicas: positivo, negativo y neutral.

### 4.3.Caso B

**Tabla 7.** Dataset caso B

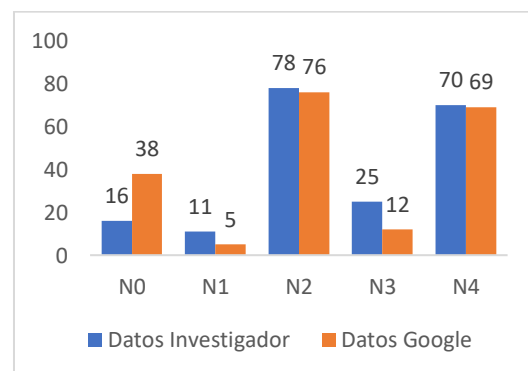
Titulo	Investigador	Google	Predicción
Learning strategies of autonomous medical students	N4	N2	0.999999523
Nursing process: what does it mean to nurses from Santa Cruz (Bolivia)?	N4	N4	0.998508155
A Case Report in Health Information Exchange for Inter-organizational Patient Transfers	N4	N4	0.970737755
A Case Study of Systems Development in Custom IS Organizational Culture	N4	N2	0.996367574
A Case Study of User-Centred Design in Four Swiss RUP Projects	N2	N2	0.999976635
A clinical decision support needs assessment of community-based physicians	N4	N4	0.999974489
A conceptual framework of challenges and solutions for managing global software maintenance	N0	N0	0.999917269

Considerando que la información que da por parte del investigador puede ser 100% exacta.

La exportación inicial total que se tiene de las bibliotecas fuentes es de 700 referencias bibliográficas. Después del proceso por parte del investigador considerando lo mencionado en el anterior párrafo se obtiene 635 referencias. Los cuales para el desarrollo de este proceso se inicia con la creación del nuevo modelo de personalización, cabe recalcar que es un resultado que fue obtenido por parte del investigador. Se procede a realizar el análisis con 200 títulos de

Como se ve en la Figura 10 existe una discrepancia de más del 50%, ya que entre las categorías de positivo-negativo, negativo-neutral y positivo-neutral; nos encontramos con el 66.6% de no coincidencias entre las dos librerías. Tenemos alrededor de 459 tweets sin una categorización igualitaria por medio de las dos librerías. Esto es una información variable debido a que damos un rango de neutralidad para el análisis de los tweets, se cabe recalcar que el análisis sin ese rango nos daría ya una valoración más acertada.

artículos. Teniendo como resultado la Figura 11.



**Figura 11.** Resultados investigador vs Natural Language

Al mencionar los resultados obtenidos por las dos partes, se obtiene:

- N0 se obtiene 38 artículos por NL vs 16 por parte del investigador,
- N1 se obtiene 11 por parte del investigador vs 5 por parte de NL
- N2 se obtiene 78 artículos vs 76 por parte del investigador

Un breve paréntesis en el análisis, como se mencionó anteriormente se debe tomar en cuenta que en estos puntos existe una cercanía

**Tabla 8.** Resultados comparativos

Etiquetas	Investigador	Google	Iguales	Diferencia	No iguales
<b>N0</b>	16	38	11	1	
<b>N1</b>	11	5	3	9	
<b>N2</b>	78	76	52	9	43
<b>N3</b>	25	12	7	6	
<b>N4</b>	70	69	59		
	200	200			

Se toma la cercanía entre cada nivel, para el respectivo análisis solo se considera las cercanías entre los niveles N0 a N1, N1 a N2 y N2 a N3. Al ver los resultados generales anteriormente explicados, tenemos una tabla general en la Tabla 8, que detalla cada uno de los resultados obtenidos durante la comparación de las dos fuentes (investigador y natural language de GCP). Considerando la diferencia de niveles en la columna “**diferencia**” se observa la cantidad de títulos de las referencias bibliográficas a clasificar en donde se obtiene los siguientes resultados:

- Entre N0 y N1: 1
- Entre N1 y N2: 9
- Entre N2 y N3: 9

No se considera N3-N4 debido a que son lejanos al tema general de búsqueda por parte del investigador. Y como dato final se obtiene que 43 títulos no tienen similitud alguna con el investigador. Para continuar con proceso se

dentro de la clasificación por parte del investigador.

- N3 se obtiene 25 artículos por NL vs 12 por parte del investigador
- N4 se obtuvo 70 artículos por NL vs 69 por parte del investigador

Donde en ambos casos la sumatoria es de 200 cada uno concordando con la cantidad inicial de datos.

considera una rubrica para los resultados de la clasificación.

**Tabla 9.** Rúbrica

RÚBRICA		
	> 0.75	1
IGUALES	<0.75 & > 0.50	0.6
	<0.5	0.3
ENTRE 1 NIVEL DE DIFERENCIA	> 0.75	0.5
	<0.75 & >0.50	0.3
NO IGUALES	> 0.75	0
	<0.75 & >0.50	0

La rúbrica establecida en la Tabla 9, permite dar una valorización a la predicción de cada título de la referencia bibliográfica debido a que al momento de realizar la clasificación por parte de nuestro modelo este retorna un valor decimal, como se observa en la Tabla 7, su etiqueta y el valor de



predicción. Estableciendo que 1 es máximo valor de predicción.

Para saber la precisión de nuestro modelo se aplica la ecuación 2, una vez ya obtenido los resultados.

$$Precision = \frac{v_P}{v_P + Fp} \quad [2]$$

Son considerados dentro de la ecuación los siguientes datos:

- VP= verdadero positivo
- FP = falso positivo

A partir de la ecuación anteriormente formulada se desarrolla de la siguiente manera:

$$P = \frac{132}{132 + 68}$$

$$P = 0.66$$

Dando como resultado un total de 0.66, equivalente al 66% de precisión donde se obtiene una diferencia del valor dado por la plataforma del 50%. Obteniendo como recomendación de 85 artículos, los cuales se descartan de etiqueta N3 y N4, referente al tema general.

En la Figura 12, se observan todos los resultados de manera general considerando el total de artículos (200), donde el 66% de los artículos clasificados son iguales al del investigador, el 12% se encuentran entre un nivel de diferencia y el 22% no tienen ninguna coincidencia, es decir, se encuentra entre más de dos niveles de diferencia.



Figura 12. Resultados finales referencias

Haciendo énfasis al 66% de artículos iguales. Se procede a realizar el respectivo análisis de los datos de artículos iguales con los del investigador tal como se observa en el Figura 13.

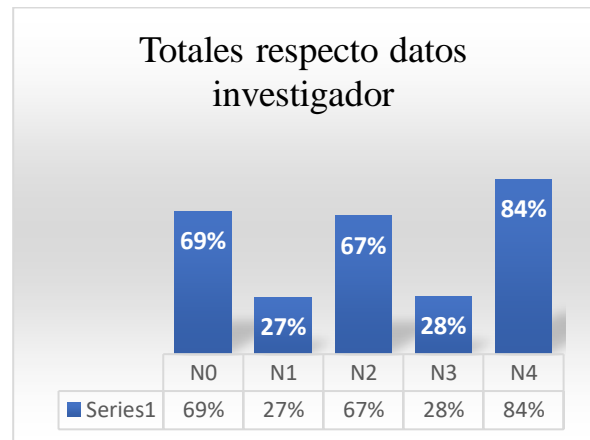


Figura 13. Resultados comparación art. Iguales con Investigador

Considerando al valor de cada etiqueta con la cantidad que se observa de datos iguales al comparar la del investigador vs la cantidad de títulos que coincidieron entre google y el investigador. Observando, se considera a los datos por parte del investigador como correctos se llega a considerar que la cantidad de artículos por parte del investigador son correctos. Se observa que las 5 etiquetas deberían cambiar a 3 etiquetas.

## **5. Conclusiones**

Considerando un aspecto del ser humano, las opiniones de estos pueden ser variantes debido a la situación y la influencia que tiene las redes sociales. El análisis dentro de las dos librerías existió la limitación del idioma, ya que el análisis de sentimientos se lo puede realizar correctamente.

GCP al ser una plataforma robusta permitió el funcionamiento de lenguaje natural ya sea por librería cliente y por un api rest, dado que la influencia de la información se dio al formato correspondiente.

Al considerar los parámetros de clasificación de textos, el nivel de acercamiento hacia el tema “Grounded Theory in Software Development”, donde se analiza por el método de SMS utilizado por el investigador y el modelo para AutoML que se basa en el proceso que el investigador da, con la diferencia de un nuevo método de clasificación de información, la cual con una precisión cerca del 50%, se debe considerar una gran cantidad de datos para que el modelo en términos simples aprenda nuestra clasificación a sí que al momento de evaluar nuevos títulos se podrá obtener un mejor resultado

Se recalca que con la precisión con la que se tiene se pudo observar que el nivel de clasificación para la etiqueta N4, su predicción es sumamente alta, obteniendo casi una misma similitud del por parte del investigador y al referirnos con títulos del campo de medicina la predicción es sumamente alta obteniendo en diferentes casos valores de 1 o 0.999, por tal motivo se considera como un nuevo modelo de clasificación la reducción de 5 a 3 etiquetas, unificando a N0 con N1, N2 con N3 y N4 quedaría manteniéndose como etiqueta única; permitiendo que esta aumente su nivel de predicción, aumentando las similitudes de los resultados de ambas partes.

Para incrementar la precisión se debe considerar grandes importaciones de datos y

nuevas preparaciones del modelo, permitiendo una mejor predicción y el porcentaje de fallo se irá disminuyendo, también se considera que con la clasificación por parte del desarrollador existe un poco de dificultad entre los niveles N2 y N3, ya que el nivel de granularidad es más profundo; permitiendo que aumente que entre los niveles sean más cercanos, debido a que el acercamiento de un título a otro del nivel cercano puede tener una granularidad sumamente igual.

Para los dos casos de análisis natural language de GCP, es una herramienta sumamente poderosa, ya que, con el correcto procesamiento, la información dada si se puede llegar a considerar como se mencionó en el modelo personalizado y la correcta selección de información, puede dar resultados sumamente buenos. En el análisis de twitter también se considera importante ya que si analizamos la información diaria solamente con la librería de la plataforma su análisis tendría mayor exactitud y no se obtendría resultados como fueron de los positivos negativos, neutrales positivos y los neutrales negativos, resultados debido al análisis entre las dos librerías. Como se recalca textblob si utiliza el procesamiento de lenguaje natural, pero a un nivel sumamente bajo, se debe considerar que la librería cloud natural language es importante, pero tiene un detalle que no puede diferenciar de un tweet si es negativamente en un aspecto específico. Dentro de este prototipo si recalca que, si es positivo o negativo el tweet, dando como resultado el comportamiento de cada una de ellas frente al hashtag buscado, se considera que para natural language #WWIII tiene un impacto negativo para los diferentes usuarios, mientras para la textblob su análisis da a un impacto neutral.

## Referencias

- [1] RD Station, «Redes Sociales,» Equipo de Marketing de Contenido, 12 03 2017. [En línea]. Available: <https://www.rdstation.com/mx/redes-sociales/>. [Último acceso: 01 12 2019].
- [2] Inteligencia del Cliente, «Procesamiento del lenguaje natural ¿qué es?,» 17 10 2017. [En línea]. Available: <https://www.iic.uam.es/inteligencia/que-es-procesamiento-del-lenguaje-natural/>. [Último acceso: 20 11 2019].
- [3] F. G. M. Rosas, «EL USO DE LA RED SOCIAL TWITTER COMO HERRAMIENTA PARA LA DIFUSION DE INFORMACION PUBLICA,» 11 2012. [En línea]. Available: [http://www.razonypalabra.org.mx/N/N81/V81/27\\_Meunier\\_V81.pdf](http://www.razonypalabra.org.mx/N/N81/V81/27_Meunier_V81.pdf). [Último acceso: 5 10 2019].
- [4] Google, «Cloud AutoML,» 15 11 2019. [En línea]. Available: <https://cloud.google.com/automl/?hl=es-419>. [Último acceso: 15 11 2019].
- [5] F.-F. Li y J. Li, «Cloud AutoML: hacer que la IA sea accesible para todas las empresas,» 17 01 2018. [En línea]. Available: <https://www.blog.google/products/google-cloud/cloud-automl-making-ai-accessible-every-business/>. [Último acceso: 5 12 2019].
- [6] A. Cortez, H. Vega y J. P. Quispe, «Revista de Ingenieria de Sistemas e Informatica vol. 6,» 07 2009. [En línea]. Available: [http://3A%2F%2Frevistasinvestigacion.unmsm.edu.pe%2Findex.php%2Fsystem%2Farticle%2Fdownload%2F5923%2F5121&usg=AOvVaw0eA1dCN2qyMo4E3Uh\\_FJPw](http://3A%2F%2Frevistasinvestigacion.unmsm.edu.pe%2Findex.php%2Fsystem%2Farticle%2Fdownload%2F5923%2F5121&usg=AOvVaw0eA1dCN2qyMo4E3Uh_FJPw). [Último acceso: 2019].
- [7] C. M. Becerra, «Análisis de sentimiento en Twitter:El bueno, el malo y el >:(,» 16 06 2016. [En línea]. Available: [https://rdu.unc.edu.ar/bitstream/handle/11086/3751/Becerra%202016\\_analisis-de-sentimiento.pdf?sequence=1](https://rdu.unc.edu.ar/bitstream/handle/11086/3751/Becerra%202016_analisis-de-sentimiento.pdf?sequence=1). [Último acceso: 10 2019].
- [8] T. Dehaene, «Game of Thrones Twitter Sentiment with Google Cloud Platform and Keras,» 22 05 2017. [En línea]. Available: <https://towardsdatascience.com/game-of-thrones-twitter-sentiment-with-keras-apache-beam-bigquery-and-pubsub-382a770f6583>. [Último acceso: 30 11 2019].
- [9] G. d. l. c. Velasco, «Modelo basado en tecnicas de procesamiento de lenguaje natural para extraer y anotar informacion de publicaciones cientificas,» abril 2014. [En línea]. Available: [https://www.google.com/url?sa=t&rc=t=j&q=&esrc=s&source=web&cd=1&cad=rja&uact=8&ved=2ahUKEwjmn5-MmeXmAhUSm1kKHxgCDRAQFjAAegQIBhAC&url=http%3A%2F%2Foa.upm.es%2F30856%2F1%2FGUILLERMO\\_DE\\_LA\\_CALLE\\_VELASCO.pdf&usg=AOvVaw3ILi5iljr0](https://www.google.com/url?sa=t&rc=t=j&q=&esrc=s&source=web&cd=1&cad=rja&uact=8&ved=2ahUKEwjmn5-MmeXmAhUSm1kKHxgCDRAQFjAAegQIBhAC&url=http%3A%2F%2Foa.upm.es%2F30856%2F1%2FGUILLERMO_DE_LA_CALLE_VELASCO.pdf&usg=AOvVaw3ILi5iljr0)

- 55KZk2nzAKeQ. [Último acceso: 30 12 2019].
- [10] Webempresa, «¿Qué es twitter? ¿Cómo funciona? ¿Cómo puedo usarlo para mi organización?», 01 03 2018. [En línea]. Available: <https://www.webempresa.com/blog/que-es-twitter-como-funciona.html>. [Último acceso: 17 02 2019].
- [11] J. C. M. Llano, «Estadísticas de redes sociales 2019: Usuarios de Facebook, Twitter, Instagram, YouTube, LinkedIn, Whatsapp y otros», 21 03 2019. [En línea]. Available: <https://www.juancmejia.com/marketing-digital/estadisticas-de-redes-sociales-usuarios-de-facebook-instagram-linkedin-twitter-whatsapp-y-otros-infografia/>. [Último acceso: 15 10 2019].
- [12] K. Petersen, R. Feldt, S. Mujtaba y M. Mattsson, «Systematic Mapping Studies in Software Engineering», de *12th international conference on evaluation and assessment in software engineering (Vol. 17, No. 1, pp. 1-10)*, 2008.
- [13] Google Cloud, «Natural Language», [En línea]. Available: <https://cloud.google.com/natural-language/>. [Último acceso: 30 11 2019].